By

Engineer Khattab Omar Abuisbae

ألمهندس خطاب عمر أبوأصبع ألسالم ألصقر

BIG DATA IS YOUR FUTURE

By

Engineer Khattab Omar Abuisbae

ألمهندس خطاب عمر أبوأصبع ألسالم ألصقر

Contents

I. Introduction.
II. To Whom I wrote This Book?
III. Acknowledgement.
IV. Personal Photos Gallery of the Author.
V. About The Author.

Chapter 1:

What's the Rate of the Salaries of people working in the field of Big Data for U.S.A?

1.1TopBest Prestige Paying Jobs in America in 2019, According to US News & World Report!

1.2Same occupation, different pay: How wages vary?

1.3The No. 1 highest paid, most in-demand job in every U.S. state

1.4How much is a Data Scientist's Salary?

1.5Data Science Jobs with their salaries

1.6Various job roles in data science

1.7Closing Remarks

1.8Salary for Skill: Big Data Analytics
1.9Is it still worth becoming a data scientist
1.10Should you think twice before training as a data scientist?
1.11What is the current entry level salary in Hadoop/big data fields in USA?
1.12Which Industry Pays the Highest Data Analyst Salary?
1.12.1What to Expect at Different Levels of Experience
1.12.2What to Expect in Different Roles
1.12.3How to Boost Your Data Analyst Salary
1.12.4Expand Your Skill Set
1.12.5Choose the Right Location
1.12.6Gain More Experience
1.12.7Negotiate Your Job Offer
1.12.8Pursue a New Degree or Certification
1.13You might also be interested in...
1.13.1Shareable Data Quotes That Will Change How You Think About Data
1.13.2Free Data Visualization Tools
1.13.3Data Analyst vs. Data Scientist
1.14 The Salaries of Data Scientists Remain Steady but Still Sexy With AI on the Horizon
1.14.1Data Scientists Salaries 2018
1.14.2Data Scientists' Average Salary 2017-2018
1.14.3Data Scientists: Top Required Skills
1.14.4Job Satisfaction by Position (Bar width corresponds to the number of respondents)

1.15 4 STEPS TO INCREASE YOUR SALARY AND BUILD A HIGH-PAYING CAREER AS A DATA ANALYST
1.16Data Analytics Job Market and Salary Trends
1.17Highest Paying Industries for Data Analysts
1.18 4 Ways to Increase Your Data Analyst Salary
1.19 Building Your Data Analytics Career
1.20 These Are the Highest Paying Jobs for the Class of 2019

1.21.1Highest Paying Jobs for New Grads
1.22What Kind of Salary Can You Expect as a Data Scientist?

1.23Wanted: Data Scientists (And Willing to Pay)
1.24Salaries by City and Region: Where They Are, Where they're headed
1.25Beyond Salary: Perks and Advantages of Being a Data Scientist
1.25.1Top Ranked Universities Ranked in Big Data
1.26Top 7 Big Data Jobs
1.2710 High-Paying Jobs That Require Knowledge of Data Analytics
1.28 What Skills Are Required for a Job in Data Analytics?
1.29 10 Jobs that Require a Knowledge of Data Analytics
1.30Data Analytics Opportunities around the Globe

Chapter 2:

Barriers, Obstacles, Challenges and Opportunities with Big Data!

Chapter 3:

An Introduction to Big Data Concepts and Terminology

Chapter 4:

DEPTH DIGING INTENSIVE Overview of Big Data Tools and Techniques, Discoveries and Pitfalls

Chapter 5:

Big Data Analytics

Chapter 6:

BIG DATA SICINETEST

Chapter 7:

DATA SICENCE

Chapter 8:

DATA ENGINEER

Chapter 9:

American Universities and Institutions Certified You in the Major of "Big Data"!

Chapter 10:

UNSTRUCTURED DATA

Chapter 11:

STRUCTURED DATA

Chapter 12:

Which American Main Organization & Institutions Using Big Data Centers?

Chapter 13:

How to Build a Big Data Center?

Chapter 14:

What Kind of Equipments You Need to Build Your Big Data Center?

Chapter 15:

IS IT WAS IT OR STILL A CENTERALATIC WAR OF BIG DATA CENTERS BETWEEN USA, CHINA, RUSSIA AND OTHER COUNTRIES?

Chapter 16:

Computers & Information Technology

Terminologies, Acronyms, Idioms and Glossaries!

Bibliography,

RESOURCES & REFERENCES

INTRODUCTION:

Information Technology fields is a vast, crucial, sophisticated, rapidly changed and information war used by every government and countries of this planet. Which is used to serve every crucial aspect of our lives?

One of the extraordinary, magnificent, attractive, controversial, interesting, newest, latest, hottest, top of the field and the most important strategic national weapon of information technology which leading to a war between governments and countries!

Big Data became the main core system of every organization, institutions, agencies, and the fields of sciences researches, technology adventure and inventions, manufacturer, marketing, sales, medical, Aviation, Airspace Invention, Elections & Voters, Weather Broadcasting, Economy, Wall Street & Stock Market, and Preventing & Arresting Criminals before Crime Occurred!

To Whom I wrote This Book?

I wrote this book for Investors, decision makers, entrepreneur, pioneers, businessmen and businesswomen, governments, scientists, universities, college's students, engineers and whom they want to start top peak demandable career of this era with highest prestige paid salaries; my buddy let me whisper in your ears? Piece-of-mind, old saying wisdom "Successful people; take the wheel not the back seat?"

ACKNOLEGEMENT

To my great parents Omar and Khadeja, brothers Belal, Mohammad and Khaled and sisters Khawla, Soha and Nisreen and the rest of my family's members and relatives, my Co-workers and friends at "Kuwait Institute for Scientific Research; Specially Professor "Adnan Shehab Aldeen" and others.

To all my instructors, Co-worker, friends and my students at "Kuwait Ministry of Education"; Specially my teacher "High School Priciple and his Deputy; "Abdullah Alokman" and "Dr. Saleh Sheshter".

To all athletics, sports celebrities and friends of "Qadesya Kuwait Sports Club and Salmya Kuwait Sports Club"; Specially my Colleagues Kuwait Soccer Celebrities at "Alromothya High School"; Sood Abu-hammd; "Jasem Yaqoob" and "Fasel Adeekheel" and Kuwait Sports Commentator Mr. "Khaled Alharban" and Kuwait Royal Families Members and my colleagues at my high school, and all my friends and brothers in Kuwait.

To all my USA Professors, Instructors, colleagues and students at "University of Dallas", "Weatherford College", "North Texas State University", "Oklahoma State University", "Oklahoma City University", "Central State University", "Los Angeles City College" and "Tree University".

To my "Democratic and Republican Parties Members"; to my friends and buddies Presidents of USA; President "Bill Clinton", "Hussein Barak Obama" and the rest of USA Presidents and Vice Presidents, Senates and Congress M embers.

To my brothers and buddies "Bill Gates", "Steve Jobs", "Larry Page ,Lawrence Edward Page" and "Sergey Brin".

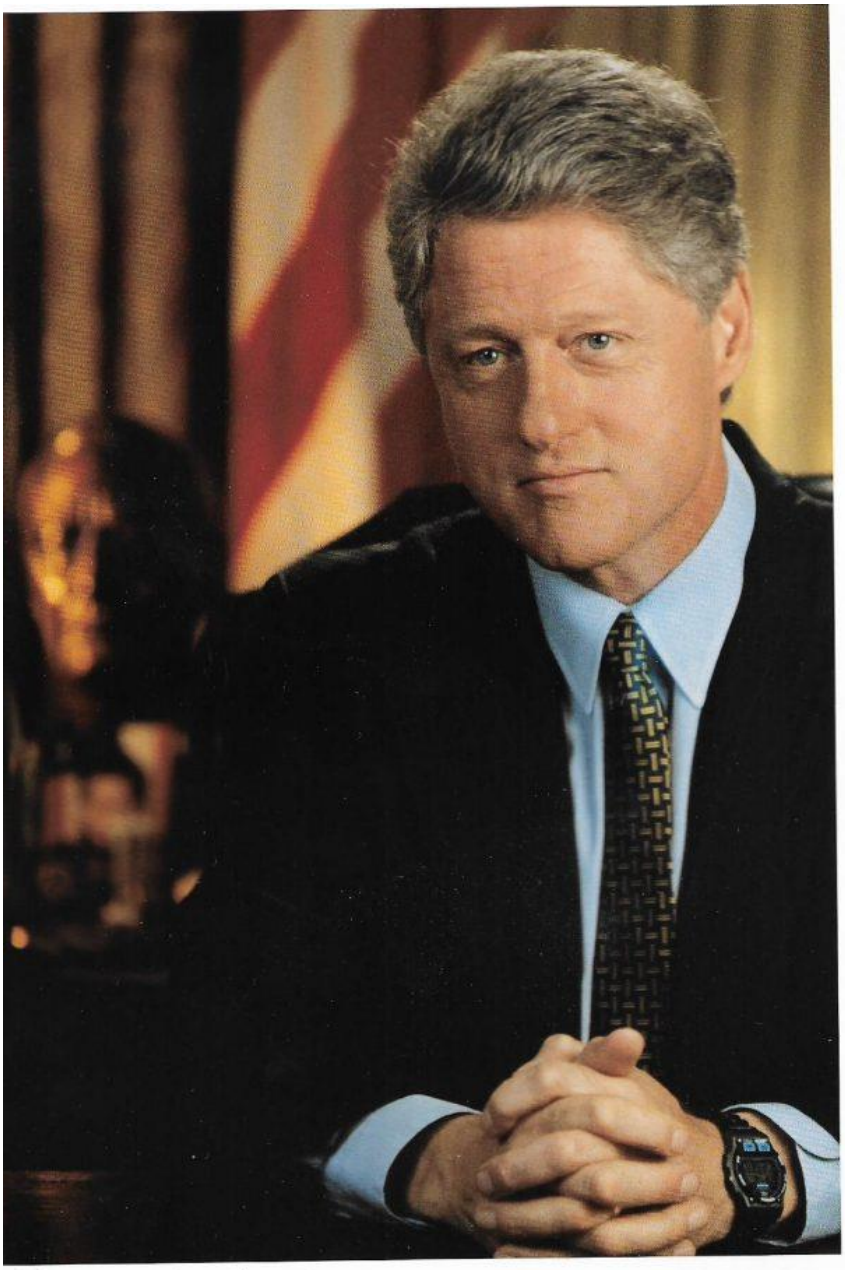
To all my friends and brothers all over USA!

1 Million Thanks with sincere warm hearted regards to the "Publisher's Staff Members" of this book.

My Best Regards & Wishes to the readers and buyers of this book all around the world!

"ALLAH" bless and mercy all of you!

Personal Photos Gallery of the Author



President Bill Clinton, Friend of Engineer Khattab Omar Abuisbae's Special Photo.

01 December 2000, Beverly Hills, Californai, USA, Engineer Khattab Omar Abuisbae's Photo.

THE WHITE HOUSE

WASHINGTON

February 1, 2000

Khattab Abuisbae
Post Office Box 3726
Torrance, California 90510-3726

Dear Khattab:

Thank you so much for your kind message. I am grateful for the words of encouragement and support I have received from people across our country and around the world.

As my Administration continues the work of moving America forward, I hope you will stay involved in the issues that affect us as individuals and as a nation. Best wishes.

Sincerely,

Bill Clinton

THE WHITE HOUSE
WASHINGTON

November 21, 2000

Khattab Abuisbae
Post Office Box 3726
Torrance, California 90510-3726

Dear Khattab:

On behalf of President Clinton, thank you for sharing your views. The President appreciates knowing your thoughts.

As this Administration continues to address the fundamental concerns of the American people and works to move the nation forward into the twenty-first century, we will keep your ideas in mind.

Best wishes.

Sincerely,

Delia A. Cohen
Special Assistant to the President
Director of Correspondence and
Presidential Messages

From Right to Left Engineer Khattab Omar Abuisbae and Ms. "Rana Kotam", Google's IT Specialist & "Strategic Planner of "Middle East and North Africa", Google Convention, Amman City, Jordan, Year 2012 Photo.

Engineer Khattab Omar Abuisbae, Istanbul, Turkey 2 September, 1978 Photo.

Engineer Khattab Omar Abuisbae at My Corporation Office, California Computers Services, Year 1993.

From Right to Left I Engineer Khattab Omar Abuisbae & My Buddy Mrs. Rebecca Olson, Executive Director of OMAN AMERICAN BUSINESS CENTER Photo1 of Year 2018

11

From Right to Left Engineer Khattab Omar Abuisbae & My Buddy Mrs. Rebecca Olson, Executive Director of OMAN AMERICAN BUSINESS CENTER Photo1 of Year 2018.

Photos of My Great Lovely Brothers, Buddie, Friends and Partners at "MargERP9+".

Photo of my Buddy and Partner at "MargERP9+" Mr. "Thakur Anup Singh"; Chairman & Managing Director.

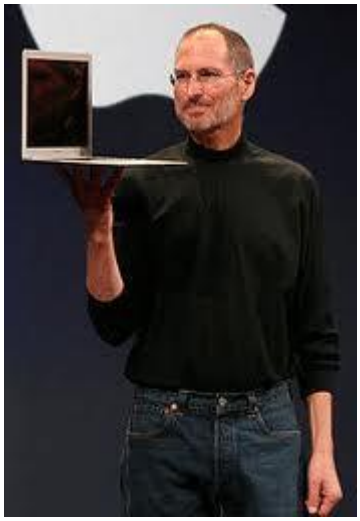
Photo of My Buddy and Partner; Mr. "Mahender Singh"; Co-founder & Managing Director (Technical) of "MargERP9+".

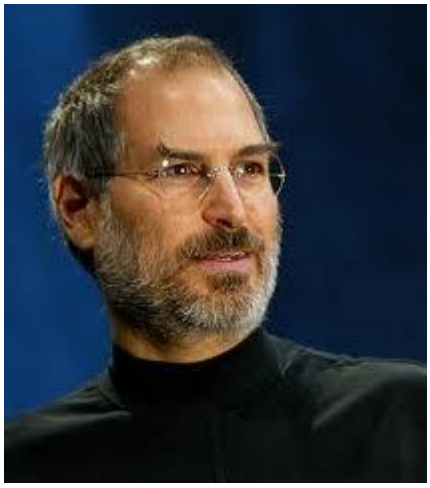
Photo of My Buddy and Partner; Mr. "Sudhir Singh"; Co-founder & Managing Director (Sales & Marketing) of "MargERP9+".

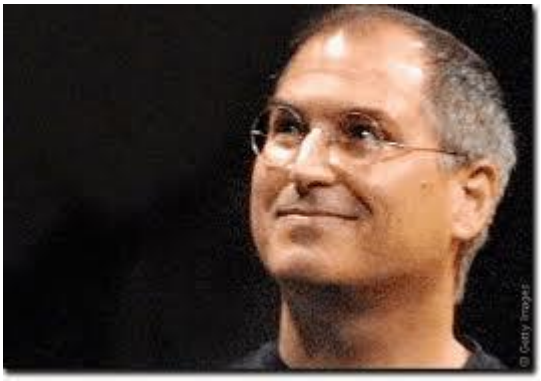
My Buddy “Bill Gates” Photo1

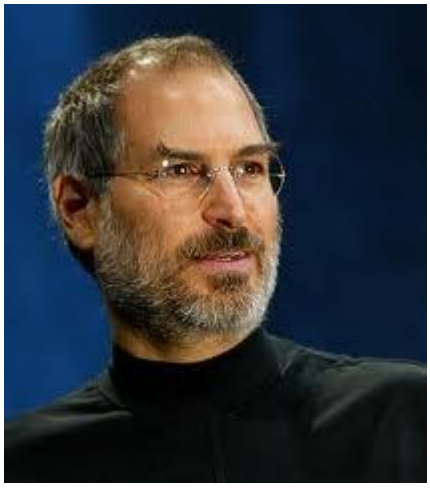
My Buddies “Bill Gates & Paul Alen” Photo.

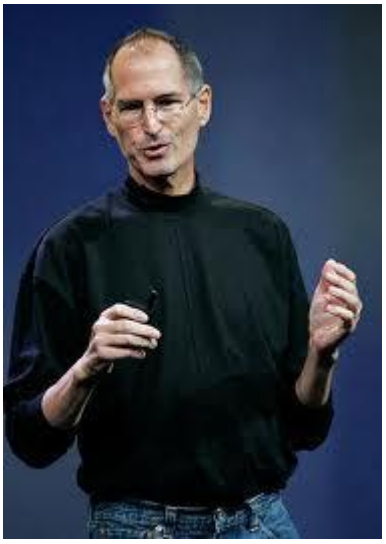
My Buddies “Steve Jobs” and “Steve Wozniak” Photos



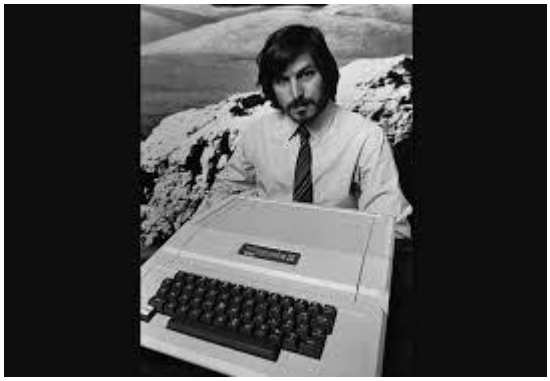


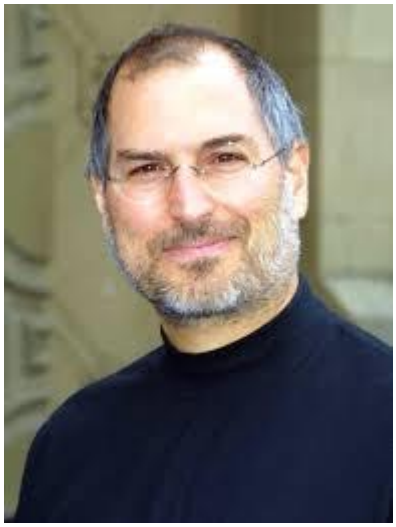






© Youtube





My Buddies “Larry Page” and “Sergey Prin” Photos

Google
I O

Google

Google

©AP

Royal Family of Kuwait Photos

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO6

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO7

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO8

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO1

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO2

SHAKHE SABAH AHAMMAD ALSABAH AMEER OF KUWAIT PHOTO3

PRINCE FAHAD ALAHMAD ALSABAH PHOTO1

PRINCE FAHAD ALAHMAD ALSABAH PHOTO3

My Sports Clubs “ALQADYSIA Kuwait Sports Club” & “SALMYA Kuwait Sports Club” Logos

ALKADESEYA CLUB OF KUWAIT'S LOGO3

ALKADESEYA CLUB OF KUWAIT'S LOGO

ALKADESEYA CLUB OF KUWAIT'S LOGO7

ALKADESEYA CLUB OF KUWAIT'S LOGO2

ALKADESEYA CLUB OF KUWAIT'S LOGO4

ALKADESEYA CLUB OF KUWAIT'S LOGO2

Salmya Kuwait Sports Club Logo1

Kuwait Institue for Scientific Research Logo

Salmya City, Kuwait, My Great and Beloved My Mother Khadeja and My Dad Omar Ragheb Abuisbae's Photo.

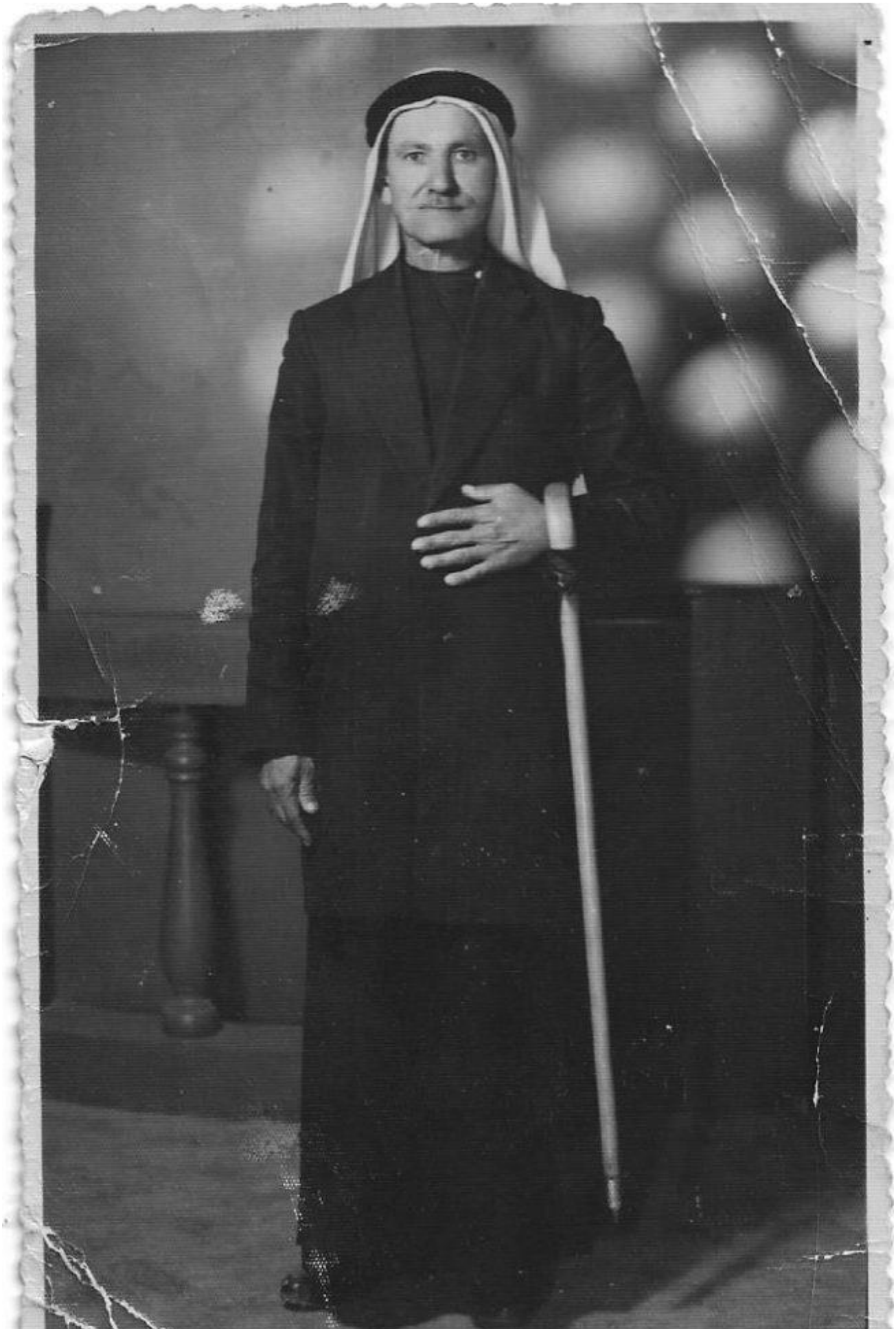
My Lovely and Beloved Greatest Uncle Ahammad Khaleel Aboosba; who the first one invented and founded "Apple Computer" under "Arab Development Establishment for Computers" at "Salama Building", 5th Circle in Amman City of Jordan the Year 1980, Photo1

My Lovely and Beloved Greatest Uncle Ahammad Khaleel Aboosba; who the first one invented and founded "Apple Computer" under "Arab Development Establishment for Computers" at "Salama Building", 5th Circle in Amman City of Jordan the Year 1980, Photo2

My Lovely and Beloved Greatest Uncle Ahammad Khaleel Aboosba; who the first one invented and founded "Apple Computer" under "Arab Development Establishment for Computers" at "Salama Building", 5th Circle in Amman City of Jordan the Year 1980, Photo3

Year 1964, Salmya City, Kuwait, at Our Home's Garden,From Left to Right My Beloved Great Grandfather Ragheb, Engineer Khattab Omar Abuisbae and Uncle Hassan's Photo

Year 1964, Salmya City, Kuwait, at Our Home's Garden,From Left to Right My Beloved Great Grandfather Ragheb, Engineer Khattab Omar Abuisbae and Uncle Hassan's Photo.

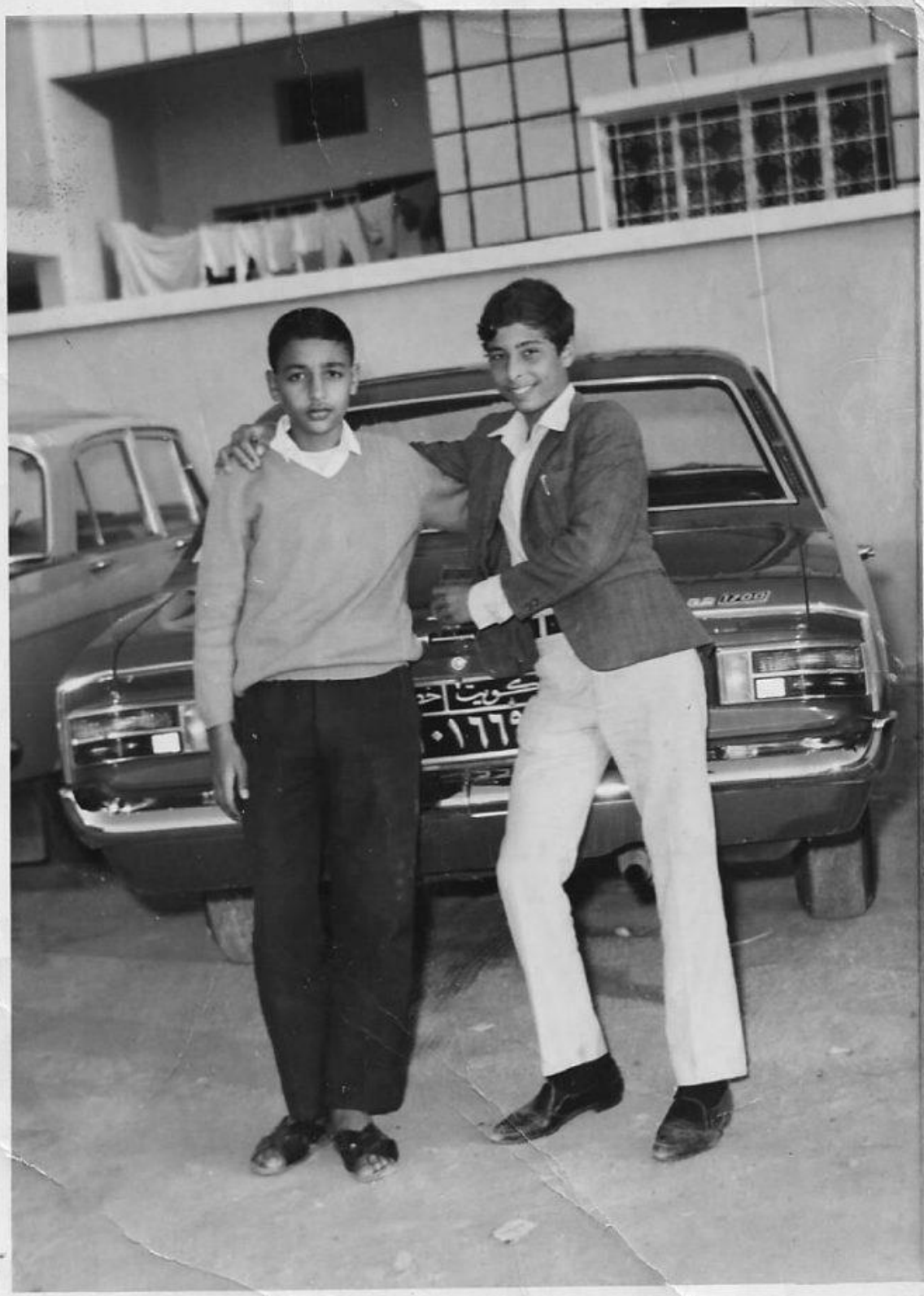


My Great Beloved Grandfather Khaleel Ahammad Aboosba, the leader of "Palestine-en Commandos" toward the "British Colonialism of Palestine" Photo.

March 14, 1958, Zarka City, Jordan/Engineer Khattab Omar Abuisbae's Photo.

Year 1960, Zarka City, Jordan, From Left to Right, My Beloved Sister Khawlla and I, Engineer Khattab Omar Abuisbae's Photo.

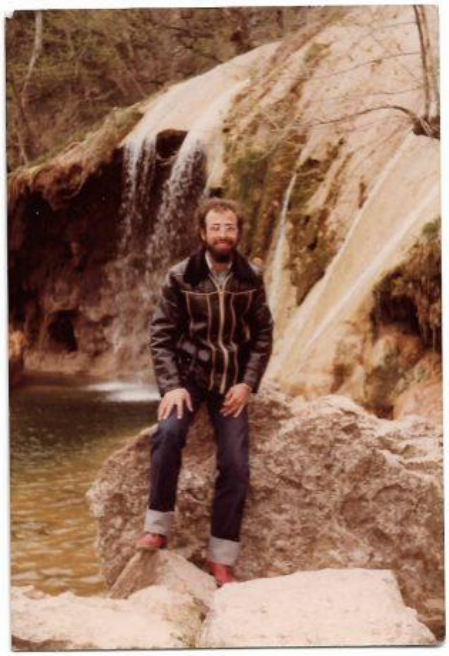
Salmya City, Kuwait,Engineer Khattab Omar Abuisbae At Age of 6's Photo.



March 1971, Salmya, Kuwait from Right to Left, Engineer Khattab Omar Abuisbae and Yousef's Photo.

OKLAHOMA CITY UNIVERSITY
Commencement Ceremony
May 10, 1986

Oklahoma City University's Commencement Ceremony of Engineer Khattab Omar Abuisbae, May 10th, 1986.



Oklahoma Falls, Year 1980, University's Trip, Engineer Khattab Omar Abuisbae's Photo1.

Oklahoma Falls, Year 1980, University's Trip, Engineer Khattab Omar Abuisbae's Photo2

SEP 80

From Left to Right, Engineer Khattab Omar Abuisbae and my Roommate Engineer Kamal Mohammad Sadah, at our apartment, Santafee Street, Weatherford City, Texas, 1st September 1980 Photo.

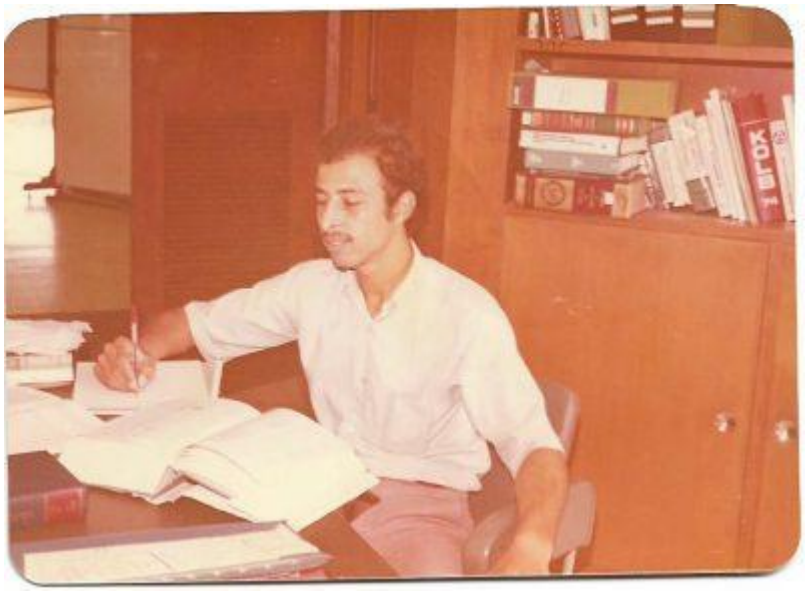
Year 1980, at my apartment of Weatherford City, Texas, USA, from Left to Right Engineers Basem Saforie,Engineer Khattab Omar Abuisbae and My Roommate Engineer Kamal Saadah's Photo.

Engineer Khattab Omar Abuisbae's Car, My Car Name "ALBARRA", Weatherford City, Texas Year 1981 Photo.

Engineer Khattab Omar Abuisbae in the middle, and his Colleagues at "Oklahoma City University's Commencement Ceremony", May 10th, 1986 Photo.

07 March 1990, Byan City, Kuwait, In the Middle Engineer Khattab Omar Abuisbae's Photo.

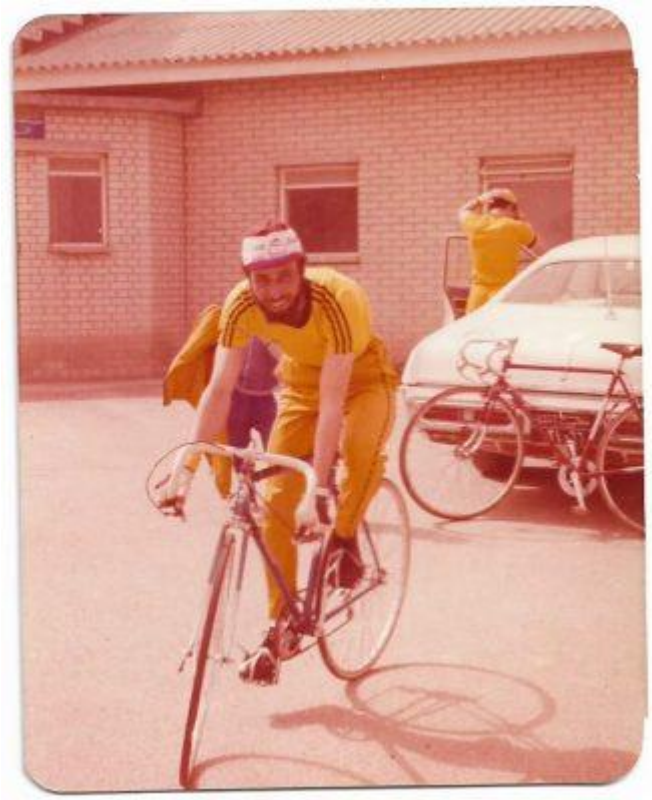
From Left to Right; Engineer Khattab Omar Abuisbae and His Friends Professors of "Information Technology", "Kuwait Computer Convention", "Kuwait National Musem Tour", Kuwait City, Kuwait of Year 1989 Photo2.



Year 1976, My Office at Kuwait Institute for Scientific Research, Engineer Khattab Omar Abuisbae's Photo.

Year 1989, Kuwait's National Cultural Museum Tour, to the Left, Engineer Khattab's Photo.

Year 1989, Kuwait's National Cultural Museum Tour, Second to the Left, Engineer Khattab's Photo.



ENGINEER “KHATTAB OMAR ABUISBAE” PREPARING HIMSELF TO START THE RACE COMPETITION FOR KOC OF KUWAIT ROAD BICYCLE CHAMPIONSHIP Represinting “Alqadsya Kuwait Sports Club” OF YEAR 1989 Photo.

Engineer Khattab Omar Abuisbae "Championship of KOC of Kuwait Road Bicycle Race", Friday 12th, March of Year 1976 Photo.

Engineer Khattab Omar Abuisbae "Championship of KOC Kuwait Road Bicycle Race", Friday 12th, March of Year 1976 Photo.

Final Finish Point of Kuwaiti Road Bicycle Race Championship, Friday 12th, March 1976, Alahmadi City, Engineer Khattab Omar Abuisbae's Photo.

KOC of Kuwait National Sports Day, Friday of March 12, 1976, Road Bicycle Race Championship Engineer Khattab Omar Abuisbae's Photo.

Friday of March 12, 1976, KOC of Kuwait National Sports Day, Road Bicycle Race Championship Engineer Khattab Omar Abuisbae's Photo.

Final Finish of KOC of Kuwait Road Bicycle Race Championship of Friday 12th, March 1976, Engineer Khattab Omar Abuisbae's Photo2.

August, 1975, Kuwait Summer National Sports Clubs Road Bicycle Race, Championship Engineer Khattab Omar Abuisbae's Photo2.

August, 1975, Kuwait Summer National Sports Clubs Road Bicycle Race, Championship Engineer Khattab Omar Abuisbae's Photo.

ABOUT THE AUTHOR:

Engineer Khattab Omar Abuisbae

1. **Born in the city of Amman, Jordan in December 25th, 1957.**
2. **I grew up in Kuwait since the year 1962 at Salmya City.**
3. **Graduated from "Alromythya High School" Now "Palestine High School" in 1976.**
4. **Summer 1976 worked at "Kuwait Institute for Scientific Research", "Marine Pollution & Desert Research Department" as "Laboratory Technician".**
5. **1987 Graduated from "Oklahoma City University with a B. S degree "Computer Systems Engineer" with GPA 2.67 out of 4.0.**
6. **1987 worked at "Kuwait Ministry of Education"**
7. **1996 Graduated from "Los Angeles City College" in "American Law" & "Management Information System" with GPA 3.97 out of 4.0.**
8. **He is a founder & owner a couple of corporations:**

1. **Discovery Computers & Electronics Engineering Corporation**
2. **California Computers Services Corporation.**
3. **Banks Core System Corporation.**
4. **Marg of USA Corporation.**
5. **Medical Doctor Software Corporation.**
6. **America DataBases Designers Corporation.**
7. **America Computer Group Corporation.**
8. **America Computers Clinic Corporation.**
9. **America Computers Accessories Corporation.**
10. **Creative World Trade Group.Online**
11. **Creative World Trade Corporation.**
12. **USATrade.Cheap Corporation.**
13. **Omar Shop Online Corporation.**
14. **eCommerce100 Corporation,**
15. **My USA Marketing Corporation.**
16. **USA Warm House Corporation.**
17. **USA Selfie Photos Corporation.**
18. **University I20 Corporation.**

- **He is a "Independent Professional Soccer Player".**
- **He is Championship of Kuwait Road Bicycle Race more than once.**
- **His Sports Clubs "Alqadsyia Kuwait Sports club" & "Salmya Kuwait Sports Club".**
- **During his high school he was "President of Palestine Students Association".**
- **At "Los Angeles City College" he was holding the position of "Administrative Legislative" at the "Student Government" under Mr. President James, with Financial Budget of $265, 000.**
- **He is a member of the American Democratic Party, Liberal.**
- **His Religion is Islam.**

Chapter 1:

Bro & Buddy Do You Know?

What's the Rate of the Salaries of people working in the field of Big Data for U.S.A?

1.1TopBest Prestiege Paying Jobs in America in 2019, According to US News & World Report!

---

1. Anesthesiologist

Mean salary: $265,990
Stress level: High
Work-life balance: Low

---

2. Surgeon

Mean salary: $251,890
Stress level: High
Work-life balance: Low

---

3. Oral and Maxillofacial Surgeon

Mean salary: $242,740
Stress level: Above average
Work-life balance: Below average

4. Obstetrician and Gynecologist

Mean salary: $235,240
Stress level: Above average
Work-life balance: Below average

The service of orthodontists and dentists are practically an essential in this day in age. The high demand for the examination, diagnoses, and treatment of diseases, injuries and malformations of teeth and the gums makes this occupation among the most highly paid professions in the United States.

5. Orthodontist

Mean salary: $229,380
Stress level: Below average
Work-life balance: Above average

6. Psychiatrist

Mean salary: $216,090
Stress level: Above average
Work-life balance: Average

7. Physician

Mean salary: $198,370
Stress level: High
Work-life balance: Average

8. Prosthodontist

Mean salary: $196,960
Stress level: Above average
Work-life balance: Below average

9. Pediatrician

Mean salary: $187,540
Stress level: Above average
Work-life balance: Average

10. Dentist

Mean salary: $174,110
Stress level: Average
Work-life balance: Above average

11. Nurse Anesthetist

Mean salary: $169,450
Stress level: Above average
Work-life balance: Below average

12. Petroleum Engineer

Mean salary: $154,780
Stress level: Above average
Work-life balance: Below average

13. IT Manager

Mean salary: $149,730
Stress level: High
Work-life balance: Below average

14. Podiatrist

Mean salary: $148,470
Stress level: Average
Work-life balance: Average

15. Marketing Manager

Mean salary: $145,620
Stress level: Above average
Work-life balance: Below average

16. Financial Manager

Mean salary: $143,530
Stress level: High
Work-life balance: Average

17. Lawyer

Mean salary: $141,890
Stress level: High
Work-life balance: Below average

18. Sales Manager

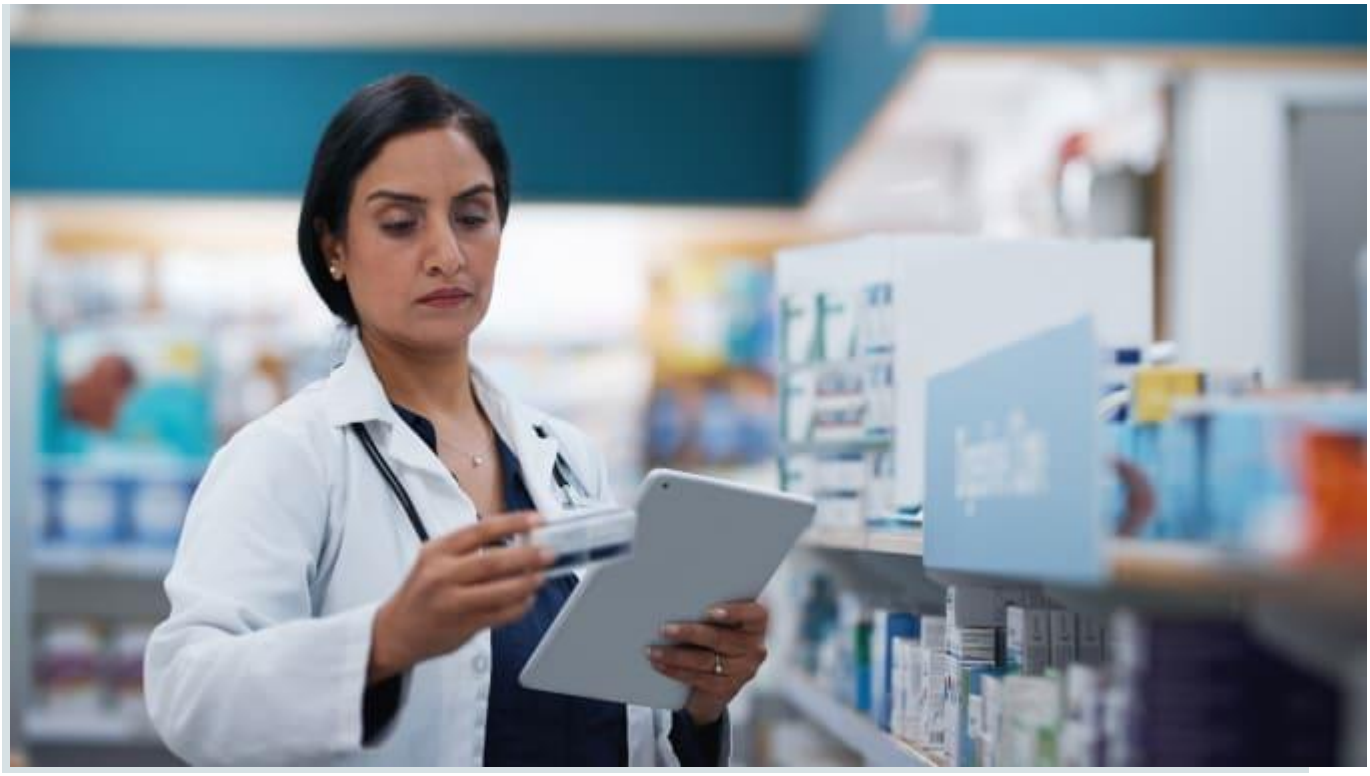
Mean salary: $137,650
Stress level: High
Work-life balance: Below average

19. Financial Advisor

Mean salary: $124,140
Stress level: Above average
Work-life balance: Average

20. Business Operations Manager

Mean salary: $123,460
Stress level: Above average
Work-life balance: Below average



21. Pharmacist

Mean salary: $121,710
Stress level: Above average
Work-life balance: Below average

22. Optometrist

Mean salary: $119,100
Stress level: Average
Work-life balance: Above average

23. Actuary

Mean salary: $114,850
Stress level: Average
Work-life balance: Average

24. Political Scientist

Mean salary: $112,030
Stress level: Average
Work-life balance: Average

25. Medical and Health Services Manager

Mean salary: $111,680
Stress level: Above average
Work-life balance: Above average

1.2 Occupational Employment Statistics

May 2018 National Occupational Employment and Wage Estimates

United States

Occupational Employment and Wages, May 2018

Computer and Information Systems Managers

Plan, direct, or coordinate activities in such fields as electronic data processing, information systems, systems analysis, and computer programming.

1.2.1National estimates for this occupation: Top

Employment estimate and mean wage estimates for this occupation:

Employment (1)	Employment RSE (3)	Mean hourly wage	Mean annual wage (2)	Wage RSE (3)
391,430	0.7 %	$73.49	$152,860	0.4 %

Percentile wage estimates for this occupation:

Percentile	10%	25%	50% (Median)	75%	90%
Hourly Wage	$41.05	$52.94	$68.53	$86.63	(5)
Annual Wage (2)	$85,380	$110,110	$142,530	$180,190	(5)

1.2.2Industry profile for this occupation: Top

Industries with the highest published employment and wages for this occupation are provided.

Industries with the highest levels of employment in this occupation:

Industry	Employment (1)	Percent of industry employment	Hourly mean wage	Annual mean wage (2)
Computer Systems Design and Related Services	91,520	4.37	$76.82	$159,790
Management of Companies and Enterprises	39,710	1.67	$74.74	$155,460
Software Publishers	14,590	3.70	$82.11	$170,790
Management, Scientific, and Technical Consulting Services	14,540	1.01	$77.44	$161,070
Insurance Carriers	13,320	1.14	$73.50	$152,870

Industries with the highest concentration of employment in this occupation:

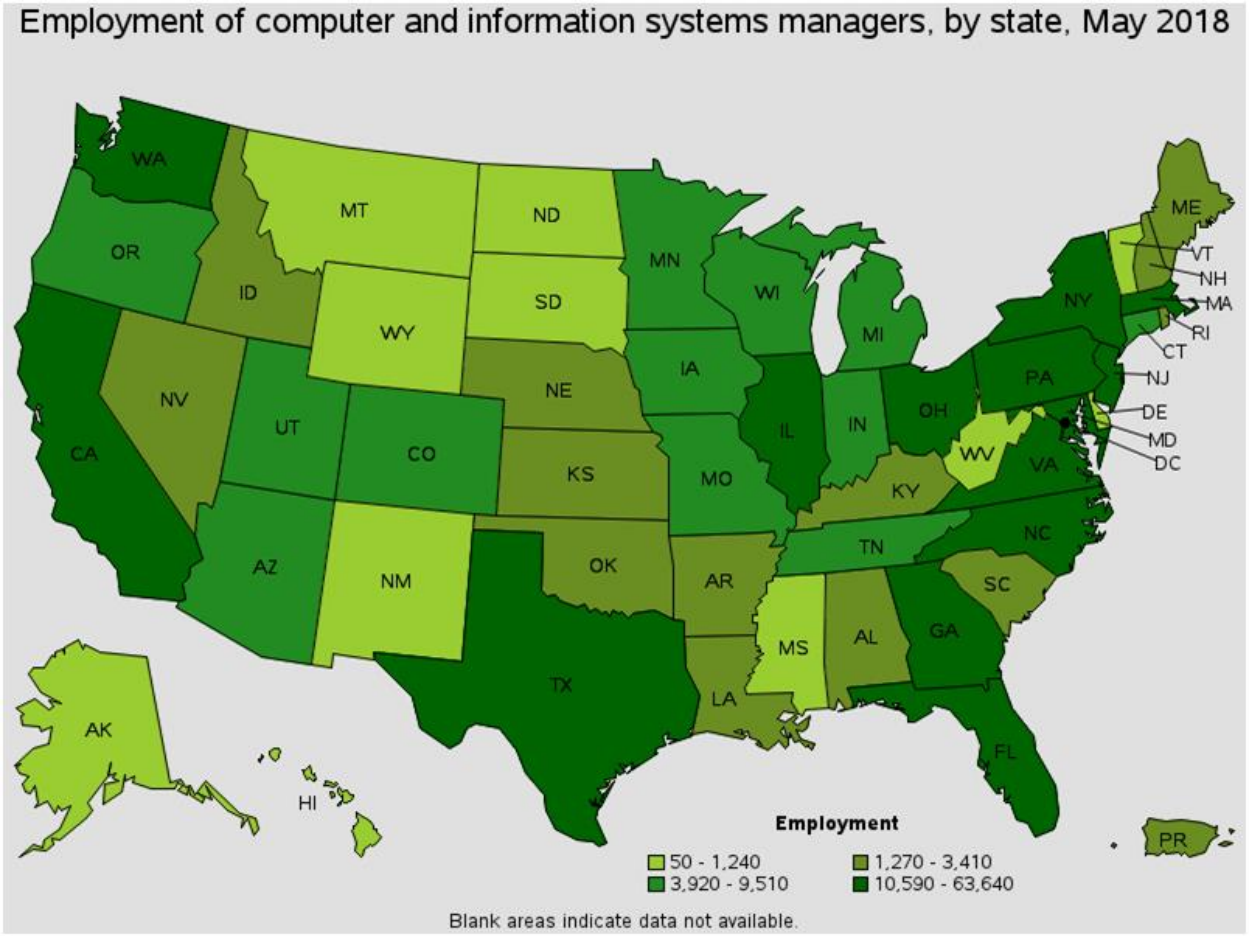
Industry	Employment (1)	Percent of industry employment	Hourly mean wage	Annual mean wage (2)
Computer Systems Design and Related Services	91,520	4.37	$76.82	$159,790
Software Publishers	14,590	3.70	$82.11	$170,790
Other Information Services	9,790	3.31	$91.71	$190,760
Data Processing, Hosting, and Related Services	10,250	3.22	$77.68	$161,570
Monetary Authorities-Central Bank	380	2.07	$83.03	$172,690

Top paying industries for this occupation:

Industry	Employment (1)	Percent of industry employment	Hourly mean wage	Annual mean wage (2)
Clothing Stores	120	0.01	$93.07	$193,580
Other Information Services	9,790	3.31	$91.71	$190,760
Securities, Commodity Contracts, and Other Financial Investments and Related Activities	5,950	0.64	$89.68	$186,540
Agents and Managers for Artists, Athletes, Entertainers, and Other Public Figures	60	0.21	$87.38	$181,750
Semiconductor and Other Electronic Component Manufacturing	2,700	0.74	$86.49	$179,890

1.2.3Geographic profile for this occupation: ***Top***

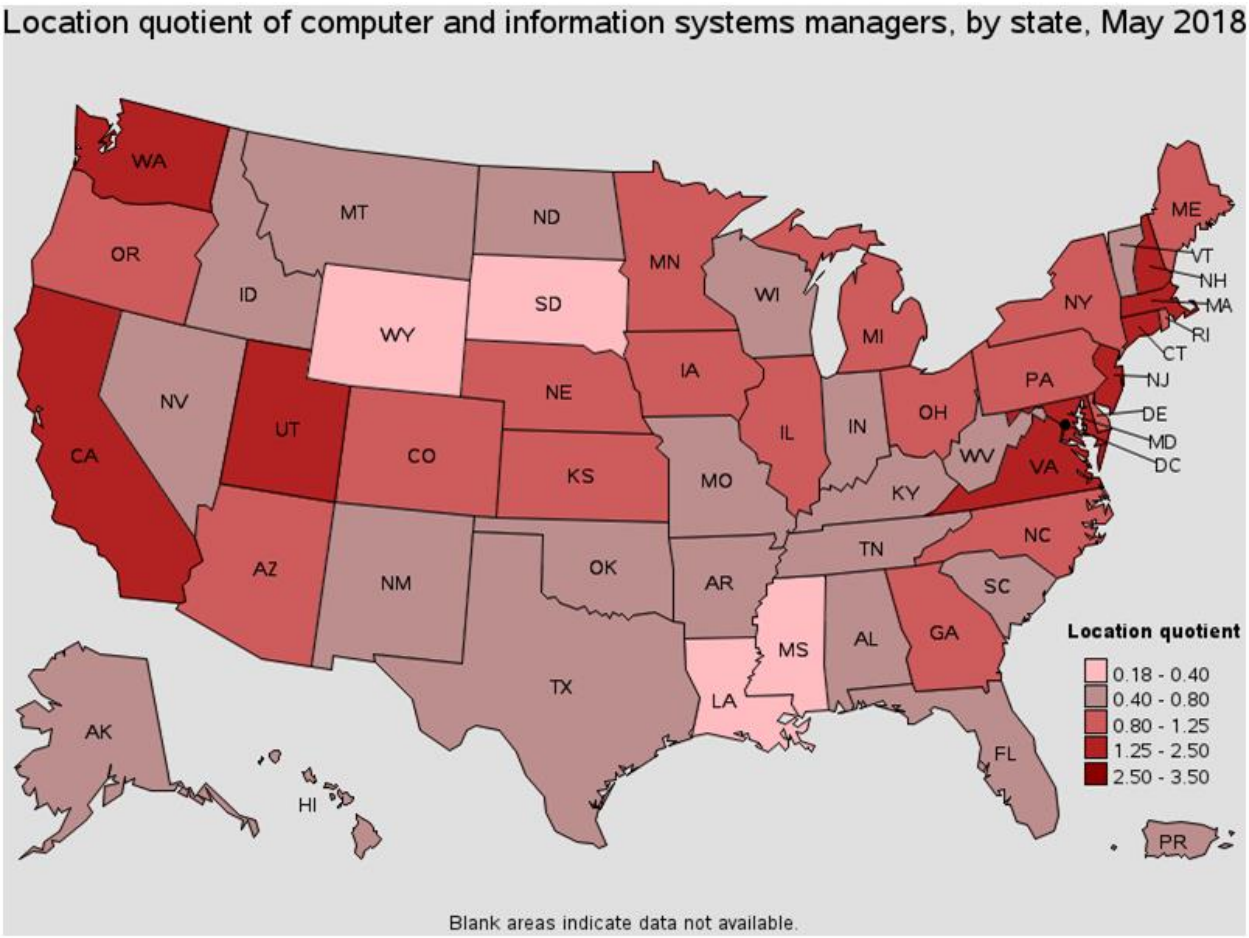
States and areas with the highest published employment, location quotients, and wages for this occupation are provided.



1.2.4States with the highest employment level in this occupation:

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
California	63,640	3.74	1.38	$86.66	$180,250
New York	27,070	2.88	1.07	$91.50	$190,310
Texas	22,350	1.84	0.68	$74.35	$154,650
Massachusetts	17,820	4.99	1.85	$75.30	$156,620
Illinois	17,770	2.97	1.10	$69.18	$143,900

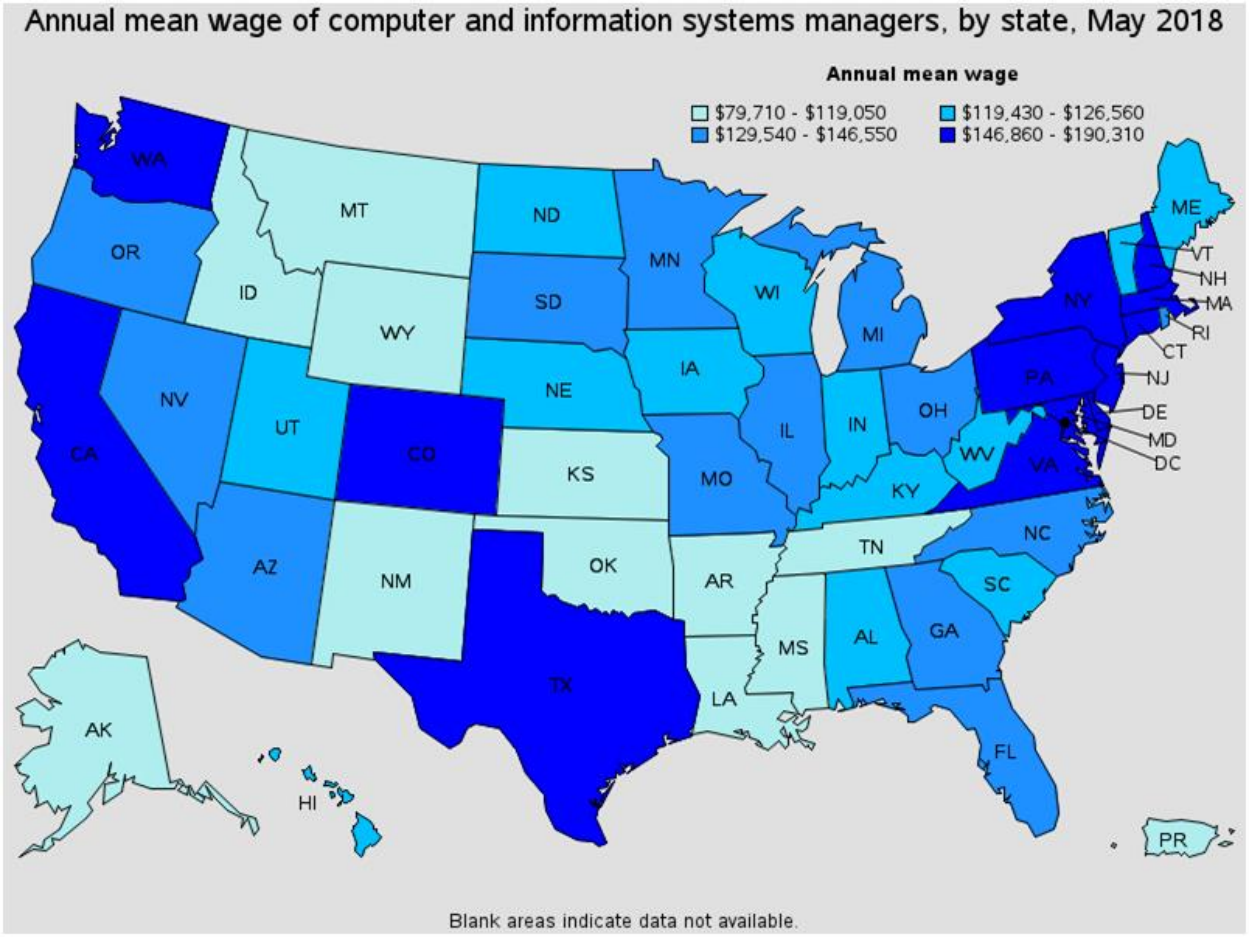
Location quotient of computer and information systems managers, by state, May 2018



Blank areas indicate data not available.

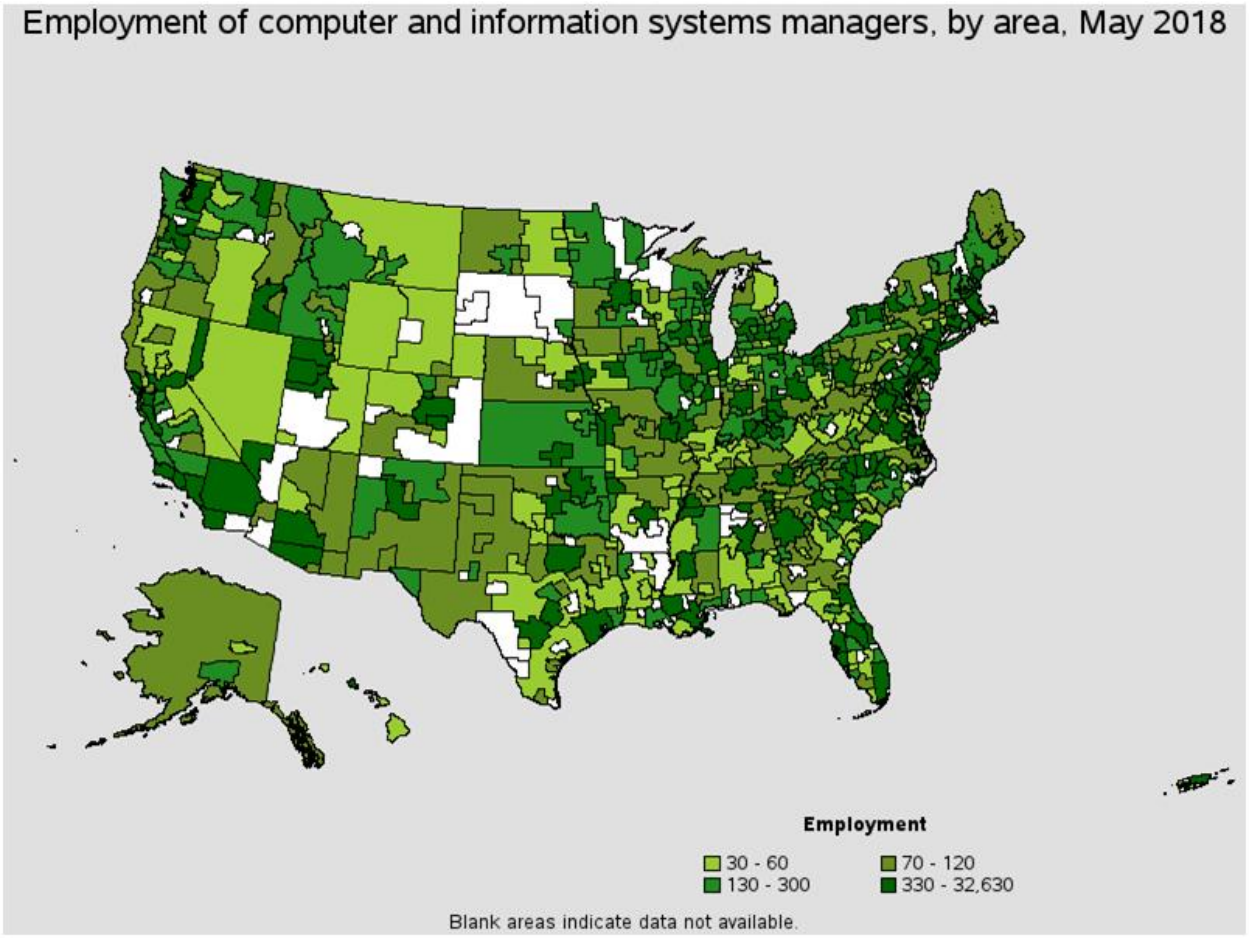
1.2.5States with the highest concentration of jobs and location quotients in this occupation:

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
District of Columbia	4,020	5.65	2.09	$81.01	$168,500
Connecticut	8,430	5.08	1.88	$70.89	$147,440
Massachusetts	17,820	4.99	1.85	$75.30	$156,620
Washington	13,690	4.20	1.55	$76.58	$159,290
Maryland	10,590	3.94	1.46	$74.46	$154,870



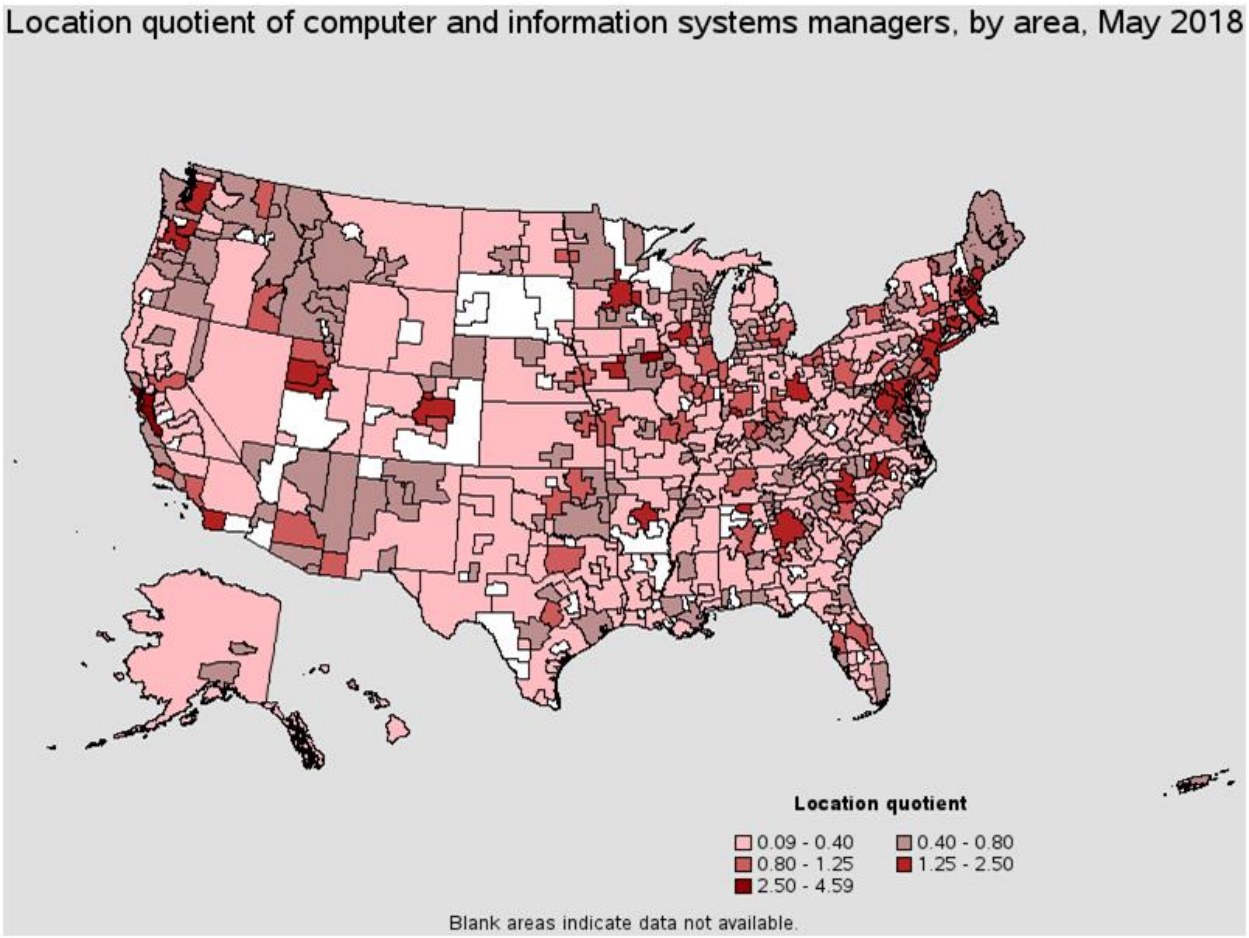
1.2.6Top paying States for this occupation:

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
New York	27,070	2.88	1.07	$91.50	$190,310
California	63,640	3.74	1.38	$86.66	$180,250
New Jersey	14,280	3.52	1.30	$84.95	$176,690
Virginia	13,450	3.51	1.30	$83.31	$173,290
Delaware	1,240	2.77	1.02	$81.60	$169,720


Employment of computer and information systems managers, by area, May 2018
Employment
30 - 60
70 - 120
130 - 300
330 - 32,630
Blank areas indicate data not available.

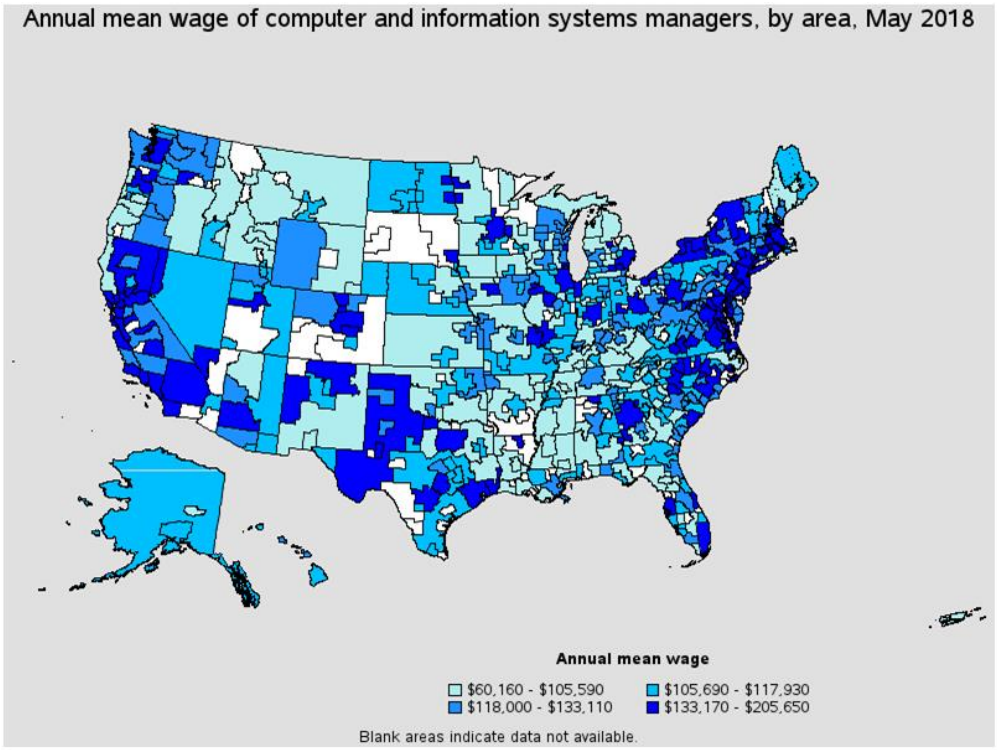
1.2.7Metropolitan areas with the highest employment level in this occupation:

Metropolitan area	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
New York-Newark-Jersey City, NY-NJ-PA	32,630	3.44	1.27	$94.09	$195,710
Washington-Arlington-Alexandria, DC-VA-MD-WV	18,220	5.81	2.15	$83.40	$173,470
Los Angeles-Long Beach-Anaheim, CA	17,990	2.94	1.09	$80.74	$167,940
San Francisco-Oakland-Hayward, CA	17,590	7.27	2.69	$93.39	$194,250
Boston-Cambridge-Nashua, MA-NH	16,530	5.99	2.22	$77.05	$160,260
Chicago-Naperville-Elgin, IL-IN-WI	15,320	3.30	1.22	$70.22	$146,050
San Jose-Sunnyvale-Santa Clara, CA	13,880	12.40	4.59	$98.87	$205,650
Atlanta-Sandy Springs-Roswell, GA	12,010	4.48	1.66	$70.75	$147,150
Seattle-Tacoma-Bellevue, WA	11,700	5.90	2.18	$79.18	$164,700
Dallas-Fort Worth-Arlington, TX	9,010	2.52	0.93	$76.81	$159,770


Location quotient of computer and information systems managers, by area, May 2018
Location quotient
0.09 - 0.40
0.40 - 0.80
0.80 - 1.25
1.25 - 2.50
2.50 - 4.59
Blank areas indicate data not available.

1.2.8Metropolitan areas with the highest concentration of jobs and location quotients in this occupation:

Metropolitan area	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
San Jose-Sunnyvale-Santa Clara, CA	13,880	12.40	4.59	$98.87	$205,650
Cedar Rapids, IA	1,080	7.67	2.84	$63.11	$131,280
San Francisco-Oakland-Hayward, CA	17,590	7.27	2.69	$93.39	$194,250
Hartford-West Hartford-East Hartford, CT	3,590	6.14	2.27	$68.12	$141,680
Boston-Cambridge-Nashua, MA-NH	16,530	5.99	2.22	$77.05	$160,260
California-Lexington Park, MD	280	5.92	2.19	$71.40	$148,520
Seattle-Tacoma-Bellevue, WA	11,700	5.90	2.18	$79.18	$164,700
Washington-Arlington-Alexandria, DC-VA-MD-WV	18,220	5.81	2.15	$83.40	$173,470
Boulder, CO	970	5.32	1.97	$77.49	$161,190
Durham-Chapel Hill, NC	1,600	5.20	1.92	$72.39	$150,570


Annual mean wage of computer and information systems managers, by area, May 2018
Annual mean wage
$60,160 - $105,590
$105,690 - $117,930
$118,000 - $133,110
$133,170 - $205,650
Blank areas indicate data not available.

1.2.9Top paying metropolitan areas for this occupation:

Metropolitan area	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
San Jose-Sunnyvale-Santa Clara, CA	13,880	12.40	4.59	$98.87	$205,650
New York-Newark-Jersey City, NY-NJ-PA	32,630	3.44	1.27	$94.09	$195,710
San Francisco-Oakland-Hayward, CA	17,590	7.27	2.69	$93.39	$194,250
Washington-Arlington-Alexandria, DC-VA-MD-WV	18,220	5.81	2.15	$83.40	$173,470
Bridgeport-Stamford-Norwalk, CT	(8)	(8)	(8)	$83.19	$173,040
Santa Maria-Santa Barbara, CA	510	2.75	1.02	$81.53	$169,590
Denver-Aurora-Lakewood, CO	5,350	3.62	1.34	$80.96	$168,400
Los Angeles-Long Beach-Anaheim, CA	17,990	2.94	1.09	$80.74	$167,940
Cumberland, MD-WV	60	1.65	0.61	$79.99	$166,370
Seattle-Tacoma-Bellevue, WA	11,700	5.90	2.18	$79.18	$164,700

1.2.10Nonmetropolitan areas with the highest employment in this occupation:

Nonmetropolitan area	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
Kansas nonmetropolitan area	300	0.76	0.28	$46.36	$96,440
Southwest Montana nonmetropolitan area	280	2.02	0.75	$50.36	$104,750
Southeast Iowa nonmetropolitan area	280	1.21	0.45	$59.10	$122,930
Southwest Maine nonmetropolitan area	270	1.42	0.52	$49.54	$103,030
West Central-Southwest New Hampshire nonmetropolitan area	260	2.57	0.95	$62.44	$129,870

1.2.11Nonmetropolitan areas with the highest concentration of jobs and location quotients in this occupation:

Nonmetropolitan area	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
West Central-Southwest New Hampshire nonmetropolitan area	260	2.57	0.95	$62.44	$129,870
Central New Hampshire nonmetropolitan area	220	2.24	0.83	$53.73	$111,760
Northern New Mexico nonmetropolitan area	160	2.05	0.76	$67.68	$140,780
Southwest Montana nonmetropolitan area	280	2.02	0.75	$50.36	$104,750
Northern Vermont nonmetropolitan area	140	1.99	0.74	$55.70	$115,850

1.2.12Top paying nonmetropolitan areas for this occupation:

Nonmetropolitan area	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
Southwest New York nonmetropolitan area	130	0.72	0.27	$70.56	$146,770
North Valley-Northern Mountains Region of California nonmetropolitan area	60	0.61	0.23	$68.85	$143,210
Northern New Mexico nonmetropolitan area	160	2.05	0.76	$67.68	$140,780
Capital/Northern New York nonmetropolitan area	80	0.57	0.21	$66.65	$138,640
Maryland nonmetropolitan area	110	1.75	0.65	$65.32	$135,860

1.2.13About May 2018 National, State, Metropolitan, and Nonmetropolitan Area Occupational Employment and Wage Estimates

These estimates are calculated with data collected from employers in all industry sectors, all metropolitan and nonmetropolitan areas, and all states and the District of Columbia. The top employment and wage figures are provided above.

The percentile wage estimate is the value of a wage below which a certain percent of workers fall. The median wage is the 50th percentile wage estimate--50 percent of workers earn less than the median and 50 percent of workers earn more than the median.

(1) Estimates for detailed occupations do not sum to the totals because the totals include occupations not shown separately. Estimates do not include self-employed workers.

(2) Annual wages have been calculated by multiplying the hourly mean wage by a "year-round, full-time" hours figure of 2,080 hours; for those occupations where there is not an hourly wage published, the annual wage has been directly calculated from the reported survey data.

(3) The relative standard error (RSE) is a measure of the reliability of a survey statistic. The smaller the relative standard error, the more precise the estimate.

(5) This wage is equal to or greater than $100.00 per hour or $208,000 per year.

(8) Estimate not released.

(9) The location quotient is the ratio of the area concentration of occupational employment to the national average concentration. A location quotient greater than one indicates the occupation has a higher share of employment than average, and a location quotient less than one indicates the occupation is less prevalent in the area than average.

1.3Same occupation, different pay: How wages vary?

Management

Managers with different job tasks and qualifications often are grouped together in the same occupation. This variation might be one of the reasons for the big wage differences in the occupations shown in table 3.

Advertising and promotions managers, for example, may have vastly different pay depending on the types of accounts they handle. A highly paid worker in this occupation might coordinate all of the advertising for a large, multinational corporation, and another worker might make much less overseeing the classified ads department of a local newspaper. As a result, the wage difference—more than $142,140—for this relatively small and competitive occupation was among the highest of any type of manager.

Table 3. Management occupations with more than $100,000 wage difference, May 2014

Occupation	Employment	Median wage	10th percentile wage	90th percentile wage(1)	Wage difference(2)
Advertising and promotions managers	29,340	$96,720	$45,060	>$187,200	>$142,140
General and operations managers	2,049,870	97,270	45,130	>187,200	>142,070
Sales managers	358,920	110,660	53,620	>187,200	>133,580

Table 3. Management occupations with more than $100,000 wage difference, May 2014

Occupation	Employment	Median wage	10th percentile wage	90th percentile wage(1)	Wage difference(2)
Public relations and fundraising managers	56,920	101,510	55,420	>187,200	>131,780
Compensation and benefits managers	16,380	108,070	58,370	>187,200	>128,830
Financial managers	518,030	115,320	62,480	>187,200	>124,720
Postsecondary education administrators	131,070	88,390	50,240	174,000	123,760
Human resources managers	116,610	102,780	60,440	183,590	123,150
Marketing managers	184,490	127,130	65,980	>187,200	>121,220
Training and development managers	29,870	101,930	57,920	178,360	120,440
Natural sciences managers	53,290	120,050	70,020	>187,200	>117,180
Chief executives	246,240	173,320	72,750	>187,200	>114,450
Computer and information systems managers	330,360	127,640	78,470	>187,200	>108,730
Purchasing managers	70,840	106,090	60,840	169,000	108,160
Medical and health services managers	310,320	92,810	55,890	161,150	105,260
Architectural and	179,320	130,620	83,580	>187,200	>103,620

Table 3. Management occupations with more than $100,000 wage difference, May 2014

Occupation	Employment	Median wage	10th percentile wage	90th percentile wage(1)	Wage difference(2)
engineering managers					
Administrative services managers	268,730	83,790	45,590	149,180	103,590
Industrial production managers	167,200	92,470	56,290	158,170	101,880

Footnotes:

(1) BLS does not publish specific estimates for percentile wages above $187,200 per year. Where the percentile wage is greater than $187,200, the wage is shown with a greater-than sign (>).

(2) Wage differences with a greater-than sign (>) were calculated using $187,200, the highest percentile wage that BLS publishes.

Source: Occupational Employment Statistics survey, BLS

1.4The No. 1 highest paid, most in-demand job in every U.S. state

High growth, highly paid jobs

Careers ranked by projected growth and salary

STATE	JOB TITLE	2018-2023 CHANGE
Alabama	Software developers	17%
Alaska	Registered nurses	16%
Arizona	Nurse practitioners	25%
Arkansas	Physical therapists	21%
California	Occupational therapy assistants	29%
Colorado	Operations research analysts	27%
Connecticut	Nurse practitioners	16%
Delaware	Financial sales agents	21%
D.C.	Software developers	26%
Florida	Medical scientists*	25%
Georgia	Film and video editors	49%
Hawaii	Physical therapists	12%
Idaho	Market researchers	22%
Illinois	Physician assistants	17%
Indiana	Physician assistants	23%
Iowa	Physician assistants	18%
Kansas	Software developers	19%
Kentucky	Physician assistants	16%
Louisiana	Software developers	29%
Maine	Software developers	15%
Maryland	Flight attendants	19%
Massachusetts	Computer scientists	18%
Michigan	Occupational therapy assistants	19%
Minnesota	Physician assistants	20%
Mississippi	Software developers	23%
Missouri	Software developers	19%
Montana	Software developers	24%
Nebraska	Physical therapists	20%
Nevada	Software developers	25%
New Hampshire	Software developers	18%
New Jersey	Physical therapist assistants	28%
New Mexico	Software developers	21%
New York	Physical therapist assistants	27%
North Carolina	Software developers	19%
North Dakota	Software developers	17%
Ohio	Physician assistants	17%
Oklahoma	Rotary drill operators	17%
Oregon	Postsecondary teachers	24%
Pennsylvania	Nurse practitioners	17%
Rhode Island	Software developers	14%
South Carolina	Software developers	26%
South Dakota	Software developers	19%
Tennessee	Software developers	25%
Texas	Nurse practitioners	25%
Utah	Producers and directors	28%
Vermont	Software developers	19%
Virginia	Physician assistants	19%
Washington	Nurse practitioners	22%
West Virginia	Software developers	22%
Wisconsin	Physician assistants	17%
Wyoming	Business operations specialists	7%

*Except epidemiologists

Source: CareerBuilder

These are the occupations with the highest median wages reported for 2018. You can view this list for any state by changing the location shown to the left.

We found **807** occupations in **United States**.

Occupation	2018 Median wages		Typical Education
	Hourly	Annual	
Prosthodontists	$84.88	$176,600	Doctoral or professional degree
Pediatricians, General	$82.00	$170,600	Doctoral or professional degree
Nurse Anesthetists	$80.75	$168,000	Master's degree
Dentists, General	$73.00	$151,800	Doctoral or professional degree
Dentists, All Other Specialists	$70.66	$147,000	Doctoral or professional degree
Computer and Information Systems Managers	$68.53	$142,500	Bachelor's degree
Architectural and Engineering Managers	$67.67	$140,800	Bachelor's degree
Airline Pilots, Copilots, and Flight Engineers	N/A	$140,300	Bachelor's degree
Petroleum Engineers	$65.95	$137,200	Bachelor's degree
Marketing Managers	$64.56	$134,300	Bachelor's degree

1.5How much is a Data Scientist's Salary?

The average data scientist salary is $113,436, according to Glassdoor. The driving factor behind high data science salaries is that organizations are realizing the power of big data and want to use it to drive smart business decisions. And because the supply of data professionals hasn't yet caught up with demand, starting salaries for these positions remain high ($50,000 to $95,000), especially for those who have an advanced degree in data science or a related field.

1.5.1Data Scientist Salary by Experience

According to O'Reilly's 2016 Data Science Salary Survey, experience is one of the most important factors in a data scientist's salary. For every year of experience, data science professionals make an average of $2,000 to $2,500 more.

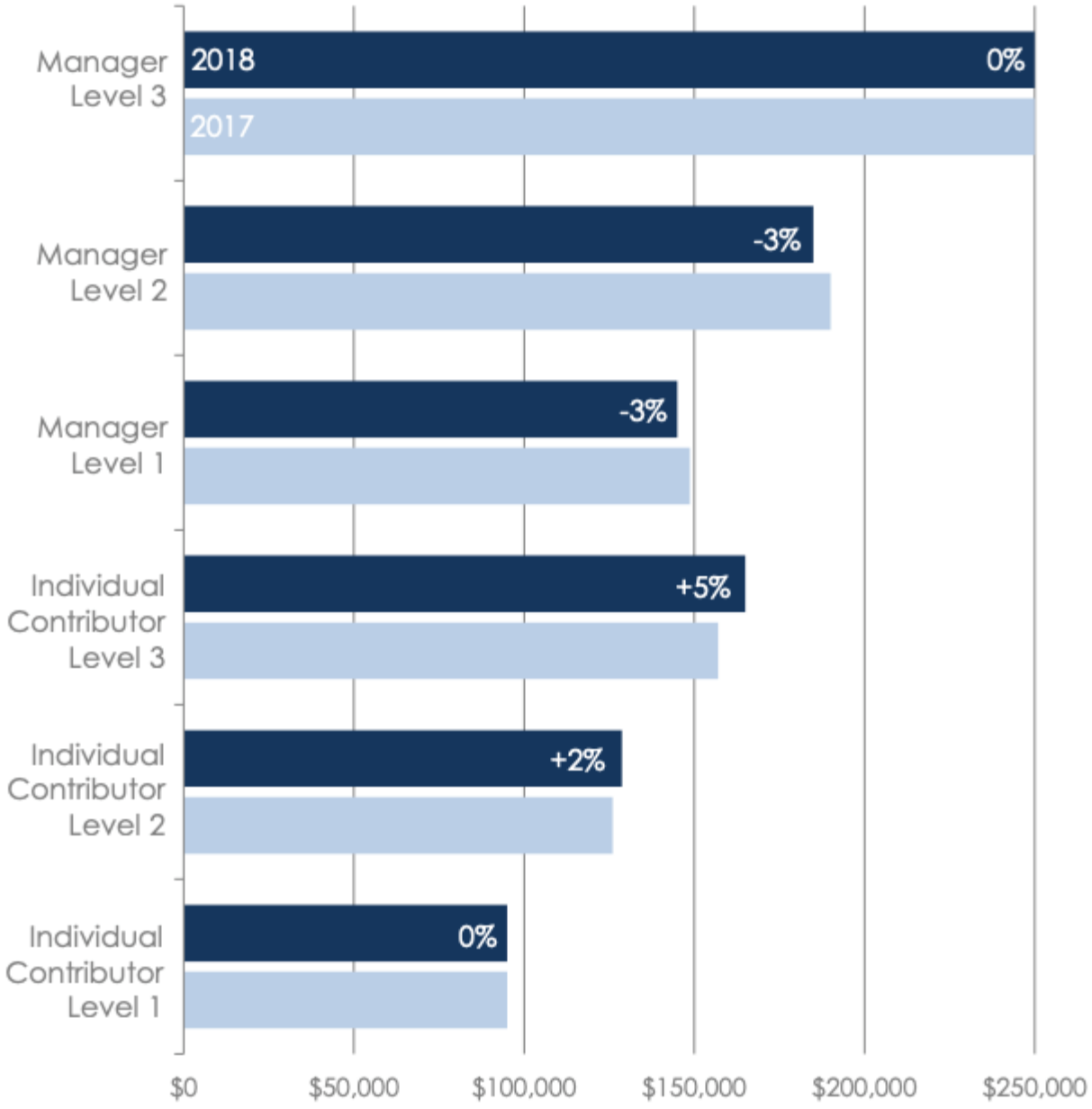
A 2018 Burtch Works study of data science salaries reported the latest salary trends based on experience:

1.5.2Entry-level data scientist salary. Despite a recent influx of early-career professionals, the median starting salary for a data scientist remains high at $95,000.

1.5.3Mid-level data scientist salary. The median salary for a mid-level data scientist is $128,750. If this data scientist is also in a managerial role, the median salary rises to $185,000.

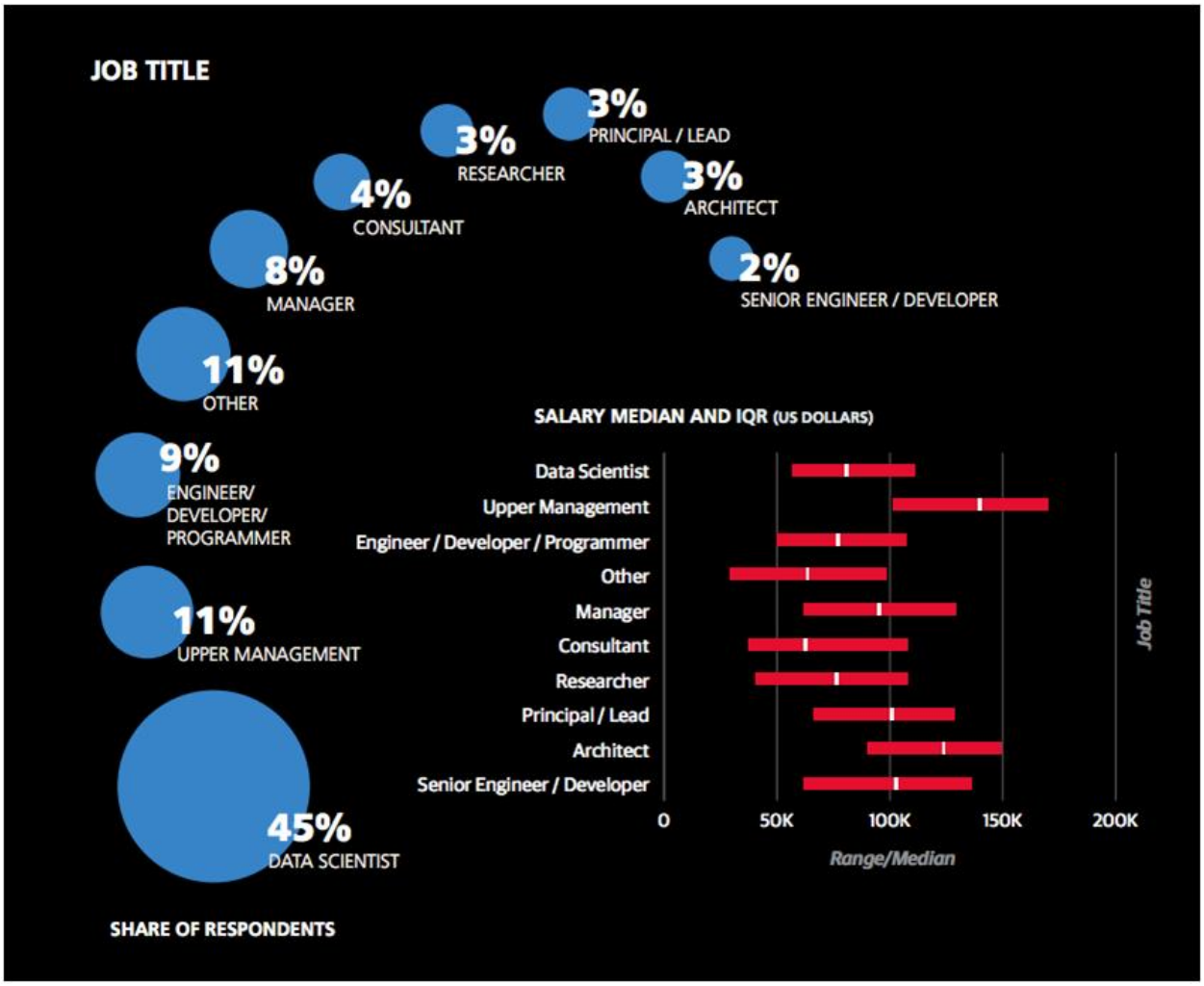
1.5.4Experienced data scientist salary. The median salary for experienced data science professionals is $165,000—while the median salary for experienced manager-level professionals is considerably higher at $250,000.

Figure 1 Comparison of Data Scientists' Median Base Salaries by Job Category



1.5.5Data Scientist Salary by Job Title

In O'Reilly's data science salary report, 45 percent of those surveyed said they hold the title of "data scientist." Another 31 percent said they are upper management; an engineer, developer, or programmer; or other. In general, the more a data science professional engages in managerial tasks—such as leading team projects, identifying business problems to be solved with analytics, or communicating with external parties—the higher the salary.



1.5.6Data scientist

Those with the job title of “data scientist” are generally experienced, expert-level professionals in data-driven organizations, according to
Salary range: $85,000-$170,000

1.5.7Data analyst

Data analysts work hands-on with data and tend to be at a point in their careers when they are focused on building up data science tools and skill sets.
Entry-level salary: $50,000-$75,000
Experienced salary: $65,000-$110,000

1.5.8Data science/analytics manager

These professionals have one to three direct reports, sharp technical and quantitative skills, and strong leadership and business aptitude.
Salary range: $90,000-$140,000

1.5.9Big data engineer

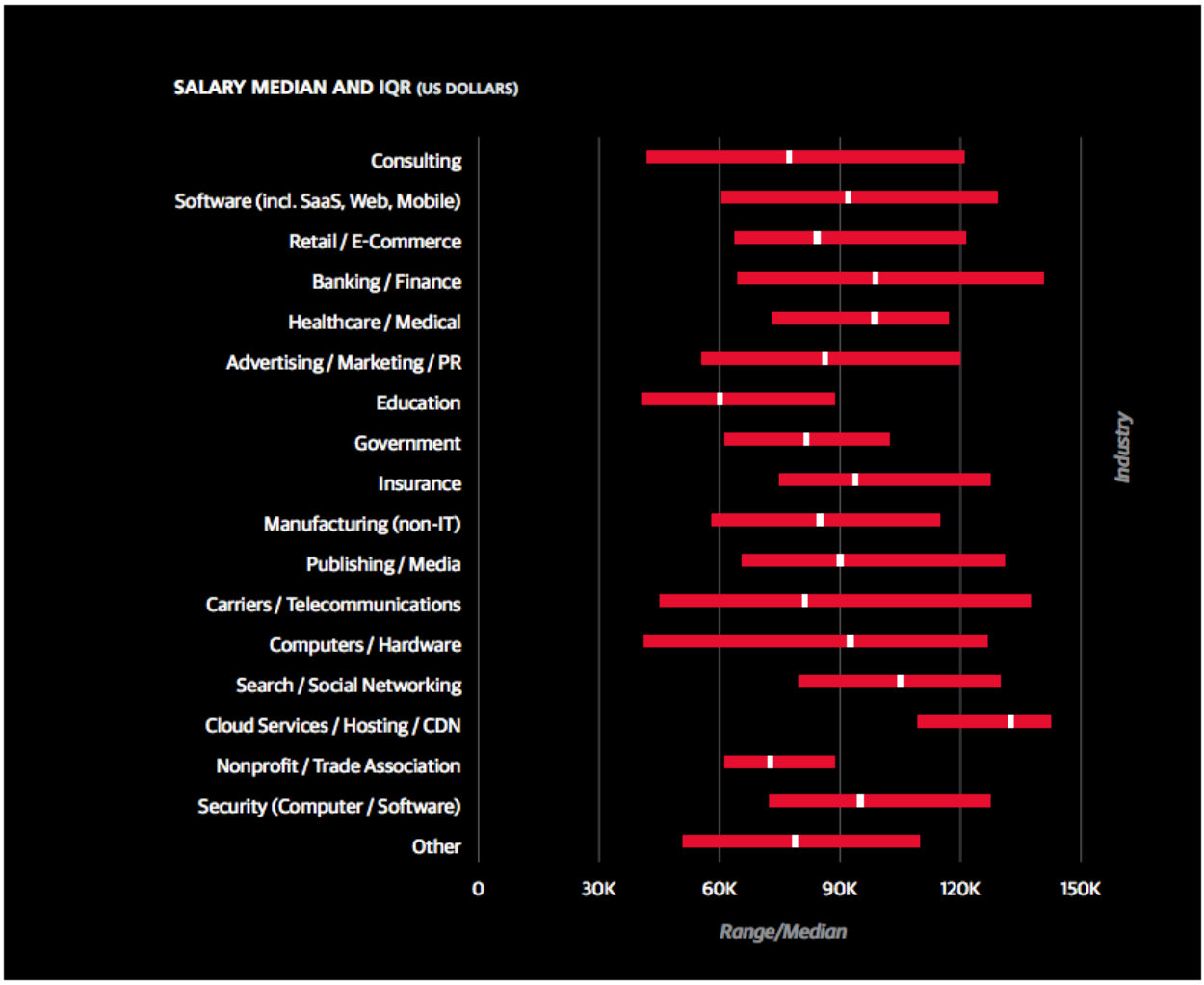
Engineers can solve problems and drive business value by building the platforms and applications that data scientists use to perform data analytics.
Junior/generalist salary: $70,000-$115,000
Domain expert salary: $100,000-$165,000

Data Scientist Salary by Industry and Company Size

The industries with the highest median data science salaries are:

1. Cloud services, hosting, and CDN
2. Search and social networking
3. Banking and finance

It’s important to note that only 12 percent of data scientists surveyed by O’Reilly in 2016 worked in these three industries. Nearly 30 percent worked in either consulting or software (SAAS, web, and mobile).

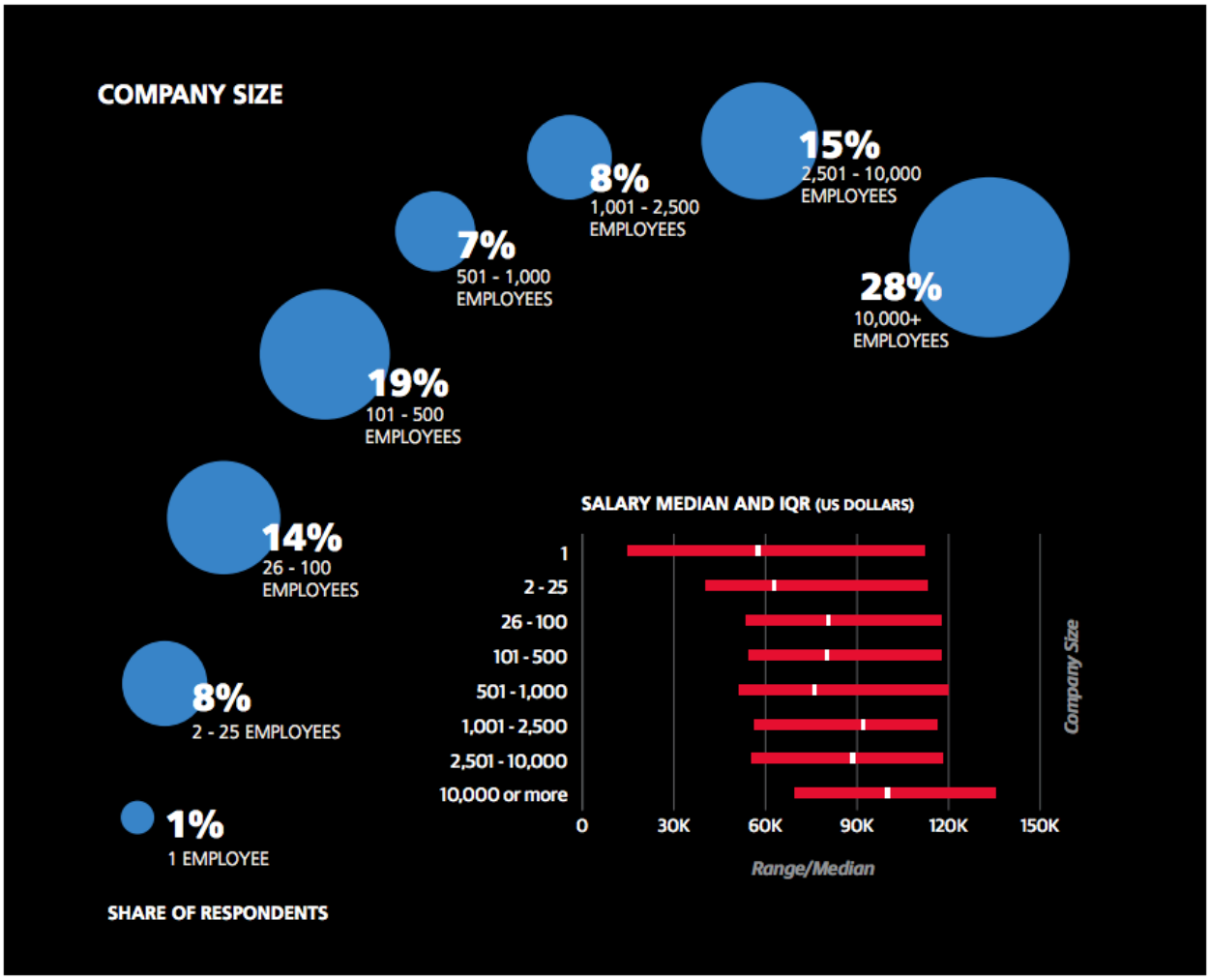


Perhaps not surprisingly, some of the highest paid data scientists work at leading tech companies. Here are average salaries at several high-profile organizations:

Google: $152,856
Apple: $145,974
Twitter: $135,360
Facebook: $134,715
PayPal: $132,909
Airbnb: $127,852
Microsoft: $123,328

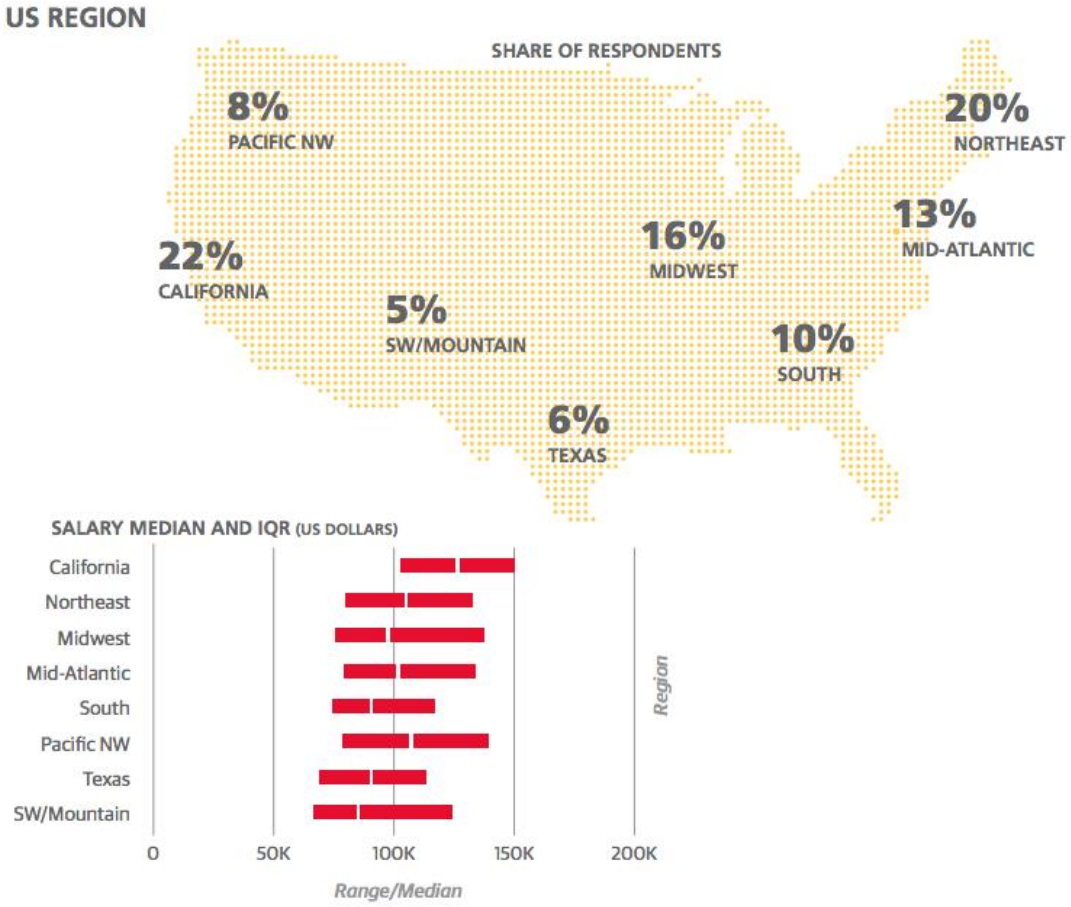
As far as company size, the larger the organization, the larger the salary. For example, in a company with at least 10,000 employees, a data scientist would

likely earn a higher income than the same role at a company with less than 1,000 employees.



1.5.10Data Scientist Salary by Region

Data scientists' salaries depend greatly on the region in which they live. The highest salaries are in California, which is also where the majority of data scientists surveyed by O'Reilly work. The Pacific Northwest has fewer data scientists than most regions of the country, but it boasts the second highest salaries.



1.5.11Data Scientist Salary by Education

Data scientists are rare, especially when it comes to having the right combination of education and skills. At the most basic level, data scientists must know how to write code. In fact, the highest data scientist salaries belong to those who code four to eight hours per week; the lowest salaries belong to those who don't code at all.

Skills That Affect Data Scientist, IT Salaries	
Apache Hadoop	▲8%
Hadoop	▲7%
Python	▲5%
Machine Learning	▲4%
Big Data Analytics	▲3%
Data Mining / Data Warehouse	▲2%
SAS	▲2%
Data Analysis	▲1%
Data Modeling	▲1%
SQL	▼0%
National Average	**$92,000**

Source: Payscale

Data scientists also need to wield the technologies that tame big data—and be able to learn the new and emerging ones—as well as open-source tools, cloud computing, and data visualization. It's not enough to simply know how to use these tools, however; it's crucial that data scientists know how to use them to derive actionable insights that will improve their organization.

Additionally, data science positions require sharp business acumen, scientific curiosity, and strong leadership and communication skills. Having them will greatly affect salary.

Data scientists tend to have advanced degrees in quantitative disciplines such as data science, applied mathematics, statistics, computer science, engineering, economics, or operations research. But only a data science master's degree will give you the precise education you need to be ready for a career in data science.

1.6Data Science Jobs with their salaries

Data Science jobs are the highest paid in the IT industry. In the past years, a barrage of data science roles came our way, so it is difficult to get a basic understanding of what they are and how they differ. Even when these roles are named differently, it doesn't help.

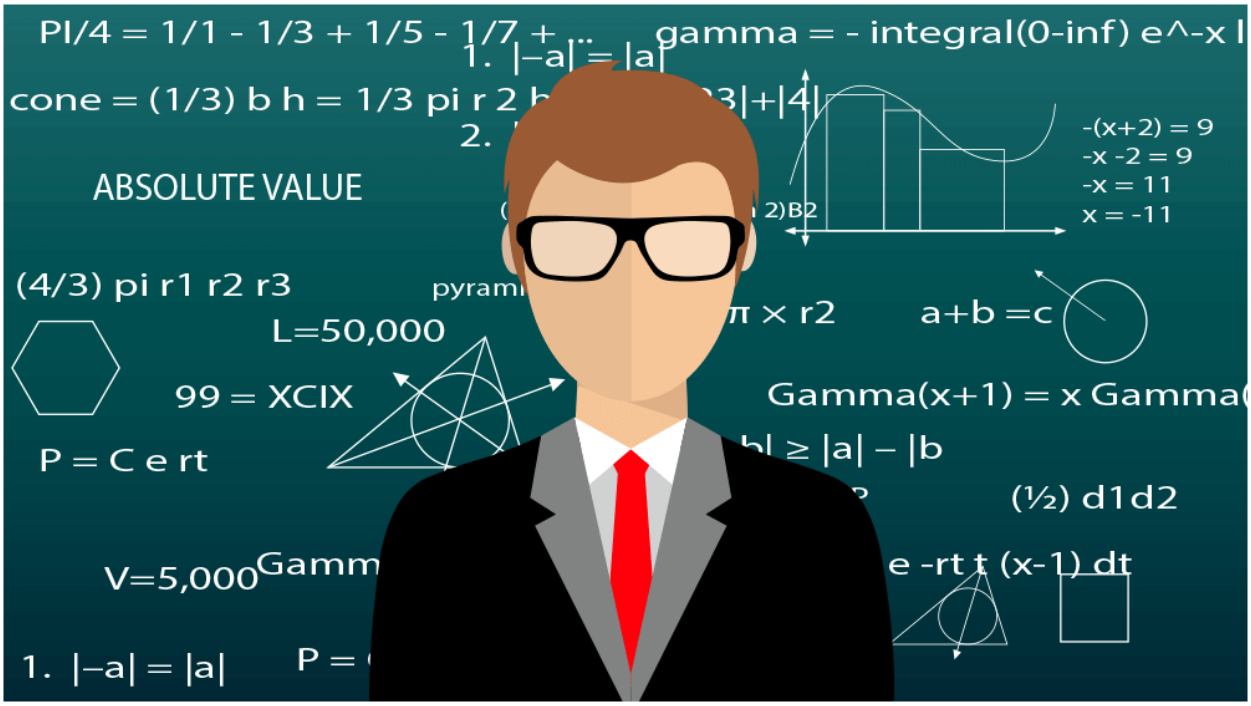
There is a considerable need for Data Scientist. Demand for Data Scientist is on the peak and they are among the highest paid people with the median base salary of **$130,000 in America itself. The predictions by** McKinsey states that in the coming years there is going to be a **50% gap** in the supply vs demand of data science professionals.You can also **master Data Science** and make Sense out of Data.

1.7Various job roles in data science:

There are a plethora of data science roles that can be difficult to understand. **The skills required for each job role differ from one another**.The data science field uses Mathematics. Statistics and Computer science disciplines and also a common tool-set that includes Python, R, Tableau, SQL, etc. The data science professionals are in huge demand today and below, we will discuss the different **job titles** offered in data science:

1.7.1. Data Scientist — A Data scientist is probably one of the hottest job titles and the most sought jobs in this field you can have in today's world. They are responsible for the handling of raw data, analyzing it with the help of various techniques as discussed above and presenting insights in a manner that are useful for predicting business problems. A Data Scientist uses **Machine Learning** and also predicts the future based on past patterns.

The average salary range (US) for a Data scientist is **$119,000.**



1.7.2 Data Analyst — Data analyst is the one who analyses data. But this process requires creating systems that help users of business to draw out insights and ensure data quality. His role is to collect, process, and perform statistical data analyses. Data Analyst finds meaningful information from available data and uses R or SAS. Not just IT industries, but all kinds of companies in the industries i.e. healthcare, automobile, finance, retail, and insurance need Data Analysts to run their business.

The average annual salary (US) for Data analysts is **$62,000.**

1.7.3 Data Architect — With the rise of big data, the role of Data Architect is increasingly gaining importance. His role is to create data management systems to integrate, protect and maintain data sources and company's information. He is responsible for database architecture, design, creation and optimization of data. He relies heavily on Microsoft Excel to do attribute mappings and to design tables. Technologies like Pig, Spark, SQL, XML, and Hive are required to be mastered by data architects.

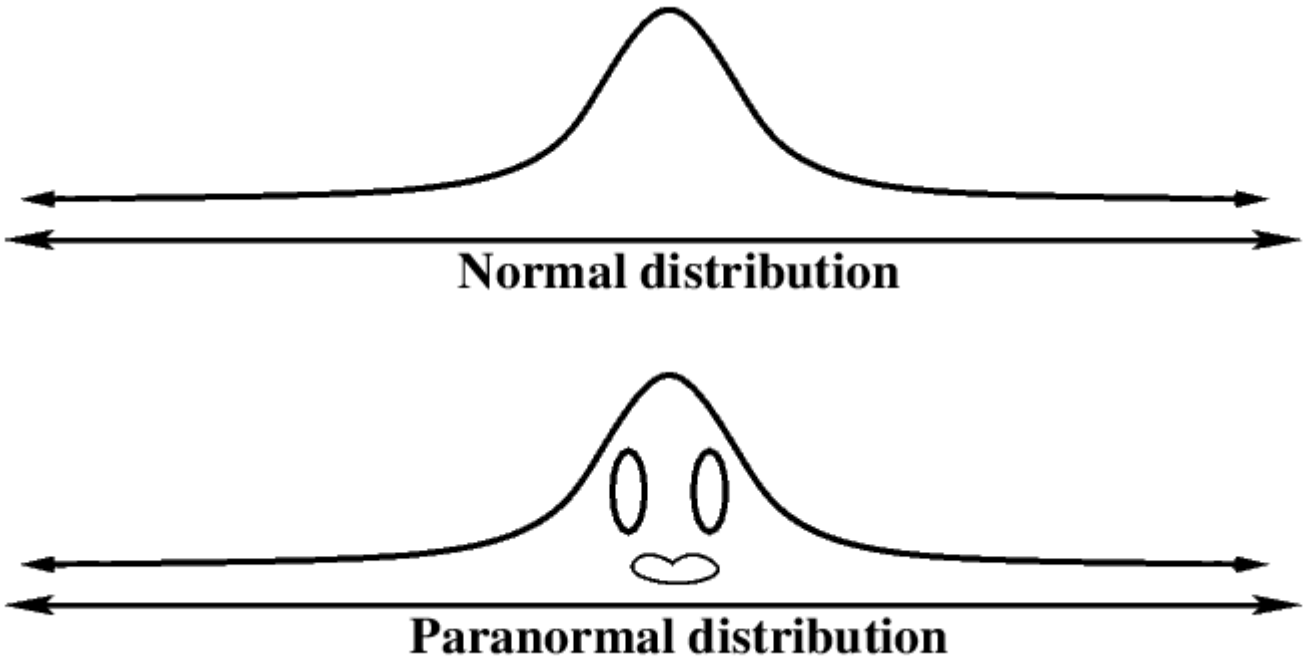
The average annual salary (US) for this career is **$100,000.**

1.7.4 Data Engineer — They are not the ones who analyze data but builds a certain software infrastructure for other professionals to do the work. They are able to do this as they have an in-depth knowledge of Hadoop and **Big Data technologies** such as MapReduce, Hive, and Pig, NoSQL technologies, SQL technologies. His role is to develop, test and maintain large scale processing systems. More than 50 percent of the work is Data Wrangling, where data engineers excel who has a background in software engineering.

The average salary (US) for this job is **$95,000.**

1.7.5 Statistician — Role of a statistician is to understand the changing market trends that affect the growth of a business. The work is to gather the data and convert it into useful information. Their skills provide them with the ability to handle all kinds of data, whether structured or unstructured. The common tools and techniques used by them comprise of SAS, SPSS, Matlab, Python, Scala, SQL, R, etc. They have a strong background of statistics which covers descriptive and inferential topics.

The average salary (US) for Statisticians is **$75,000.**



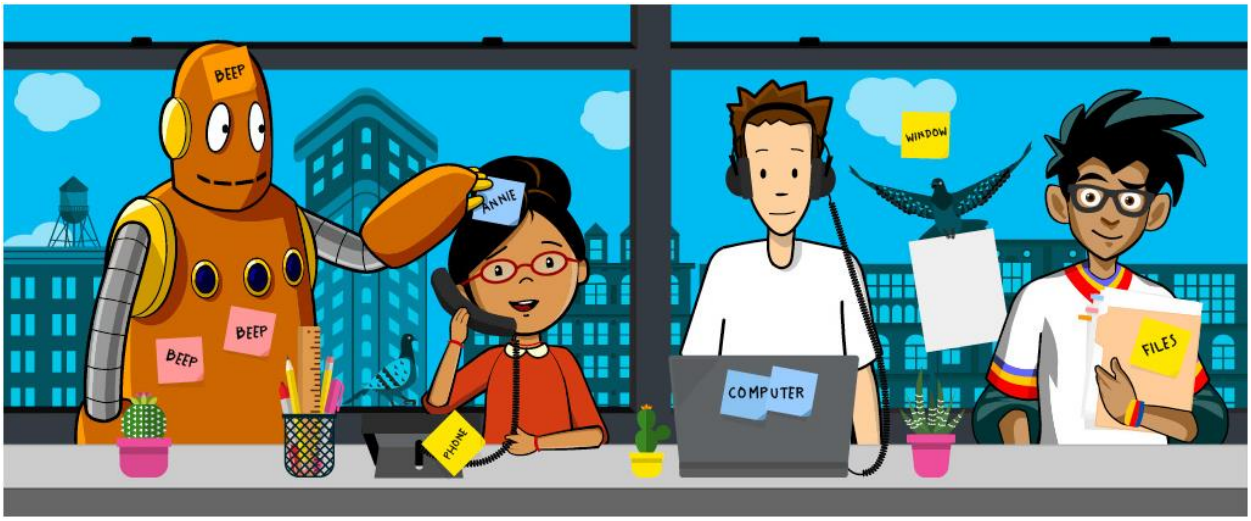
Source:Google

1.7.6 Database Administrator — These are trained professionals who are responsible for storing and organizing the company's data. All the activities related to protecting, monitoring, maintaining and securing of data is performed and directed by Database Administrators. This helps in making databases available to users including employees at all times. The various tools used for this purpose are MySQL Workbench, Toad, Adminer, DatabaseSpy, etc.

The typical average salary (US) for this position is **$68,000.**

1.7.7 Business Analyst — The role of a business analyst is similar to that of a data analyst. However, business analysts have specialized knowledge of the business domain. They use their skills to guide the business in improving its products and services. They gather and analyze the basic business needs and requirements and are responsible to bring and manage change in the business. Tools used by a business analyst are Microsoft Office which includes PowerPoint, Excel, Word and Visio, Google Docs, Sheets and Slides, SmartDraw, Balsamiq, etc.

The average salary (US) for this data science job is **$65,000.**



1.7.8. Data & Analytics Manager — Data & analytics managers are responsible for developing processes for effective data analysis, interpretation, and reporting. They provide directions to the team of data analysts and are also responsible for making hiring decisions. They need to have 7–10 years of experience in the field of data science. They can advise organizational leaders on making changes according to the market trends. They must have good knowledge of Computer Programming Languages, Database Systems, **Big Data tools**, and Machine Learning.

This career has an average salary (US) of **$116,000.**

"I'm done! I'm so bad at playing cards, I even lost money playing Planning Poker."

1.8Closing Remarks:

While looking for your data science dream job, take into account that the roles differ and that is why you need to make sure you know each role in detail and then opt whichever you think fits right for you and your skills. If one chooses to make his **career in the field of Data science,** he knows that he will enjoy his job with good pay and

demand. The career opportunities are expected to grow for a long time as Data permeates our life and companies try to understand the data generated.

1.8.1Salary for Skill: Big Data Analytics

$108k of USA

Avg. Salary **Show Hourly Rate**

*These Jobs Average are from $46K to $177K

Data Scientist

$98k

Analytics Manager

$97k

Director of Analytics

$136k

Data Engineer

$96k

Senior Data Scientist

$129k

Software Engineer

$104k

Data Analyst

$66k

1.9Is it still worth becoming a data scientist

With salaries flattening and competition rising, there are signs the prospects for data scientists may be less stellar than once thought.

.

Figures produced by Glassdoor Economic Research show a year-on-year fall in US data scientist wages in February and March of this year."On Glassdoor, we've seen pay for data scientists actually shrink 1.2 percent in March 2019," said Glassdoor senior economist Daniel Zhao.
"This is a continuation of a longer running trend--data scientist wage growth has been well below the national average for the last year."

1.9.1ADVERTISING

Glassdoor is not alone in noticing the trend, with a similar tailing off of salaries evident in data collected by Stack Overflow over the past year.

"Data scientist salaries are moving closer to the mainstream of software developer salaries in general," said Stack Overflow data scientist Julia Silge, adding there was "much

less of a difference" between the pay of the two groups when controlling for education level.

As for the reason for the salary squeeze, for Glassdoor's Zhao, it's clear that there are now more candidates for data scientist roles than there are jobs available.

"As data science has risen in prominence, enrolments in data science programs and bootcamps have exploded. This huge increase in workers for limited entry-level jobs is holding down wages," he said.

While analyst reports often discuss the sharp uptick in demand for data science skills, anecdotal evidence from those in the industry suggests that demand may be being outstripped by the large numbers of new entrants to the field, thanks to the explosion in the number of data science courses offered by online learning hubs like Fast.ai and Coursera.
Vicky Boykis, senior manager for data science and engineering at CapTech Ventures, wrote that she and others she knows in the industry have seen more than **a fivefold increase in the numbers applying for junior data science roles.**"It can be very hard for someone with a new degree in data science to find a data science position, given how many new people they're competing with in the market," she wrote.
But how can suggestions of there being an oversupply of data scientists be reconciled with frequent **reports of a data science skills shortage**? Zhao says it's important to understand that while businesses may be struggling to find

the skills they need, that doesn't mean there's not enough entry-level talent.
"There might be a skills shortage, but not an applicant shortage. It's not unusual for entry-level or internship openings in data science to receive hundreds of applicants. When employers talk about shortages, they're generally talking about a lack of experienced professionals," he said, adding this largely stemmed from the newness of data science as a mainstream field.

One confounding factor to bear in mind, however, is that comparing salary figures for data scientists over time is made difficult by how poorly defined the data scientist role is.

"Companies are increasingly using the data scientist title for other similar roles such as data analyst or statistician," said Zhao.

"This muddling of job titles is changing the composition of the data scientist workforce and holding down wages as a result."

1.10 Should you think twice before training as a data scientist?

With slowing salary growth among data scientists and signs there may be a glut of junior talent, should aspiring data scientists pause for thought?

Boykis' advice is **to consider getting into the field by the "back door"**, by starting out in a tangentially related field like a junior developer or data analyst and working your way towards becoming a data scientist, rather than aiming straight for data scientist as a career.
Stack Overflow's Silge has a slightly different interpretation of why salaries are levelling out and believes people shouldn't necessarily be deterred from entering the industry.

"I think that what we're seeing is a little bit of the standardization and the professionalization of data science," she said.

"The past ten years have been a bit of the Wild West when it comes to data science. 'How do you become a data scientist?', it's been a really open question.

"I see the industry moving towards some consensus around 'What does it mean to be a data engineer? and 'What does it mean to be a data scientist?'.

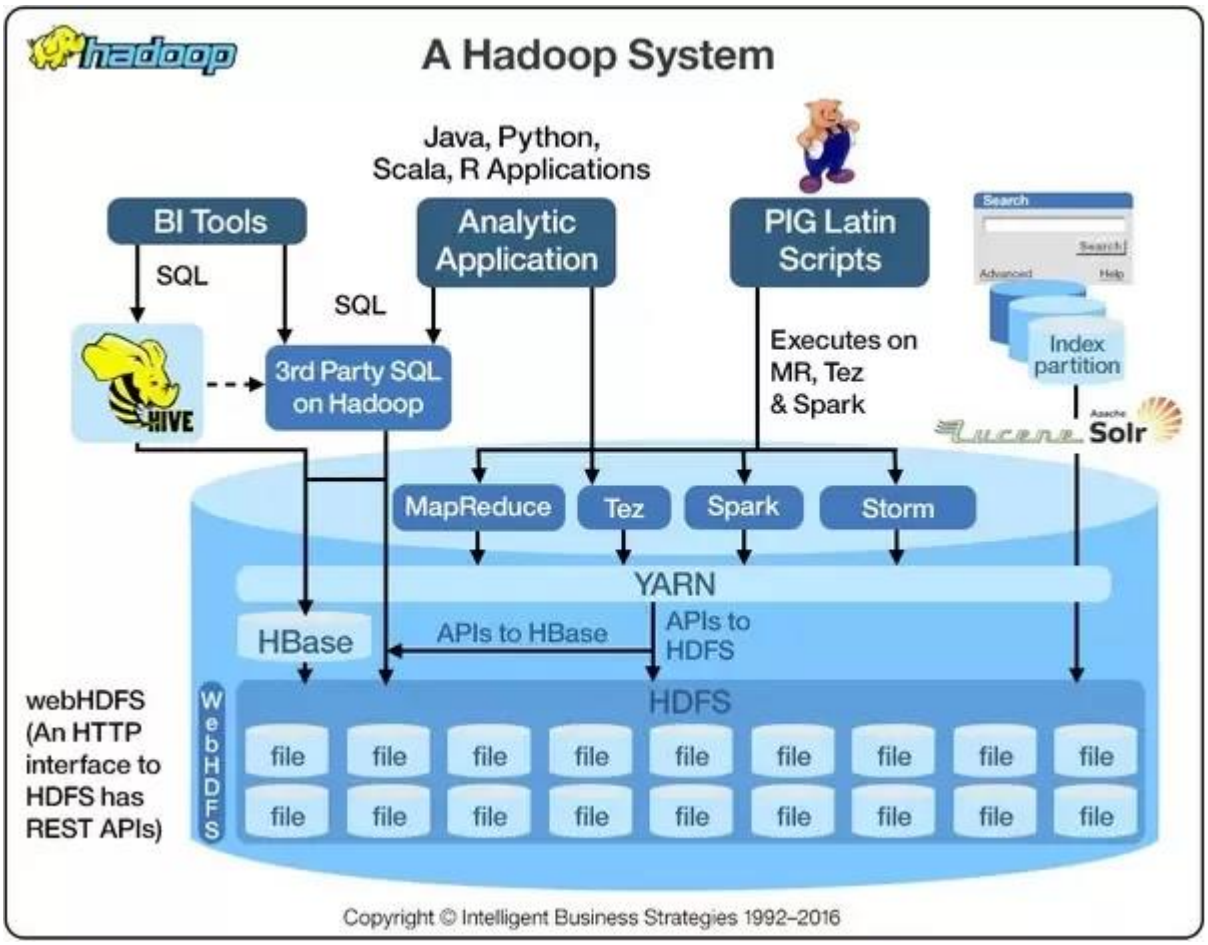
"When you get to that stage it becomes easier to hire for those roles, and when these roles are easier to hire for you don't have the crazy salary situation we had before."

Glassdoor's Zhao is also quick to point out there are still many aspects of being a data scientist that make it an attractive role -- **not least the fact that US data scientists are still taking home $95,459 in median annual pay**."One thing to keep in mind is that this isn't necessarily bad news for aspiring data scientists," he said.
"Data scientists still have one of the highest-paying and highest-job-satisfaction jobs in the United States."

However, he cautions new entrants to the field to go into it with their eyes open.

"But it does mean that competition amongst applicants is and will continue to be fierce in the coming years."

1.11 What is the current entry level salary in Hadoop/big data fields in USA?



IDC Infographic sponsored by SAP
Worldwide
ecosystem
ANALYTICS & BIG DATA
Growth Opportunities for SAP Partners
Analytics & Big Data Market Opportunities
The IT industry is witnessing a shift to what IDC calls the Third Platform of computing. It is made up of four key areas:
Cloud
Big Data/ Analytics
Mobility
Social Business
2020
2013
90%
FROM 2013 THROUGH 2020, 90% OF IT INDUSTRY GROWTH WILL BE DRIVEN BY 3RD PLATFORM TECHNOLOGIES THAT, TODAY, REPRESENT JUST 22% OF ICT SPENDING.
Customers clearly need help with Analytics and Big Data
69% of companies looking for skills in advanced analytics / predictive analytics in the next 12 months.
63% of IT and business executives are not familiar with the phrase "Big Data."
68% of companies do not have a stated Business Intelligence/ Analytics strategy.
Top 3 IT Challenges To Delivering a Successful Business Analytics Solution
Lack of Sufficiently Skilled IT Staff & Cost of Technology
Managing Data Quality
Data Integration
FROM 2005 TO 2020, THE DIGITAL UNIVERSE WILL GROW BY A FACTOR OF 300
FROM 130 EXABYTES TO 40,000 EXABYTES

6 High-Paying Careers for 2015

Robert Half *Salary Guides* reveal tech and creative positions with rising starting salaries:*

1 MOBILE APPLICATIONS DEVELOPER
$107,500 - $161,500
10.2% projected increase

4 USER EXPERIENCE DIRECTOR
$110,500 - $178,000
6.8% projected increase

2 BIG DATA ENGINEER
$119,250 - $168,250
9.3% projected increase

5 INTERACTIVE CREATIVE DIRECTOR
$100,500 - $180,250
5.7% projected increase

3 WIRELESS NETWORK ENGINEER
$99,000 - $137,500
9.1% projected increase

6 WEB DESIGNER (5+ years)
$80,000 - $112,500
4.8% projected increase

More information about the positions listed above can be found in the ***Robert Half Technology 2015 Salary Guide***, which includes a wide range of information technology job descriptions, and ***The Creative Group 2015 Salary Guide***, which focuses on interactive, design and marketing jobs. Learn more and download the guides at **rht.com/salary-center** and **creativegroup.com/salary-center**.

*All salaries listed are U.S. national averages based on data published in the 2015 *Salary Guides* from Robert Half Technology and The Creative Group. Actual projected salary ranges may vary depending on location.

© 2015 Robert Half International Inc. An Equal Opportunity Employer M/F/Disability/Vet

Robert Half® Technology

THE CREATIVE GROUP®

1.11.1Take a look at the salary ranges

- **Suggested Salary Range for Data Engineer – Junior/Generalist**

$70,000 – $115,000

- **Suggested Salary Range for Big Data Engineer – Domain Expert**

$100,000 – $165,000

- **Suggested Salary Range for Analytics Manager – 1-3 Direct Reports**

$90,000 – $140,000

- **Suggested Salary Range for Analytics Manager – 4-9 Direct Reports**

$130,000 – $175,000

- **Suggested Salary Range for Analytics Manager – 10+ Direct Reports**

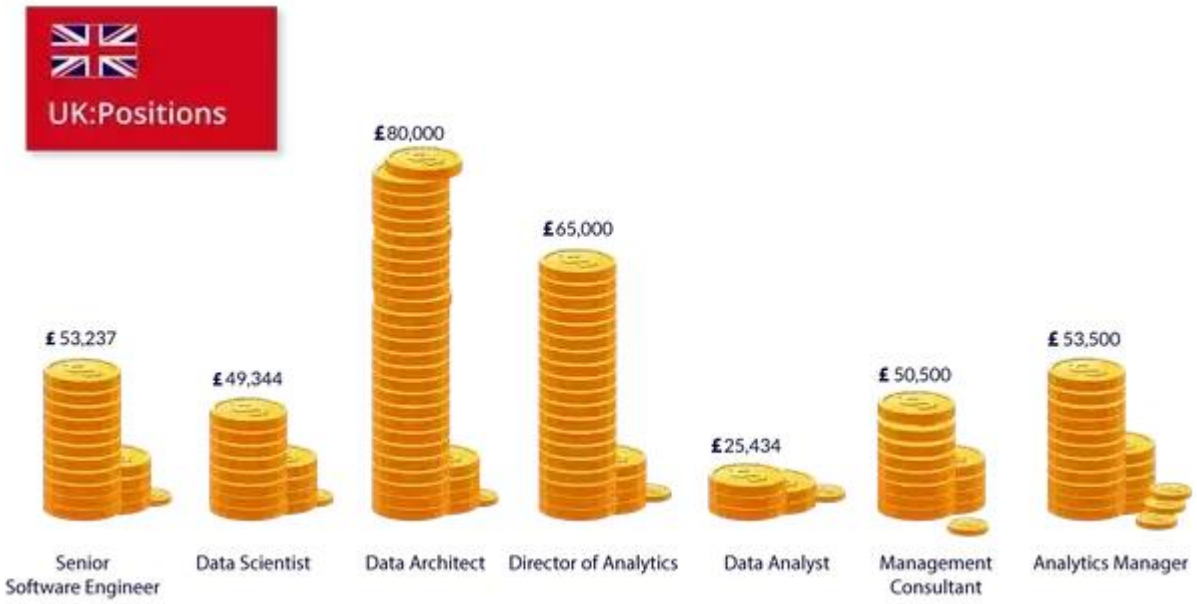
$160,000 – $240,000

1.11.2US: ***Major companies that are hiring for Hadoop***

- ***Amazon Inc:*** *$78,264 - $161,178*
- ***International Business Machines (IBM) Corp.:*** *$72,052 - $163,043*
- ***Capital One Financial Corp:*** *$90,200 - $183,994*
- ***Microsoft Corp:*** *$98,735 - $158,117*
- ***Booz, Allen, and Hamilton:*** *$54,248 - $172,310*
- ***Facebook Inc:*** *$92,110 - $199,332*
- ***Deloitte Consulting LLP:*** *$71,768 - $185,550*
- ***General Electric Co (GE):*** *$72,200 - $221,250*
- ***Expedia, Inc.:*** *$88,275 - $137,500*
- ***UnitedHealth Group:*** *$60,000 - $140,283*
- ***Google, Inc.:*** *$66,977 - $156,111*
- ***Accenture:*** *$78,906 - $183,125*
- ***J.P. Morgan Chase & Co. (JPMCC):*** *$92,371 - $182,322*
- ***Cisco Systems Inc:*** *$83,957 - $151,894*
- ***Comcast Cable, Inc.:*** *$73,899 - $157,812*

- ***eBay Inc.:*** *$110,738 - $213,679*
- ***American Express Co. (AMEX):*** *$85,569 - $140,482*
- ***The Nielsen Company:*** *$110,011*
- ***Citibank:*** *$94,259*
- ***Deloitte:*** *$92,500*
- ***EY (Ernst & Young):*** *$95,000*
- ***Uber Technologies, Inc.:*** *$111,910*
- ***Paypal, Inc.:*** *$142,482*
- ***Humana, Inc.:*** *$128,482*
- ***Apple Computer, Inc:*** *$132,635*
- ***Verisk Analytics:*** *$90,000*
- ***Johnson & Johnson:*** *$117,447*
- ***Wells Fargo & Co.:*** *$126,403*
- ***Oracle Corp.:*** *$143,415*
- ***Aetna, Inc.:*** *$85,059*

1.11.3England: *Position-wise salary distribution*



Experience
£ 25,458
£ 38,763
£ 47,184
£ 68,329
£ 82,782
Less than 1 year
1-4 years
5-9 years
10-19 years
20 years or more

1.12Which Industry Pays the Highest Data Analyst Salary?

Springboard analyzed salary information to determine what the typical data analyst salary is, which industry pays most, and how you can maximize your earning potential.

People send an average of 188 million emails every minute. We watch 4.5 million YouTube videos and fire off 18.1 million text messages in the same timespan. In fact, we produce over **2.5 quintillion bytes of data** each day.

This means that more and more companies are looking for talented analysts to process and interpret this valuable data. By 2020, **data science and analytics (DSA) job openings** are predicted to grow to 2.7 million, representing a $187 billion market opportunity.

Data analytics is an attractive career for those with a mathematics, statistics, or business background. The data analyst position often involves data mining, using languages like SQL/CQL, R, and Python to extract insights from large data sets, and communicating those ideas through visualizations and reports.

The need for data analysts is clear. But how much can you earn as an entry-level data analyst? What is the typical career progression? Can you increase your pay grade by targeting a different job title, transitioning to another industry, or moving to a new city?

This guide will show how you can maximize your data analyst salary and why it is so important to invest continuously in your personal and professional development.

1.12.1What to Expect at Different Levels of Experience

Earnings in data analytics are well above the median income of $47,000 for a full-time U.S. worker, but actual salaries change dramatically based on experience level.

Zippiareports that the average entry-level business / data analyst salary for someone with up to two years of experience and a bachelor's degree is $54,000. After two to four years on the job, you can expect an average salary of $70,000. A senior analyst with around six years of experience commands a higher salary of about $88,000, but a specialization can make the data analyst salary soar past the $100,000 mark. (Zippia analyzed its own data plus salary information from the U.S. Labor Department, PayScale

Typical Business Analyst Salaries

Position	Salary
Entry Level Business Systems Analyst	$56,043
Intermediate Level Business Systems Analyst	$71,865
Senior Business Systems Analyst	$86,635
Expert Business Systems Analyst	$100,312
Business Systems Analyst- Consultant	$113,202
Entry Level Business Data Analyst	$54,272
Intermediate Level Business Data Analyst	$70,123
Senior Business Data Analyst	$87,713
Business Process Consultant	$106,234
Business Development Director	$170,307

Data analyst salaries **increase steadily over the first five to 10 years**, but afterward, it can prove difficult to obtain raises without broadening your skill set.

For this reason, many data analysts end up switching to careers in data science or data engineering. As demonstrated by the differences between data analysts and data scientists, data scientists tend to have more specialized skills in advanced programming, data modeling, and machine learning, which justify a pay increase.

According to the **University of Wisconsin**, the average data scientist earns $113,000, with salaries typically starting between $50,000 and $95,000.

(It's important to note that there continues to be some fluidity when it comes to data-related job titles. Analysts may decide to apply for both data analyst and data scientist job postings, depending on their skills and knowledge. It's wise to pay more attention to the job requirements than the title itself.)

Data science positions may specify knowledge of languages and programs like Java, Python, predictive modeling algorithms, R, Business Objects, Periscope, ggplot, D3, Hadoop, Couch, MongoDB, and Neo4J.**O'Reilly's Data Science Survey** shows that knowledge of Spark and Scala can add a $15,000 salary boost, while mastery of D3 results in an extra $15,000.

PayScale predicts that five to 10 years of additional experience in data science is equal to $20,000 more in annual salary, 10 to 20 years of experience translates into

$30,000 in additional salary, and more than 20 years of experience results in a $55,000 pay boost over time. Hard work definitely pays off in the long run.

Another career path for people interested in building data pipeline architecture and infrastructure is that of a **data engineer. On Glassdoor**, data engineers report salaries between $110,000 to $155,000, with average salaries nearing $138,000. Skills, ranging from software development to Scala, result in anywhere from a 2-17% salary boost, according to **PayScale data**.

Skills That Affect Data Engineer Salaries

Scala	▲15%
Apache Spark	▲14%
Data Warehouse	▲13%
Java	▲12%
Data Modeling	▲11%
Linux	▲10%
Apache Hadoop	▲9%
Amazon Web Services (AWS)	▲8%
ETL (Extract, transform, load)	▲6%
Big Data Analytics	▲4%
Software Development	▲1%

Transitioning from data analysis to data engineering usually requires coding, machine learning/artificial intelligence, and/or natural language processing experience. Many job postings also look for a master's or Ph.D. in computer science, math, engineering, or statistics.

1.12.2What to Expect in Different Roles

As data analytics becomes a fundamental part of every industry, there is an increasing number of career paths to choose from. Data analysts can add tremendous value to different industries and progress within those fields. And this can impact the data analyst salary.

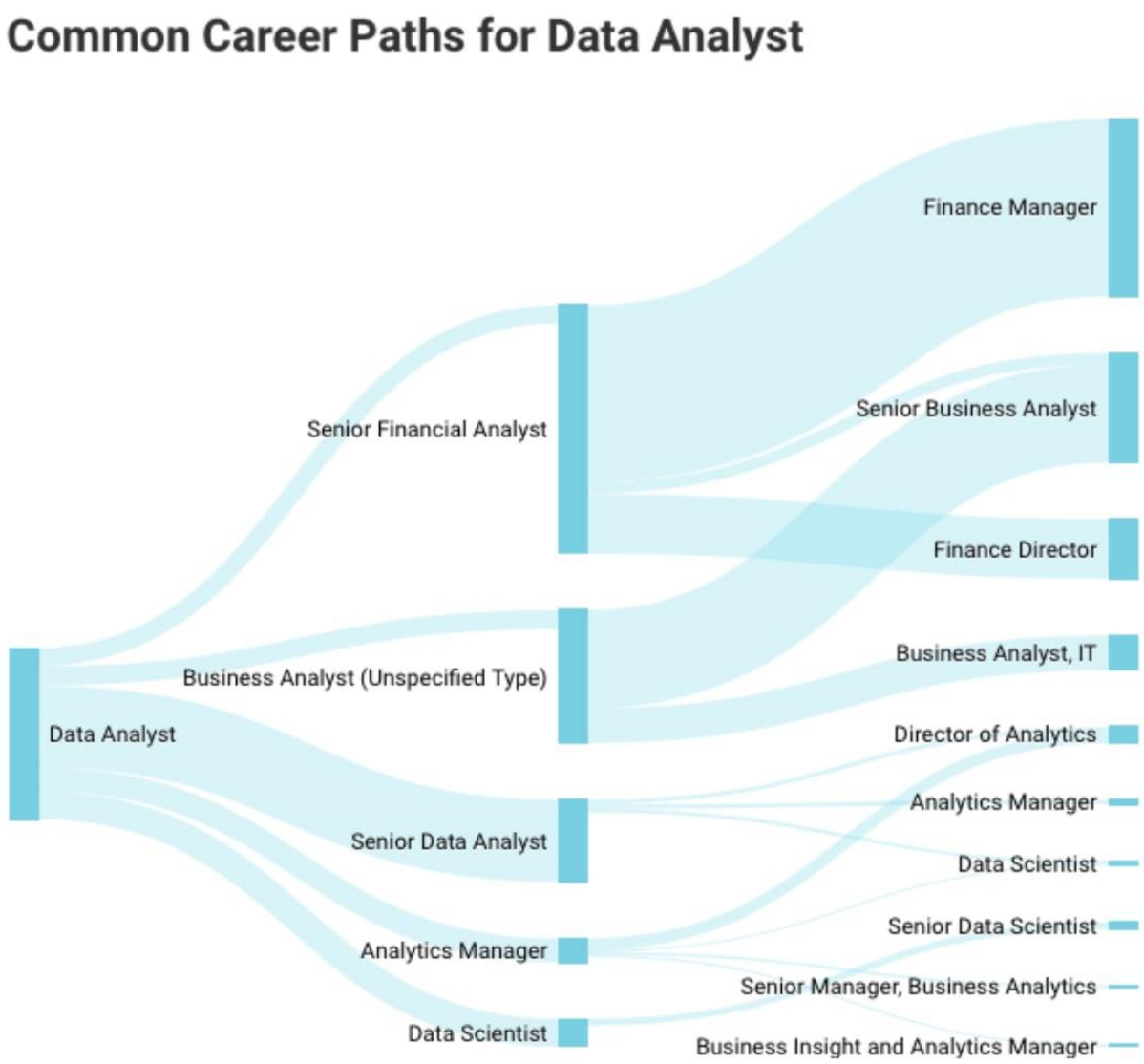
Data analysts who turn data-driven insights into actionable business recommendations are often called business analysts. They use **tools like Excel**, Tableau, and SQL. Business analyst salaries range from $54,700 to $69,000 at the entry level.

Financial analysts work with financial data to predict company performance and analyze macro and microeconomic trends, creating financial models and forecasts. **The2017 Robert Half Salary Guide for Accounting and Finance** calculates that most entry-level financial analysts at large companies make $52,700 to $66,000 and up to $50,000 in bonuses and commissions.

Market research analysts use sales, competitor research, and customer survey data to scope out the landscape for a particular product or service and how to optimize segmentation, targeting, and positioning efforts. This requires communication and storytelling abilities in addition to strengths in **quantitative data analysis**. The **starting salary range for this career** is between $51,000 and $65,000.

Healthcare data analysts manage analytics within hospital settings to streamline administrative tasks and improve the efficiency of operations. These **professionals earn** $46,000 to $80,000 and often have a particular interest in the healthcare industry.

Data analytics consultants apply the same data analysis skills; however, they advise several clients instead of working in-house for one company. They may evaluate an organization's data capabilities and technical systems, create new dashboards and reports, and suggest improvements to business processes**. Salaries in this position** vary from $58,000 to $112,000.



With data analysis skills, you can choose from other career paths that may not have “data analyst” in the title, but require similar experiences and qualifications.

For example, transportation logistics specialists usually work for delivery/shipping companies, airlines, and urban planning organizations. They use analytics to understand the most efficient transportation routes, given variables such as traffic, weather, and time of day. Salaries for transportation logistics

specialists **usually start around $79,000**.Digital marketing managers also use tools like Google Analytics to understand the effectiveness of different advertising/marketing channels and track how investment in campaigns translates into customer acquisition and click-through rates. According to **Glassdoor**, digital marketing managers make a base pay of around $73,000.

Operations analysts work with data collection, management, and reporting, as it pertains to streamlining and optimizing a company's internal processes and procedures. Salaries for operations analysts start at $54,000, according to **Payscale**, but increase significantly as people progress to the managerial level.

IT systems analysts need to understand the technical requirements of an IT system, usually for testing and deploying software. They conduct research on pain points and figure out how to use custom tools to change technical configurations, design new workflows, and analyze the impact of system changes on stakeholders. Systems analysts usually start with salaries of around **$68,000**.

As you can see, data analysis is a very versatile skill set that makes you attractive to employers across diverse fields. As a data analyst, you can easily craft a career around your functional and industry areas of interest and expertise.

1.12.3How to Boost Your Data Analyst Salary

If you are interested in increasing your data analyst salary, there are many ways to do so, beyond choosing an attractive industry. Here are some other strategies for improving your hireability in the market.

1.12.4Expand Your Skill Set

By learning new tools like Spark and Scala, D3, cloud computing software such as Amazon Web Services, and open-source tools, you can increase your earning potential dramatically. The most valuable analytical skills are MapReduce (average salary of $115,907), PIG ($114,474), machine learning ($112,732), Apache Hive ($112,732), and Apache Hadoop ($110,562).

1.12.5Choose the Right Location

Jobs in large cities like Chicago, San Francisco, and New York tend to pay the most. Generally, companies benchmark salaries against competitors when recruiting talent. When deciding whether or not to relocate, think about state taxes and the cost of living in the places you are considering. Some remote positions may pay less, but the flexibility of working from home in a cheaper city or country could constitute a more attractive offer.

Data Analysts are highest in demand in SF Bay Area, London, and New York. Browse and compare average salaries in locations where this role is also popular:

1.	**Seattle**	**$130,352**	8.	Toronto	C$99,326
2.	SF Bay Area	$124,290	9.	San Diego	$96,458
3.	Los Angeles	$122,736	10.	Chicago	$96,020
4.	Washington D.C.	$118,475	11.	Austin	$93,571
5.	New York	$115,828	12.	Dallas/Ft Worth	$83,125
6.	Denver	$110,761	13.	London	£57,248
7.	Boston	$105,510	14.	France	€49,643

1.12.6Gain More Experience

According to IBM, 81% of DSA jobs require **three to five years of experience**. In addition to experience at traditional companies, you can also take on f**reelance projects** to boost your portfolio through popular platforms like Fiverr, Upwork, Kaggle, and Toptal. Supplementary experience may allow you to enter in at a more senior or managerial level, increasing responsibilities and your base data analyst salary.

1.12.7Negotiate Your Job Offer

When applying for a data analyst position, it is often helpful to cast a wide net with the goal of receiving multiple offers. Then you can use compensation information as leverage to negotiate a higher salary at your dream company.

Glassdoor estimates that the average American employee could earn an extra 13%, or $8,000, by understanding their market value, but only 39% negotiated their last offer. The typical worker is leaving a lot of money on the table.

When comparing offers and considering your negotiating power, make sure to consider the full package offered by each company, including perks, bonuses, and benefits, in addition to career paths, mentorship, and cultural fit.

1.12.8Pursue a New Degree or Certification

According to IBM, only 6% of data analyst postings require a master's degree or higher, whereas 39% of data scientist and advanced analytics postings look for graduate-level education. Even without returning to formal education, there are many ways to increase your skill set through online courses and bootcamps.

Springboard offers a free Data Analytic learning path for people looking for a short but intensive introduction to the topic. If you already have some experience and are looking for more intensive training as a data analyst, consider the **Data Analytic Career Track**. This mentor-guided bootcamp includes a job guarantee—find a job within six months of graduating or we'll refund your tuition.

This type of training shows future employers that you are eager to learn and develop new areas of expertise—a major plus in a field that is constantly evol

1.13You might also be interested in...

1.13.1Shareable Data Quotes That Will Change How You Think About Data

Data Quotes The amount of data generated in real time is immense. This has created oceans of data from which companies can derive real business value and make better business decisions. The big data market is predicted to grow by 20% this year, and by 2020, every human is expected to generate 1.7 megabytes (of […]

1.13.2Free Data Visualization Tools

Data may be the buzzword of the decade (and the oil of the 21st century), but without the right storytelling tools, data is just data—boring, confusing, and uninspiring. Thankfully, it's easier than ever before to find the data visualization tools you need to start transforming numbers and statistics into workable strategies and business goals—and on a […]

1.13.3Data Analyst vs. Data Scientist

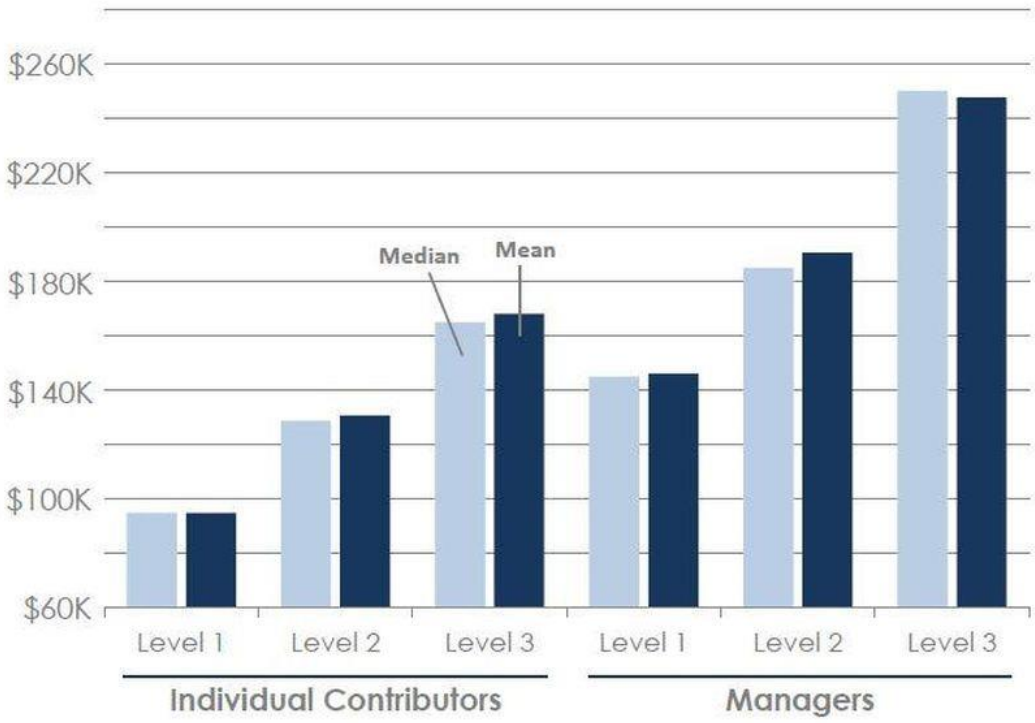
If you have an analytical mindset and love decoding data to tell a story, you may want to consider a career as a data analyst or data scientist. After all, they are two of the hottest jobs in tech (and pay pretty well, too). Harvard Business Review even awarded "data scientist" the title of "sexiest […]

1.14The Salaries of Data Scientists Remain Steady but Still Sexy With AI on the Horizon

The just-published 5th-annual Burtch Works Study: Salaries of Data Scientists provides fresh insights into the compensation trends for those holding **the sexiest job of the 21st century:**Similar to last year's figures, the median base salaries of data scientists have remained relatively steady. After rising rapidly in 2015 and 2016, median base salaries at all job levels changed by a single-digit percentage point or not at all from March 2017 to March 2018.

- The median base salaries range from $95,000 at level 1 (0-3 years of experience) to $165,000 at level 3 (9+ years) for individual contributors and from $145,000 at level 1 (1-3 reports) to $250,000 at level 3 (10+ reports) for managers.

Figure 4 Median and Mean Base Salaries of Data Scientists by Job Category



1.14.1Data Scientists Salaries 2018

BURTCH WORKS

- If you are a data scientist, you don't need to become a manger to earn a high salary. Becoming a subject matter expert can be a separate career track that doesn't have to lead to people management.
- Data scientists holding a PhD earn higher median base salaries than those holding a Master's degree in every job category. This trend is most pronounced among individual contributors—at level 1, data scientists with a PhD earn a median base salary of $102,000 while those with a Master's degree earn a median base salary of $92,500.
- Data scientists earn base salaries up to 36% higher than other predictive analytics professionals. In all job categories except the higher-level managers, data scientists realize a double-digit percentage advantage in base salary compared to others in predictive analytics. (Burtch Works distinguishes data scientists from other predictive analytics professionals by "the depth and breadth of data scientists' coding skills," allowing them "to exploit data regardless of its source, size, or format.")
- The salary gap between data scientists working for technology companies and those in other industries is

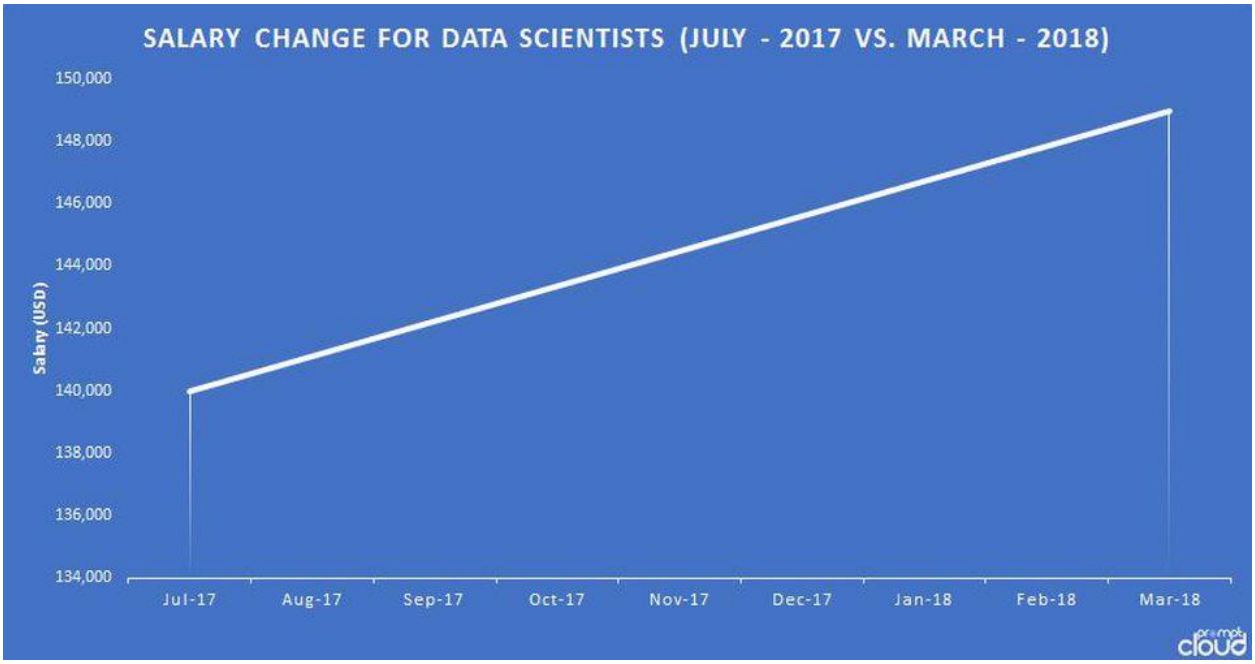
beginning to narrow. Yet another indicator that today, all companies are tech companies…

- Previously, H-1B residency status held constant at 9% (2015 and 2016) and 8% (2017) of the data scientists covered by the Burtch study, but this year's data shows a jump to 13% of data scientists holding an H-1B visa. This may be due, says Burtch, "to many companies realizing that it is important to employ an aggressive hiring strategy for data science and gain access to a wider talent pool by supporting the visa transfer process. Additionally, the lengthening green card process could be at play."

Sexy salaries attract new entrants into a new profession and a new profession evolves to accommodate the increasing demand for its services. There has been a shift to requiring Master's degrees instead of PhD's which in turn helped drive the **proliferation of data science Master's programs**. The pool of available data scientists has increased due to the influx of existing statisticians and analytics professionals upgrading their skills (and their job descriptions) and newly-minted graduates of data science educational programs (including many online courses).

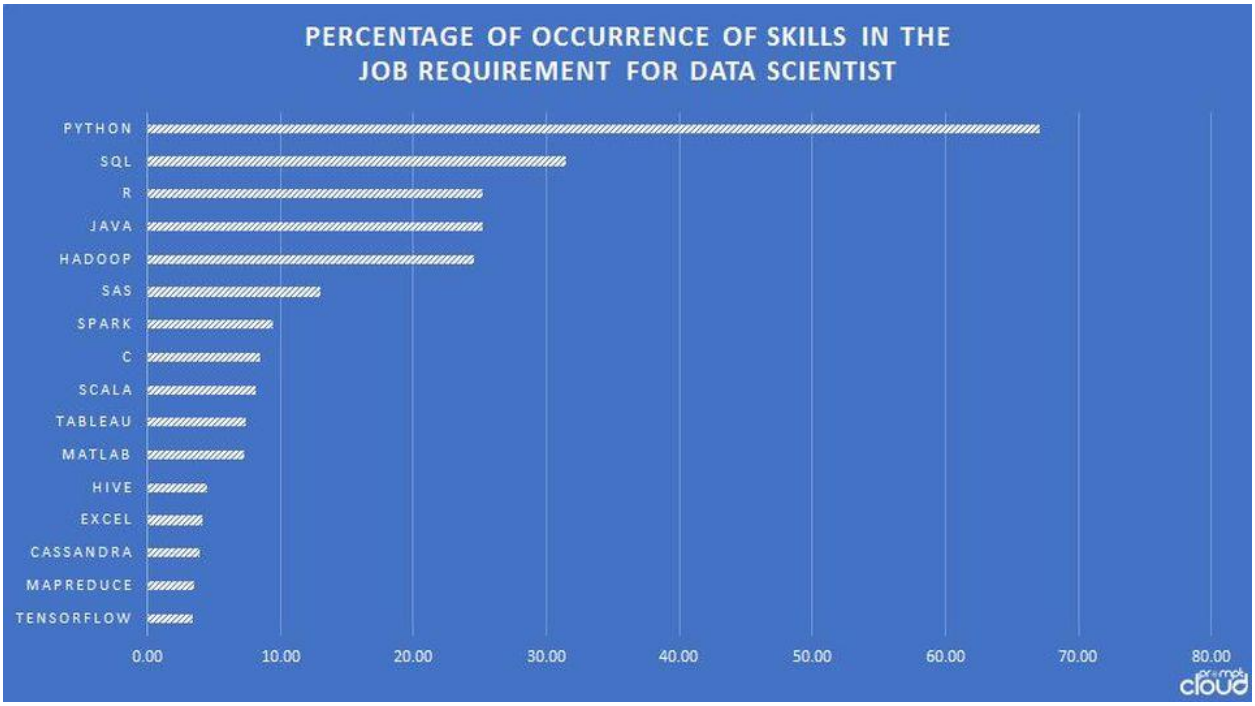
The supply of data scientists has increased, but so has the demand. As a result, salary growth has moderated but the salaries offered to data scientists still remain at the "sexiest" level. Confirming the Burtch study in this regard, is an analysis conducted by Web data provider **PromptCloud**, exclusive to *Pressed Data*. Analyzing job listings for data scientists on Dice.com in July 2017 and March

2018, PromptCloud has found that the average salary offered has increased 6.43%.



1.14.2Data Scientists' Average Salary 2017-2018 PROMPTCLOUD

The job listings also confirm the Burtch study's contention that data scientists are distinguished from other analytics professionals by their coding proficiency. An analysis of the required skills listed in these jobs descriptions shows that the Python programming language leads by a wide margin other popular statistical analysis, data management, and coding tools.



1.14.3Data Scientists: Top Required Skills PROMPTCLOUD

Python was first implemented in the late 1980s, but it was its 3rd version, released 10 years ago, that became—with the rise of "big data"—a widely popular tool among data scientists. Computer programming skills, especially in Python, contributed to the emerging definition and institutionalization of the new discipline, and with it, specialized training programs and how they were differentiated from their immediate predecessors, the "business analytics" programs.

One such new program was the **Master of Data Science program at the Illinois Institute of Technology (IIT). Shlomo Argamon**, Professor of Computer Science at IIT and director of the program from its inception five years ago, has also observed the broadening

of both the demand for and the supply of data scientists. Says Argamon:

“The awareness of the usefulness of data science has grown, and as tools become more widely available, companies in a greater variety of industries are making use of data scientists, either in-house or via a growing number of data science consulting firms.”

And as the Burtch study also noted, Argamon has increasingly seen the graduates of the IIT program finding employment not only with tech firms and data science consultancies, but also “in a wide variety of industries such as insurance, media, manufacturing, banking, and more.” And he adds: “I should note that, despite the hype, it seems that most firms understand that they cannot just plug in some tools and expect results, that they need actual data scientists to get the insights they seek.”

The increase in the diversity of employment opportunities has been matched by the increasingly varied background of the people entering the data science profession. Argamon: “We’ve always sought a diversity of backgrounds among our students, as we see that as strengthening the learning experience for budding data scientists – our goal is to produce well-rounded data scientists who can effectively communicate with both technical and non-technical colleagues and clients, and who can learn new domains and collaborate with diverse experts.”

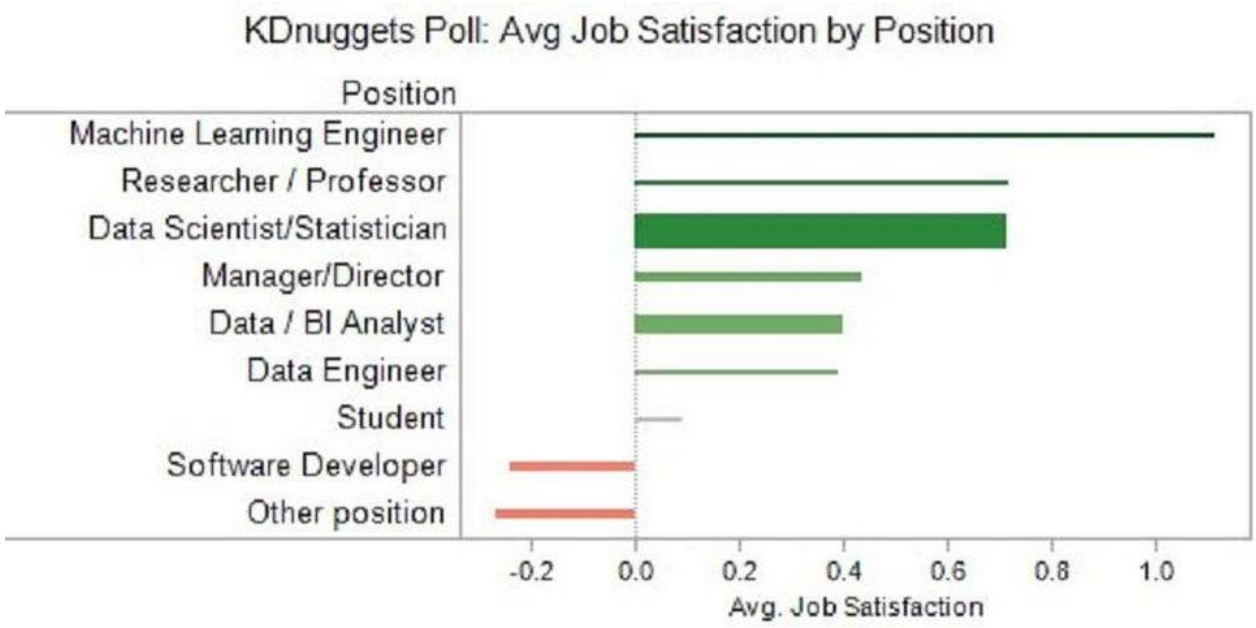
Over the last five years, Argamon has observed the growing attraction of data science to entrants from many different disciplines and experiences: “While our earliest students were almost entirely from computer science, engineering, and statistics backgrounds, we have seen growth in the numbers of our incoming

students from other backgrounds, with majors such as business, history, biology, chemistry, economics, and anthropology. These different ways of thinking, we believe, will contribute to the breadth and effectiveness of these students and others in their cohort as data scientists."

At IIT, as in similar data science programs, an integral part of the curriculum is practical experience. In addition to putting to test in real-life situations what they have learned in the classroom, these experiences highlight to the students (and would-be students) the value of data science for business and society.

In 2016, IIT's basketball coach, Todd Kelly, turned to the data science program for help in what was probably the very first use of *Moneyball*-type methods in Division III sports. In 2013, IIT's record was 0-25 and the next year, Kelly's first as a coach, it improved to 2-23. After two IIT data science students spent the summer of 2016 collecting and analyzing data about the team, the **IIT Scarlet Hawks** finished the next season with a 22-6 record, making it all the way to the U.S. Collegiate Athletic Association Division 1 National Basketball Championship.

This kind of results make data scientists happy about the work they do. A recent **KDnuggets job satisfaction survey** found "that while Data Scientist may no longer be the sexiest profession, it is still very satisfying." The jobs with the highest satisfaction, KDnuggets reports, are Machine Learning Engineer, Researcher/ Professor, and Data Scientist/Statistician. Is this a leading indicator that "machine learning" is the next "data science" in terms of sexiest salaries, satisfaction, and skill shortage?



1.14.4Job Satisfaction by Position (Bar width corresponds to the number of respondents)

KDNUGGETS

Burtch Works reports that they see a shift in job requirements toward machine learning and AI skills: "Now that many teams have grown, the need for generalist 'unicorns' has given way to a higher demand for data science specialists that can work with other analytics professionals and data engineers on the team. For example, we now regularly receive requests for experts in specific areas such as NLP (natural language processing) or image processing using tools like TensorFlow."

AI specialists already command **super-sexy salaries** and the latest in machine learning and artificial intelligence is being added to data science training and career **transition programs such as Insight Data Science**. And so goes the cycle of evolving disciplines and job categories, all attempting to improve the methods and tools for data analysis and for finding new uses for computer-generated data.

1.15 4 STEPS TO INCREASE YOUR SALARY AND BUILD A HIGH-PAYING CAREER AS A DATA ANALYST

Most successful data analysts have to wear many hats. Organizations in virtually all industries increasingly rely on

data to drive the decision-making process and identify business inefficiencies that need to be addressed.

“Analytics professionals are like a Swiss army knife,” says Paul Wallenberg, unit manager of tech services at IT recruiting firm LaSalle Network. “They’re not just a subset of IT, they’re embedded in departments across the company. These skills and roles set you apart from IT because these they give you the fluidity to bounce between departments, making you more marketable and valuable.”

There’s good news for these professionals in the analytics field: Job opportunities are growing and the demand for qualified individuals to fill them is high. Because of this demand, these professionals are compensated well.

1.16 Data Analytics Job Market and Salary Trends

According to the Bureau of Labor Statistics, the field is projected to grow 23 percent from 2016 to 2026—much faster than the average for all occupations.

But finding qualified individuals to fill these roles isn't easy. These professionals are responsible for helping organizations make sense of their technical data in order for executives to make better business decisions, and that requires a unique skill set, which combines technical savvy with business knowledge.

With more than 2.3 million job openings in 2017, experts predict that by 2020, that number will swell to more than 2.7 million—and some of these jobs are taking twice as long to fill than the national benchmark average. This industry need, coupled with the shortage of talent, means data analysts are well compensated. According to Robert Half Technology's (RHT) 2019 Salary Guide, these professionals earn between $81,750 and $138,000 depending on their experience, education, and skill set.

1.17Highest Paying Industries for Data Analysts

While there is a need for data analysts in nearly all industries, some fields are experiencing higher demand—and increasingly competitive compensation standards—for skilled data professionals. According to RHT's Salary Guide, the top industries in 2019 are:

- Financial services
- Manufacturing
- Software as a Service (SaaS)
- Healthcare

To maximize your earning potential, career experts suggest that you focus on four areas: mastering today's in-demand data skills, considering an advanced degree, seeking extra experience, and knowing the range in which your salary lies.

1.18 4 Ways to Increase Your Data Analyst Salary

1.18.1 Master the Necessary Skills

The skills that are in demand today aren't always the ones that will be in demand tomorrow. That's why it's important to keep tabs on industry trends and make sure you're up to date on the latest skills, says John Reed, senior executive director at RHT.

Experience with databases like Microsoft SQL Server, Oracle, and IBM DB2 are essential skills for managing data, he says. Mastery of Microsoft SQL Server and Oracle, in particular, could raise your salary by as much as eight percent and six percent respectively, according to RHT 's 2017 Salary Guide.

Acquiring additional programming skills—particularly R and Python—can also increase your earning potential, says Blake Angove, director of technology services at IT recruiting firm LaSalle Network.

"When you have in-demand skills, you're in a better position to ask for higher pay," Reed says. "Your degree can certainly give you a lot of knowledge and skills needed as a data analyst through math, computer science, data modeling courses, and more. And if you'd like to expand your skillset, consider taking an online course."

1.18.2 Pursue an Advanced Degree or Certificate

Businesses are having a difficult time finding qualified individuals with the right mix of skills, training, experience, and education, which is why salaries for these professionals are so competitive.

Entry-level data analysts typically earn a bachelor's degree from a science, technology, engineering, or math (STEM) field, but pursuing an advanced can fast-track you to a promotion and higher salary, Angove says. In fact, 39 percent of advanced analytics jobs require a master's degree or PhD, according to IBM.

Seek out a program with a focus on math, science, databases, programming, modeling, and predictive analytics, he suggests. Northeastern offers a Master of Professional Studies in Analytics, for example, which equips students with a deep understanding of the mechanics of working with data, along with the capacity to identify and communicate data-driven insights that influence decision-making.

"There's a definite trend in obtaining formal degrees in analytics," says Paul Wallenberg, unit manager of tech services at IT recruiting firm LaSalle Network. "If you're interested in perhaps marketing yourself as a data scientist down the road, for example, having a PhD is highly preferred, especially at top-tier organizations, which pay the highest."

Northeastern's MPS in Analytics, has three concentrations:

- **Statistical Modeling**, which is intended for students pursuing careers as a data scientist or quantitative analyst;

- **Evidence-Based Modeling**, for students interested in developing an understanding of how advanced analytics can be used to reduce decision-making uncertainty; and
- **Informational Design**, for those who want to broaden and deepen their understanding of standardized data flow, management, and reporting processes.

Northeastern also offers master's degrees in business analytics, data science, and informatics, depending on the area you want to specialize in.

"We not only focus on the technology and techniques, but we're also focused on how to read and communicate the data," says Uwe Hohgrawe, assistant teaching professor and lead faculty for the MPS in Analytics. "If you're more interested in expanding your knowledge specifically in analytics, then a master's program is for you."

1.18.3 Seek Extra Experience

Whether you're considering a graduate degree or are already enrolled in a master's program, there are steps you can take now to boost your earning potential, Reed says. To begin, gain hands-on experience through internships or project-based roles while you're in school to build up your resumé.

"These experiences can help you earn a higher salary as a new tech professional, since you'll already have some of the skills and knowledge needed," he says.

Research companies and job openings in your area and network with industry professionals to see what opportunities

are available. With more real experience and tangible projects, employers may boost your salary.

“One of the best ways to increase salary is to show an employer how you’ve brought value to a previous employer or project,” Reed adds.

Additionally, becoming involved in your local analytics community is a great way to get more visibility, which could lead to inquiries from head hunters with jobs that could boost your salary, Wallenberg says.

He suggests researching local groups on the networking site MeetUp. “If you’re involved in a Hadoop, R, or Python MeetUp group in your city, you’ll probably connect with people who can help you land higher-paying jobs,” he says.

Other sites to connect with others on include GitHub, an open-source development platform, and Stack Overflow, an online community for developers.

1.18.4 Don’t Sell Yourself Short

If you’re already employed as a data analyst and are seeking a promotion or bump in pay, prepare yourself for that conversation with your manager. There are a variety of salary calculators and resources available to check how your compensation stacks up in the market, including Salary.com, PayScale, and RHT ’s annual Salary Guide.

These resources are essential to have the most productive conversations with hiring managers, Reed says. “If you find you need to negotiate for higher pay, present your research

during a salary discussion with your potential or current employer in order to help strengthen your case. Understanding what your industry peers in your geographic location are making is extremely important to knowing if your compensation is up to par.”

1.19Building Your Data Analytics Career

There is no shortage of opportunities for aspiring data analytics professionals in the current job market. With more businesses than ever relying on data for decision-making purposes, the demand for skilled data analysts is continuing to rise. If you think a career in data analytics is right for you, consider the ways you can prepare yourself to break into the field and build a successful, high-paying career.

There are also ways for *current* data analysts to kickstart their careers and increase their earning potential. If you want to take the next step in your data analytics career, there are steps you can take to develop your skills that could translate to an increase in salary.

1.20 These Are the Highest Paying Jobs for the Class of 2019

- Starting pay for Data Scientist graduates reported at $95,000
- Physical therapist is the only non-technical job on the list

The Class of 2019 is getting ready to enter one hot job market. And grads with a background in slicing and dicing data have an edge.
Data scientist was the highest-paying entry-level job last year, according to Glassdoor research. Young adults in this field earned a median annual

base salary of $95,000. That's higher than young Wall Street workers received as investment-banking analysts.

1.20.1Highest Paying Jobs for New Grads

Entry-level Job	Base Salary
Data Scientist	95,000
Software Engineer	90,000
Product Manager	89,000
Investment Banking Analyst	85,000
Product Designer	85,000
UX Designer	73,000
Implementation Consultant	72,000
Java Developer	72,000
System Engineer	70,000
Software Developer	68,000

Businesses from fashion giants to hedge funds to venture capital firms are building teams focused on figuring out how to turn the information they capture into valuable insights.

“There’s such a high demand across industries for people with data-science skills,” said Amanda Stansell, senior research analyst at Glassdoor.

Recruitment for this profession will continue as more companies accumulate data and figure out what to to do with it, Stansell added.

Tech jobs dominated Glassdoor’s list of the top 25 highest-paying entry-level positions. Software engineer ranked second, with a median base salary of $90,000. The only non-

technical and non-business profession on the list was physical therapists with a median salary of just under $64,000.

Job openings for the information-technology sector have increased more than 4% in the past year, according to the S&P 500 LinkUp Jobs Index. During the same period, postings for the financial industry dropped 10%.

Coding boot camps have proliferated in recent years amid talk of a skills shortage, and some companies have worked to combat the gap by partnering with locals schools to provide mentorship and training.

Graduates looking to pursue one of these jobs don't need to limit themselves geographically. In addition to the coastal tech hubs, several other cities offer unique tech opportunities, including Huntsville, Alabama, and Detroit, Michigan.

1.21What Kind of Salary Can You Expect as a Data Scientist?

Data scientists are key players in the big data revolution that has been changing the way businesses, government agencies, and non-profits do almost everything. With the analytical and quantitative skills required to unravel the mysteries hidden in big data, data scientists deliver the fuel required to drive organizational goals and offer surprising insights about the behavior of people and systems that before remained obscure and ignored.

So it should come as no surprise that working with big data can offer up big paychecks. Companies may be willing to pay an arm and a leg to a talented data scientists, but it's only because the value they bring to the table offers even bigger profits. McKinsey Global Institute's December 2016 analysis of how big data is impacting business brings to light some major disparities emerging between industry leading enterprises and average companies, a dynamic that is creating a winner-take-all scenario in some market verticals.

When it's a case of ratcheting-up salaries for data scientists or risking irrelevance, you can be sure companies are going to spend what it takes to stay competitive.

In recent years, data storage prices have plummeted, allowing many organizations to save and catalog enormous sets of unstructured data. This has resulted not only in increased job opportunities for data scientists but increased salaries, as well. The longer valuable data sits there, unexplored, un-mined, the more of the benefits of collecting that data are squandered.

Companies are willing to pay a premium to get accomplished and experienced data scientists on the job ASAP.

1.22Wanted: Data Scientists (And Willing to Pay)

A 2017 study by IBM and job market data firm Burning Glass Technologies projects a 39% increase in demand for data scientists and data engineers by 2020. That's a total of 2.7 million jobs, which McKinsey estimates will be around 250,000 more openings than there are qualified candidates to fill them.

Increasing requirements for specializations further narrows the window for top talent in a way that these reports can't fully quantify. The risk for firms, and job seekers, is that the skills required for a particular position might not be met by candidates even if technically, there are enough data scientists available.

This has lead to a market that has not only boosted salaries for individuals, but has driven up the valuation of entire technology firms heavy in big data talent. In 2014, Google acquired artificial intelligence firm DeepMind for more than $600 million, a price that breaks down to almost $7 million per employee… many of them data scientists who simply couldn't be hired any other way.

But companies are certainly making an effort to attract top talent with big paychecks. O'Reilly's 2016 Data Science Salary Survey found that U.S.-based data scientists enjoyed a median salary of $106,000. Among respondents, the real story wasn't simply the current base salary, though; for around half of them, the number had jumped by 20% year-over-year, and for a lucky 12%, their salary had at least *doubled*.

Pay Hikes are Going to Leaders and Specialists in the Field

In their 2017 salary survey, Burtch Works, an executive talent recruiting firm, breaks the numbers down even further, classifying data scientists according to three different career levels and based on whether or not they are in a management role:

Position	Median Salary	Percentage Increase YoY
Entry-level Data Scientist	$95,000	-2%
Mid-career Data Scientist	$126,000	1%
Advanced Data Scientist	$150,000	3%
Entry-level Manager	$148,750	6%
Mid-level Manager	$190,000	0%
High-level Manager	$250,000	4%

Not only is the range surprising, but the fact that entry-level positions dropped slightly reflects the increasing demand for specialization… a skill set that isn't available yet to most newly minted data scientists. The demand for managerial talent at the entry level, however, echoes a finding in the McKinsey report that some 45% of organizations report their biggest obstacle to leveraging big data is in "…designing an appropriate organizational structure to support data and analytics activities."

McKinsey also finds that the issue is more prominent in some industries than others, with manufacturing, public sector, and retail organizations lagging in the adoption and leveraging of data science tools.

Another study by McKinsey projects that by 2018, the U.S. may face a 50 to 60 percent gap between supply and demand of "deep analytic talent." There are a number of industries already experiencing this shortage, including insurance, finance, pharmaceuticals, and aerospace.

1.23Salaries by City and Region: Where They Are, Where they're headed

The Robert Half Technology 2017 Salary Guide revealed a 6.4% increase in data scientist salaries between 2016 and 2017… less than the 9% seen the year before, but still the single largest bump among all job categories the company tracks. According to Robert Half, the salary range for data scientists in 2017 was $116,000 – $163,500.

Salary ranges for data scientists in some of the country's largest cities (as of 2016) include:

- **New York, NY: $152,600 – $215,250**
- **Dallas, TX: $118,265 – $166,819**
- **Los Angeles, CA: $139,520 – $196,800**
- **Miami, FL: $116,630 – $164,512**
- **Seattle, WA: $129,601 – $182,809**
- **Pittsburgh, PA: $106,820 – $150,675**
- **Saint Louis, MO: $109,000 – $153,750**
- **Atlanta, GA: $115,119 – $162,382**
- **Chicago, IL: $134,070 – $189,112**

The 2017 Burtch Works publication "Salaries of Data Scientists" broke down salaries according to geographical region. For example, in every job category, data scientists earned more on the West Coast and the Northeast than those in the Middle U.S. West Coast data scientists posted the highest salaries in the nation, bringing home a median base salary at mid-career of $140,000; that's 8% more than their

counterparts in the Northeast and 14% more than those in the Middle U.S.

The Burtch Works study broke down data scientist salaries further by seniority/position, with Level 1 being entry-level positions and Level 3 being senior positions:

- **Level 1**
  - **Northeast: $90,000**
  - **Middle U.S.: $88,500**
  - **West Coast: $102,500**
- **Level 2**
  - **Northeast: $130,000**
  - **Middle U.S.: $120,000**
  - **West Coast: $140,000**
- **Level 3**
  - **Northeast: $160,000**
  - **Middle U.S.: $155,000**
  - **West Coast: $167,500**

This is leveling out, however… the major increases in salary from the pervious year's numbers happened primarily in the Northeast and Midwest, with the Midwest experiencing the highest jump, while West Coast data scientists actually saw median salary increases correct and pull back a bit.

Burtch Works also revealed annual, average salaries for data scientist managers, by region and by level:

- **Level 1**
  - **Northeast: $150,000**
  - **Middle U.S.: $140,000**

    - **West Coast: $153,800**
- **Level 2**
    - **Northeast: $191,500**
    - **Middle U.S.: $180,000**
    - **West Coast: $200,000**
- **Level 3**
    - **Northeast: $240,000**
    - **Middle U.S.: $250,000**
    - **West Coast: $275,000**

1.24Beyond Salary: Perks and Advantages of Being a Data Scientist

Salaries have been fluctuating somewhat unpredictably as organizations identify and look to fill specialty roles, and in response to the influx of new job-seekers entering the field. To some extent, the field of data science is sill defining itself and its role within certain industries. After all, the field was first identified and classified just ten years ago in 2008.

Since data science is all about discovering new ways to leverage the massive amount of data organizations have been accumulating, it isn't completely clear yet in all verticals exactly what the next discovery will yield and how it will be applied. To some degree, the role remains open to interpretation among analysts and hiring managers alike. In some instances, the true value of a data scientist may not be fully understood until they are actually on the job.

For that reason, starting salaries are not the only measure of demand, and they shouldn't be the only factor you consider in your job search.

Burtch Works, for example, notes that 85% of survey respondents are eligible for bonuses over and above the base salary. Median bonus numbers were reported to be from $10,000 to $29,500.

There is also much value to be found in having a job that allows for mobility. O'Reilly found that 64% of data scientists found it easy or very easy to locate a new position. That provides enormous bargaining power when it comes to

negotiating a salary, not to mention real quality-of-life benefits when it comes to choosing where and how you want to live.

A solid education is a must-have for a prospective data scientist and an advanced degree is a big investment. But with salaries in the field booming, it's an investment that will pay big dividends for years to come.

1.24.1Top Ranked Universities Ranked in Big Data:

Featured Programs:

- **Syracuse University - M.S. in Applied Data Science: GRE Waivers available**
- **SMU - Master of Science in Data Science - Bachelor's Degree Required.**
- **UC Berkeley - Master of Information and Data Science Online - Bachelor's Degree Required.**
- **Syracuse University - Master of Information Management Online**
- **Villanova Business - Master's in Analytics and Study Data Mining, Predictive Analytics Online**

1.25 Top 7 Big Data Jobs

Ranking of Big Data Jobs for 2019

Organizations in most sectors of the economy have begun to collect, organize, store, and interpret massive amounts of data related to their operations. In a recent survey by NewVantage Partners, 91.6% of the executives queried said that they were increasing their investments in Big Data. 91.7% of these respondents said they felt their investments were necessary to transform their operations into more agile and competitive businesses.

This is great news for job candidates who are seeking employment within one of the top big data job fields.

Where the Big Data Jobs Are

Due to the rapid expansion of big data applications, demand for workers has increased dramatically in recent years. LinkedIn reports a significant shortage of data science workers nationally and in major cities in their 2018 Workforce Report: “Nationally, we have a shortage of 151,717 people with data science skills, with particularly acute shortages in New York City (34,032 people), the San Francisco Bay Area (31,798 people), and Los Angeles (12,251 people).”

Top 7 Big Data Jobs

1.25.1. Data Scientist

According to PayScale, there are plentiful opportunities for talented information technology (IT) data scientists capable of mining and interpreting complex data for large corporations. Partnering with cross-functional IT teams, they compile and create various models of statistical data to inform their development of systems-related recommendations and action plans.

Preferred Qualifications - IT data scientists should have advanced knowledge of different data mining techniques such as clustering, regression analysis, decision trees, and support vector machines. An advanced degree (such as a Master's or Ph.D.) in computer science is usually required for this kind of position, in addition to previous years of work experience in a related field.

Salary: Cited by Glassdoor as #1 in their 50 Best Jobs in America 2019 list, data scientists can demand an average salary of $108,000. An overall job satisfaction score of 4.3/5 was a key factor in this top rating.

1.25.2. Data Engineer

Data engineers, as PayScale points out, utilize their computer science and engineering strengths to aggregate, analyze, and manipulate massive data sets. Common tasks include creating and translating computer algorithms into prototype code, developing technical processes to improve

data accessibility, and designing reports, dashboards, and tools for end-users.

Preferred Qualifications – Employers typically require that job candidates for data engineering positions have successfully completed a college degree in computer science, engineering, or a related field. They also prefer applicants who have three to five years' experience in the field. Desired technical proficiencies include Linux systems knowledge, proficiency in SQL database design, and a solid command of coding languages such as Java, Python, Kafka, Hive, or Storm. Soft skills include excellent written and verbal communication capabilities as well as the ability to work both independently and on teams.

Salary: Glassdoor's 50 Best Jobs in America 2019 ranked data engineers in 8th position with an average salary of $106,000 and job satisfaction score of 3.9/5.

1.25.3. Data Analyst

PayScale describes how data analysts gather actionable information about various topics by designing and implementing large-scale surveys. It's their job to recruit survey participants, compile and interpret submitted data, and relay their findings in traditional charts and reports as well as digital formats.

Preferred Qualifications - Individuals looking for data analyst jobs must be knowledgeable in computer programs such as Microsoft Excel, Microsoft Access, SharePoint, and SQL databases. Data analysts also need good

communication and presentation skills, with the ability to effectively translate often-complicated information to company stakeholders.

Salary: Data Analysts, ranked 31st on 50 Best Jobs in America 2019, take home an average salary of $60,000, and have a job satisfaction score of 3.9/5.

1.25.4. Security Engineer

Security engineers play a vital role in IT disaster planning, aversion, and mitigation. They lessen corporate risk exposure by setting up computer firewalls, detecting and responding to intrusions, and forensically pinpointing systems security issues. They also create and implement test plans for new or updated software and hardware and establish multi-layered defense protocols for computer networks. (Source: PayScale).

Preferred Qualifications - A bachelor's degree in engineering, computer science, or a related field is needed for this position, along with several years of relevant work experience and, ideally, industry security certifications. In addition to their technical understanding of computer languages and operating systems, security engineers should also possess solid problem-solving and mathematical abilities and the capacity to work independently.

Salary: Security engineers ranked 17th on 50 Best Jobs in America 2019, with an average salary of $102,000 and a job satisfaction score of 3.8/5.

1.25.5. Database Manager

Database managers, trained and highly skilled in project management and multitasking, perform diagnostics and repair of sophisticated databases. They also review business requests for data and data usage, evaluate data sources to improve data feeds and help to design and install storage hardware. (Source: PayScale)

Preferred Qualifications – Job advertisements for database managers generally list a bachelor's degree in information technology and a minimum of five years in a database leadership position as preferred qualifications. Job candidates for database management roles should also be well-versed in different database software such as MySQL and Oracle.

Salary: Database managers earned an average of $73,545 according to Glassdoor.

1.25.6. Data Architect

Data architects use their knowledge of data-oriented computer languages to organize and maintain data in relational databases and corporate repositories, developing data architecture strategies for each subject area of the enterprise data model. (Source: PayScale)

Preferred Qualifications – Common job skills and keyword attributes that employers seek in data architects include advanced technical proficiency (particularly in languages like SQL and XML), excellent analytical acumen, creative

visualization, and problem-solving skills, and strong detail orientation. Most data architects have earned at least a bachelor's degree (and, often, an advanced degree) in a computer science-related field.

Salary: Data architects earned some of the highest salaries in the big data sector, an average of $113,078 according to PayScale.

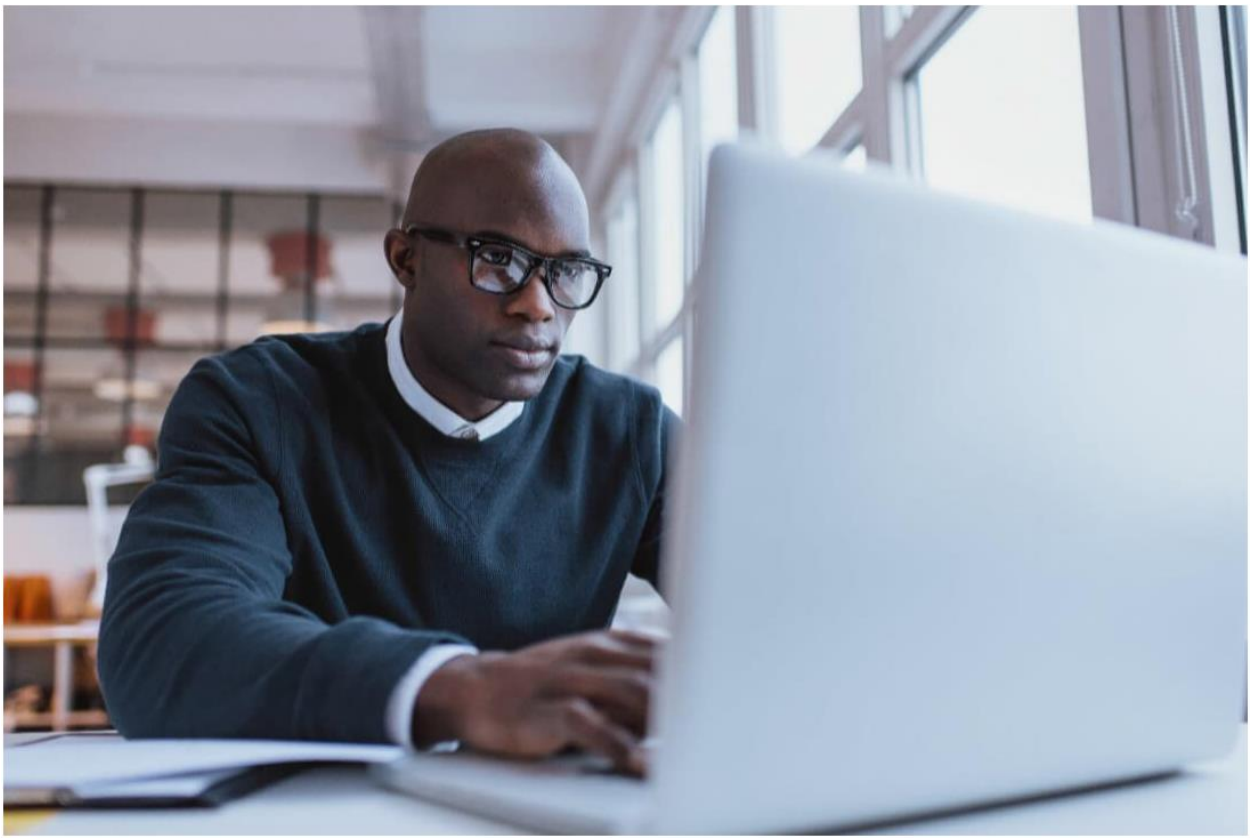
1.25.7. Technical Recruiter

Technical recruiters specialize in sourcing and screening talented big data, IT, and other technical professionals for candidate pools. They work closely with corporations to assess their hiring needs, and then search the market for the strongest candidates for specific job openings. They also support the professional candidates they recruit throughout the job application, interviewing, hiring, and onboarding processes. (Source: PayScale).

Preferred Qualifications - Technical recruiters need to have advanced knowledge of the technical qualifications that employers seek in their prospective personnel. Good "people skills" are also essential, since successful technical recruiters need to build a positive rapport with job candidates during the screening and interviewing stages.

Salary: Glassdoor ranked technical recruiters as #28 on their List of the 50 Best Jobs in America, with a job satisfaction score of 4.1/5 and an average salary of $48,000.

1.26 10 High-Paying Jobs That Require a Knowledge of Data Analytics



There’s never been a better time to learn data analytics and enter the workforce as a data scientist. The job landscape is promising, opportunities span multiple industries and the nature of the job often allows for remote work flexibility and even self-employment.

Plus, many data analytics experts boast a high median salary, even at entry-level positions.

With technology reaching new heights and a majority of the population having access to an internet connection, there’s no

denying that Big Data and data analytics have become hot topics in recent years – and a growing need. **According to IBM, the number of jobs for data professionals in the U.S will increase to 2,720,000 by 2020.**
Demand for knowledgeable data analytics professionals currently outweighs the supply, meaning that companies are willing to pay a premium to fill their open job positions.

But the skillset and job opportunities within data science go beyond the tech and digital spaces. Let's take a look at what you need to know as a data scientist – and what you'll learn when you take our courses.

1.27 What Skills Are Required for a Job in Data Analytics?

As you delve into the 10 jobs we have here and start applying for positions in the data analytics field, you'll notice many of them require the same foundational skills. Make sure you've mastered these before you start sending your cover letter and portfolio to potential employers.

And, if you find a skill that you still need to learn, remember that you can take an affordable, self-paced data science course that will help you learn everything you need to know for a successful career in data science.

1.27.1 Python

Python is currently one of the most commonly used programming languages.

Having a solid understanding of how to use Python for data analytics will probably be required for many roles. Even if it's not a required skill, knowing and understanding Python will give you an upper hand when showing future employers the value that you can bring to their companies.

If you're ready to advance your programming language proficiency, learn how to manipulate and analyze data, understand the concept of web scraping and data collection, and start building web applications, consider enrolling in our Python for Data Science: Fundamentals Course.

1.27.2 SQL (Structured Query Language)

Working with data sources is a necessary aspect of data analytics.

Early in your career, you'll need at least a basic understanding of SQL. SQL (pronounced sequel) is often a major component of these positions. When you go to interview, listen for hiring managers' mentions of this programming language when asking about your work with databases.

The experience you'll get in our SQL courses will give you a good foundation. Like Python, SQL is a relatively easy language to start

learning. Even if you are just getting started, a little SQL experience goes a long way.

Knowing the basics of SQL will give you the confidence to navigate large databases, and to obtain and work with the data you need for your projects. You can always seek out opportunities to continue learning once you get your first job.

1.27.3 Data Visualization Skills

Knowing how to visualize data and communicate results is a huge competitive edge for job seekers.

On the job market, these skillsets have high demand (and high pay)! Regardless of the career path you're looking into, being able to visualize and communicate insights related to your company's services and bottom line is a valuable skillset that will turn the heads of employers.

In this way, data scientists are a bit like data translators for other people in the organization that aren't sure what conclusions to draw from their datasets.

At Dataquest, students are equipped with specific knowledge and skills for data visualization in Python and R using data science and visualization libraries.

1.28 10 Jobs that Require a Knowledge of Data Analytics

Before you take the time to learn a new skillset, you'll likely be curious about the earning potential of related positions. Knowing how your new skills will be rewarded gives you the proper motivation and context for learning.

Lots of employers are hiring for these positions, both remote and onsite, worldwide. Here are a few positions worth looking into – and their median incomes, according to popular job search websites.

1.28.1 IT Systems Analyst

Systems analysts use and design systems to solve problems in information technology.

The required level of technical expertise varies in these positions, and that creates opportunities for specialization by industry and personal interests. Some systems analysts use existing third-party tools to test software within a company, while others develop new. proprietary tools from their understanding of data analytics and the business itself.

A typical systems analyst in the US makes around $68,807.

1.28.2 Healthcare Data Analyst

Healthcare data analysts have the opportunity to improve the quality of life for many people by helping doctors and scientists find answers to the questions and problems they encounter on a daily basis.

The amount of data coming from the healthcare industry is growing rapidly, be it with the increased popularity of wearables like Apple Watch or through enhanced medical testing in clinics, hospitals and labs. Plus, with a rise in regulations and restrictions on how that data can be stored, retrieved, and processed, demand for proficient data analysts is on the rise as well.

The median salary for a Healthcare Data Analyst is $61,438.

1.28.3 Operations Analyst

Operations analysts are usually found internally at large companies, but may also work as consultants.

Operations analysts focus on the internal processes of a business. This can include internal reporting systems, product manufacturing and distribution, and the general streamlining of business operations.

It's more important for professionals in these roles to have general business savvy, and they often have technical knowledge of the systems they're working with. Operations analysts are found in every type of business, from large grocery chains, to postal service providers, to the military and can make upwards of $75,000 annually. Due to the versatile nature of this data analytics job and the many industries you may find employment in, the salary can vary widely.

1.28.4 Data Scientist

Much like analysts in other roles, data scientists collect and analyze data and communicate actionable insights. Data scientists are often a technical step above of data analysts, though. They are the ones who are able to understand data from a more informed perspective to help make predictions. These positions require a strong knowledge of data analytics including software tools, programming languages like Python or R, and data visualization skills to better communicate findings.

These positions are challenging – and rewarding, with an average salary of $91,494. The demand for data analytics experts with technical backgrounds is at an all-time high.
Dataquest has multiple learning paths that are tailored to provide you with everything you need to hone your technical skills, including the Data Scientist Path that will help you become a certified data scientist.

1.28.5 Data Engineer

Data engineers often focus on larger datasets and are tasked with optimizing the infrastructure surrounding different data analytics processes.

For example, a data engineer might focus on the process of capturing data to make an acquisition pipeline more efficient. They may also need to upgrade a database infrastructure for faster queries. These high-level data analytics professionals are also well-paid, with median salaries being comparable to data scientists at $90,963.

1.28.6 Quantitative Analyst

A quantitative analyst is another highly sought-after professional, especially in financial firms. Quantitative analysts use data analytics to seek out potential financial investment opportunities or risk management problems.

The median salary for quantitative analysts is $82,879. They may also venture out on their own, creating trading models to predict

the prices of stocks, commodities, exchange rates, etc. Some analysts in this industry even go on to open their own firms.

1.28.7 Data Analytics Consultant

Like many of these positions, the primary role of an analytics consultant is to deliver insights to a company to help their business. While an analytics consultant may specialize in any particular industry or area of research, the difference between a consultant and an in-house data scientist or data analyst is that a consultant may work for different companies in a shorter period of time.

They may also be working for more than one company at a time, focusing on particular projects with clear start and end dates.

These positions are best for those who like change, and those who have a narrowed interest and expertise in a field of study. Analytics consultants are also in a great position to work remotely, another enticing factor about this role to consider.

Compensation varies widely by industry, but $78,264 is a representative salary for the role.

1.28.8 Digital Marketing Manager

Digital marketing also requires a strong knowledge of data analytics. Depending on your other complementary skills and interests, you could find yourself in a specific analytics role within a company or agency, or simply applying your data science expertise as a part of a larger skillset.

Marketers often use tools like Google Analytics, custom reporting tools and other third party sites to analyze traffic from websites and social media advertisements. While these examples require a basic understanding of data analytics, a skilled data scientist has the ability to create a long-term career in marketing.

A lot of money could be wasted on campaigns that do not drive traffic, so marketing professionals will continue to need analysts to make smart decisions about how to leverage existing resources.

While digital marketing positions have a wide range, the higher end salary for an advanced digital marketing manager is $97,000.

1.28.9 Project Manager

Project managers use analytics tools to keep track of a team's progress, track their efficiency, and increase productivity by changing processes.

Project managers need at least a working understanding of data analytics, and often more.

These positions are found internally at large corporations, and frequently in management consulting. Another example of a career trajectory for project managers could be moving into product and supply chain management, which companies rely on to maintain profit margins and smooth operations.

A typical salary for a project manager is around $73,247.

1.28.10 Transportation Logistics Specialist

A transportation logistics specialist optimizes transportation of physical goods, and could be found in large shipping companies, like Amazon, UPS, naval transport companies, airlines and city planning offices.

A data analytics background is especially helpful in this job because transportation logistics specialists need to reliably identify the most efficient paths for products and services to be delivered. They must look at large amounts of data to help identify and eliminate bottlenecks in transit, be it on land, sea or in the air.

With seasoned professionals in this industry making around $79,000 per year, a transportation logistics specialist is an appealing career path for individuals who are detail-oriented, technical and forward thinkers.
A data analytics background also helps transportation logistics specialists, among others, to focus on the most important issues, seeing potential problems and solutions in context and communicating those effectively.

1.29Data Analytics Opportunities around the Globe

These are just a few of the many high-paying jobs which require knowledge of data analytics. Specific figures from this article are for the median salaries in the United States, all cities included.

Salaries in each city may vary and reflect local demand and general cost-of-living expenses. Boston, Portland, and Denver, for example, have become hotspots for data analytics positions. While the numbers included in this article represent a typical salary in the United States, opportunities for data analytics professionals can be found all around the globe. Many of them can even be done remotely, allowing you the highly-desired opportunity to work from anywhere in the world on a competitive US salary.

Whether your goal is to get a full-time job in a new industry, advance in your existing career, or work for yourself in the data analytics field, Dataquest can prepare you for the opportunity. With the portfolio-building missions and projects in Dataquest's Data Analyst path, a community of mentors, and a strong alumni network, you'll have all you need to become a certified data analyst and be set up to get the job of your dreams.

How Much Do Data Scientists Really Earn?

- Posted by Bernard Marr on June 28, 2015 at 6:00am

- View Blog

Last year, there was much made of an unconfirmed report that a developer at Google was making $3 million a year. It made many people speculate that the high demand for data scientists coupled with the low supply was creating a salary bubble.

IT World couldn't confirm the $3 million developer story, but it did some number crunching from publically available data and determined that the average developer at Google makes around $145,000 per year, including stock and bonuses, while the highest paid senior engineer might make as much as $1 million including stocks and bonuses.
But that's Google. What can a data scientist expect to make elsewhere?

According to Glassdoor, the current U.S. average salary for a data scientist is $118,709, but it varies widely based on a number of factors.

One major factor that will determine salary levels is your actual title and relevant job skills and responsibilities.

- **Database Administrator (DBA): $50,000–$120,000**
  DBAs are responsible for the upkeep of the databases that store all that data. Salaries depend on the person's level of experience and the complexity of the system being maintained.
- **Data Analyst: $50,000–$110,000**
  A data analyst is a quant-focused professional and generally has a BS or MS degree. Entry-level analysts can expect to make $50k and up, while senior level analysts can reach six figures.
- **Data Scientist: $85,000–$170,000**
  A data scientist is an experienced, expert-level professional (there's no such thing as an entry-level data scientist) and are paid accordingly. Some of the variants in pay come from the topics and applications in which the person is well versed.
- **Data Scientist/Analyst Manager: $90,000–$240,000**
  When a person rises to the level of manager, pay is often based on the number of people directly reporting to that manager. Low end would be 1–3 subordinates, while the high end would be 10 or more employees.
- **Data Engineer**: **$70,000–$165,000**
  The data engineer is the person designing the systems that record and store the data. They must

have a wealth of programming knowledge as well as expertise in informational architecture.

Another factor is the type of company you work for. According to an O'Reilly survey from 2013:

- Startups pay the most, with a median salary of $130k.
- Public companies paid a median salary of $110k.
- Government and education sector jobs were the lowest at $80k.

Glassdoor also has interesting data from anonymously reported salaries that shows averages at different companies. For example, data scientists at Groupon right now are reportedly earning the most at an average of $164k while Booz Allen Hamilton is paying closer to $94k for the same job title.

Data Scientist Salaries | More details for Data Scientist Jobs | Salary

Likewise, the type and number of tools a person uses on the job can increase the salary range. Salaries show data scientists who use the Hadoop cluster got paid more; average salary for those who don't use it was $85k compared to $125k for those who use at least six tools or more.

Salaries depend on so many factors that it's impossible to point to a single number and say, "This is what you should be earning." But it can be helpful if you are a data scientist (or are considering the field) to understand where your own salary fits into national averages.

Are you a data scientist or do you employ one? How does this information tally with your experience in the market? I'd love to hear your real-world experiences in the comments below.

As always, I am always keen to hear your views on the topic and invite you to comment with any thoughts you might have.

About : Bernard Marr is a globally recognized expert in analytics and big data. He helps companies manage, measure, analyze and improve performance using data. His new book is: Big Data: Using Smart Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance

Chapter 2:

Barriers, Obstacles, Challenges and Opportunities with Big Data!

Challenges and Opportunities with Big Data

A community white paper developed by leading researchers across the United State:

This white paper was created through a distributed conversation among many prominent researchers listed below. This conversation lasted a period of approximately three months from Nov. 2011 to Feb. 2012. Collaborative writing was supported by a distributed document editor.

Divykant Agrawal, UC Santa Barbara

Philip Bernstein, Purdue University

Susan Davidson, University Of Pennsylvania

Umeshwar Dayal, HP

Michael Franklin, UC Berkeley

Johannes Gehrke, Cornell University

Laura Haas, IBM

Alon Halevy, Google

Jiawei Han, UIUC

H. V. Jagadish, University of Michigan (Coordinator)

Alexadros Labrinidis, University of Pittsburgh

Sam Madden, MIT

Yannis Papkonstantinou, UC San Diego

Jignesh M. Patel, University of Wisconsin

Raghu Ramakrishnan, Yahoo!

Keneth Ross, Columbia University

Cyrus Shahabi, University of Southern California

Dan Suciu, University of Washington

Shiv Vaithyanathan, IBM

Jennifer Widom, Stanford University

1. Introduction

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

Scientific research has been revolutionized by Big Data [CCC2011a]. The Sloan Digital Sky Survey [SDSS2008] has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is

an entire discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.

Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance.

It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [CCC2011c], by making care more preventive and personalized and basing it on more extensive (home-based) continuous

monitoring. McKinsey estimates [McK2011] a savings of 300 billion dollars every year in the US alone.

In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data) [CCC2011d], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative [MGI2011]), computational social sciences (a new methodology fast growing in popularity because of the dramatically lowered cost of obtaining data) [LP+2009], financial systemic risk analysis (through integrated analysis of a web of contracts to find dependencies between financial entities) [FJ+2011], homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged information and other events, known as Security Information and Event Management (SIEM)), and so on.

In 2010, enterprises and users stored more than 13 exabytes of new data; this is over 50,000 times the data in the Library of

Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [PCAST2010] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b].

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just

the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. For example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all be laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user

interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it [NYT2012]. We can drastically increase this number by supporting

many levels of engagement with the data, not all requiring deep database expertise.

Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis.

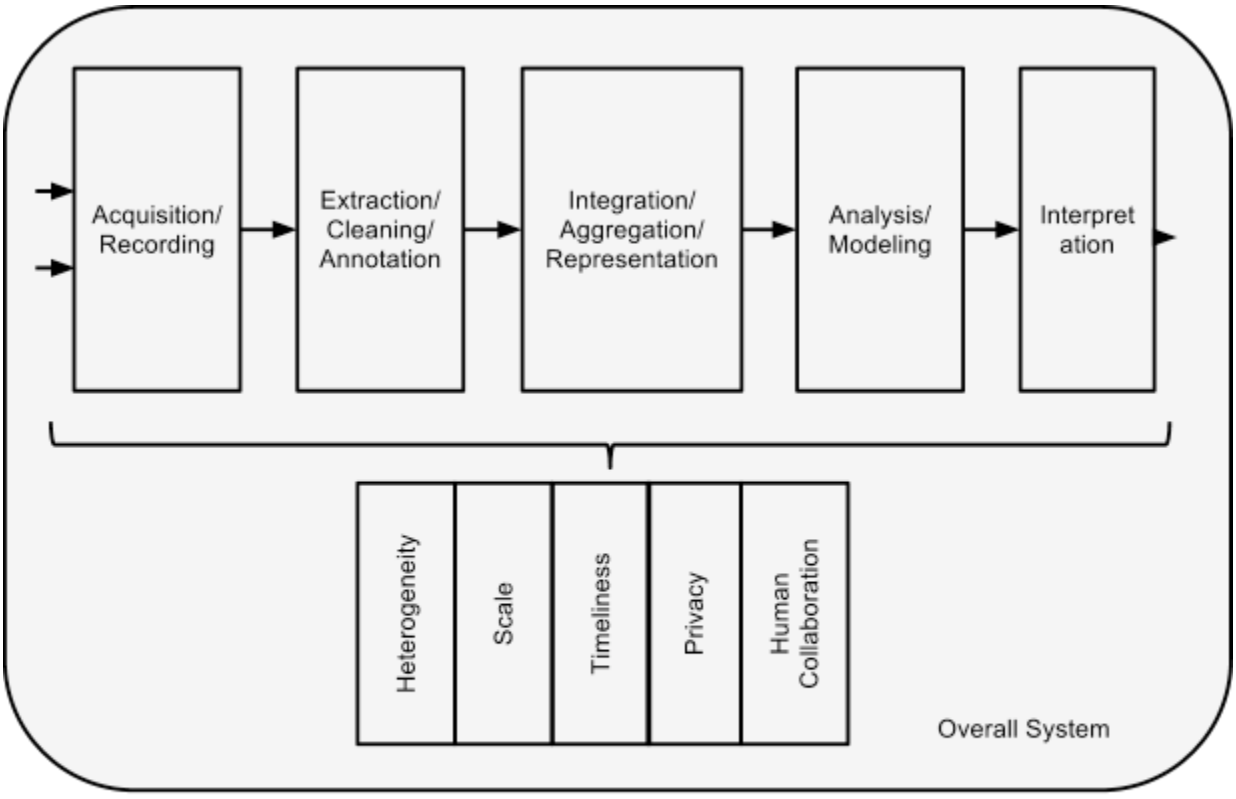


Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Fortunately, existing computational techniques can be applied, either as is or with some extensions, to at least some aspects of the Big Data problem. For example, relational databases rely on the notion of logical data independence: users can think about what they want to compute, while the system (with skilled engineers designing those systems) determines how to compute it efficiently. Similarly, the SQL standard and the relational data model provide a uniform, powerful language to express many query needs and, in principle, allows customers to choose between vendors, increasing competition. The challenge ahead of us is to combine these healthy features of prior systems as we devise novel solutions to the many new challenges of Big Data.

In this paper, we consider each of the boxes in the figure above, and discuss both what has already been done and what challenges remain as we seek to exploit Big Data. We begin by considering

the five stages in the pipeline, then move on to the five cross-cutting challenges, and end with a discussion of the architecture of the overall system that combines all these functions.

2. Phases in the Processing Pipeline

Data Acquisition and Recording

Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce petabytes of data today.

Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artifact that deserves attention? In addition, the data collected by these sensors most often are spatially and temporally correlated (e.g., traffic sensors on the same road segment). We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack. Furthermore, we require "on-line" analysis

techniques that can process such streaming data on the fly, since we cannot afford to store first and reduce afterward.

The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. For example, in scientific experiments, considerable detail regarding specific experimental conditions and procedures may be required to be able to interpret the results correctly, and it is important that such metadata be recorded with observational data. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data provenance. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. For example, a processing error at one step can render subsequent analysis useless; with suitable provenance, we can easily identify all subsequent processing that dependent on this step. Thus we need research both into generating suitable metadata and into data systems that carry the provenance of data and its metadata through data analysis pipelines.

Information Extraction and Cleaning

Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of

electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements (possibly with some associated uncertainty), and image data such as x-rays. We cannot leave the data in this form and still effectively

analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge. Note that this data also includes images and will in the future include video; such extraction is often highly application dependent (e.g., what you want to pull out of an MRI is very different from what you would pull out of a picture of the stars, or a surveillance photo). In addition, due to the ubiquity of surveillance cameras and popularity of GPS-enabled mobile phones, cameras, and other portable devices, rich and high fidelity location and trajectory (i.e., movement in space) data can also be extracted.

We are used to thinking of Big Data as always telling us the truth, but this is actually far from reality. For example, patients may choose to hide risky behavior and caregivers may sometimes mis-diagnose a condition; patients may also inaccurately recall the name of a drug or even that they ever took it, leading to missing information in (the history portion of) their medical record. Existing work on data cleaning assumes well-recognized

constraints on valid data or well-understood error models; for many emerging Big Data domains these do not exist.

Data Integration, Aggregation, and Representation

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other

purposes. Witness, for instance, the tremendous variety in the structure of bioinformatics databases with information regarding substantially similar entities, such as genes. Database design is today an art, and is carefully executed in the enterprise context by highly-paid professionals. We must enable other professionals, such as domain scientists, to create effective database designs, either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so that databases can be used effectively in the absence of intelligent database design.

Query Processing, Data Modeling, and Analysis

Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. As noted previously, real-life medical records have errors, are heterogeneous, and frequently are distributed across multiple systems. The value of Big Data analysis in health care, to take just one example application domain, can only be realized if it can be applied

robustly under these difficult conditions. On the flip side, knowledge developed from data can help in correcting errors and removing ambiguity. For example, a physician may write "DVT" as the diagnosis for a patient. This abbreviation is commonly used for both "deep vein thrombosis" and "diverticulitis," two very different medical conditions. A knowledge-base constructed from related data can use associated symptoms or medications to determine which of two the physician meant.

Big Data is also enabling the next generation of interactive data analysis with real-time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, to populate hot-lists or recommendations, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today.

A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process and bringing the data back. This is an obstacle to carrying over the interactive elegance of the first generation of SQL-

-driven OLAP systems into the data mining type of analysis that is in increasing demand. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

Interpretation

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results.

This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision-maker diligence — rather than accept the stated solvency of a financial institution at face value, a decision-maker has to examine critically the many assumptions at multiple stages of analysis.

In short, it is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data. By studying how best to capture, store, and query provenance, in conjunction with techniques to capture adequate metadata, we can create an infrastructure to provide users with the ability both to interpret analytical results obtained and to repeat the analysis with different assumptions, parameters, or data sets.

Systems with a rich palette of visualizations become important in conveying to the users the results of the queries in a way that is best understood in the particular domain. Whereas early business intelligence systems' users were content with tabular presentations, today's analysts need to pack and present results in powerful visualizations that assist interpretation, and support user collaboration as discussed in Sec. 3.5.

Furthermore, with a few clicks the user should be able to drill down into each piece of data that she sees and understand its

provenance, which is a key feature to understanding the data. That is, users need to be able to see not just the results, but also understand why they are seeing those results. However, raw provenance, particularly regarding the phases in the analytics pipeline, is likely to be too technical for many users to grasp completely. One alternative is to enable the users to "play" with the steps in the analysis – make small changes to the pipeline, for example, or modify values for some parameters. The users can then view the results of these incremental changes. By these means, users can develop an intuitive feeling for the analysis and also verify that it performs as expected in corner cases. Accomplishing this requires the system to provide convenient facilities for the user to specify analyses. Declarative specification, discussed in Sec. 4, is one component of such a system.

3. Challenges in Big Data Analysis

Having described the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases. These are shown as five boxes in the second row of Fig. 1.

Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data

require further work.

Consider an electronic health record database design that has fields for birth date, occupation, and blood type for each patient. What do we do if one or more of these pieces of information is not provided by a patient? Obviously, the health record is still placed in the database, but with the corresponding attribute values being set to NULL. A data analysis that looks to classify patients by, say, occupation, must take into account patients for which this information is not known. Worse, these patients with unknown occupations can be ignored in the analysis only if we have reason to believe that they are otherwise statistically similar to the patients with known occupation for the analysis performed. For example, if unemployed patients are more likely to hide their employment status, analysis results may be skewed in that it considers a more employed population mix than exists, and hence potentially one that has differences in occupation-related health-profiles.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress.

Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But,

there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

First, over the last five years the processor technology has made a dramatic shift --- rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes don’t directly apply for intra-node parallelism, since the architecture looks very different; for example, there are many

more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node. Furthermore, the move towards packing multiple sockets (each with 10s of cores) adds another level of complexity for intra-node parallelism. Finally, with predictions of "dark silicon", namely that power consideration will likely in the future prohibit us from using all of the hardware in the system continuously, data processing systems will likely have to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how we design, build and operate data processing components.

The second dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger

clusters (that are required to deal with the rapid growth in data volumes). This places a premium on declarative approaches to expressing programs, even those doing complex machine learning tasks, since global optimization across multiple users' programs is necessary for good overall performance. Reliance on user-driven program optimizations is likely to lead to poor cluster utilization, since users are unaware of other users' programs. System-driven holistic optimization requires programs to be sufficiently transparent, e.g., as in relational database systems, where declarative query languages are designed with this in mind.

A third dramatic shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, hard disk drives (HDDs) were used to store persistent data. HDDs had far slower random IO performance than sequential IO performance, and data processing engines formatted their data and designed their query processing methods to "work around" this limitation. But, HDDs are increasingly being replaced by solid state drives today, and other technologies such as Phase Change Memory are around the corner. These newer storage technologies do not have the same large spread in performance between the sequential and random I/O performance, which requires a rethinking of how we design storage subsystems for data processing systems. Implications of this changing storage subsystem potentially touch every aspect of data processing,

including query processing algorithms, query scheduling, database design, concurrency control methods and recovery methods.

Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge as described in Sec. 2.1, and a timeliness challenge described next.

There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding

qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from

both perspectives to realize the promise of big data.

Consider, for example, data gleaned from location-based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address these privacy concerns. An attacker or a (potentially malicious) location-based server can infer the identity of the query source from its (subsequent) location information. For example, a user's location information can be tracked through several stationary connection points (e.g., cell towers). After a while, the user

leaves "a trail of packet crumbs" which could be associated to a certain residence or office location and thereby used to determine the user's identity. Several other types of surprisingly private information such as health issues (e.g., presence in a cancer treatment center) or religious preferences (e.g., presence in a church) can also be revealed by just observing anonymous users' movement and usage pattern over time. In general, Barabási et al. showed that there is a close correlation between people's identities and their movement patterns [Gon2008]. Note that hiding a user location is much more challenging than hiding his/her identity. This is because with location-based services, the location of the user is needed for a successful data access or data collection, while the identity of the user is not necessary.

There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but gets larger and changes over time; none of the prevailing techniques results in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. Wikipedia, the online encyclopedia, is perhaps the best known example of crowd-sourced data. We are relying upon information provided by

unvetted strangers. Most often, what they say is correct. However, we should expect there to be individuals who have other motives and abilities – some may have a reason to provide false information in an intentional attempt to mislead. While most such errors will be detected and corrected by others in the crowd, we need technologies to facilitate this.

We also need a framework to use in analysis of such crowd-sourced data with conflicting statements. As humans, we can look at reviews of a restaurant, some of which are positive and others critical, and come up with a summary assessment based on which we can decide whether to try eating there. We need computers to be able to do the equivalent. The issues of uncertainty and error become even more pronounced in a specific type of crowd-sourcing, termed participatory-sensing. In this case, every person with a mobile phone can act as a multi-modal sensor collecting various types of data instantaneously (e.g., picture, video, audio, location, time, speed, direction, acceleration). The extra challenge here is the inherent uncertainty of the data collection devices. The fact that collected data are probably spatially and temporally correlated can be exploited to better assess their correctness. When crowd-sourced data is obtained for hire, such as with "Mechanical Turks," much of the data created may be with a primary objective of getting it done quickly rather than correctly. This is yet another error model, which must

be planned for explicitly when it applies.

4. System Architecture

Companies today already use, and appreciate the value of, business intelligence. Business data is analyzed for many purposes: a company may perform system log analytics and social media analytics for risk assessment, customer retention, brand management, and so on. Typically, such varied tasks have been handled by separate systems, even if each system includes common steps of information extraction, data cleaning, relational-like processing (joins, group-by, aggregation), statistical and predictive modeling, and appropriate exploration and visualization tools as shown in Fig. 1.

With Big Data, the use of separate systems in this fashion becomes prohibitively expensive given the large size of the data sets. The expense is due not only to the cost of the systems themselves, but also the time to load the data into multiple systems. In consequence, Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads. The challenge here is not to build a system that is ideally suited for all processing tasks. Instead, the need is for the underlying system architecture to be flexible enough that the components built on top

of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads. The effects of scale on the physical architecture were considered in Sec 3.2. In this section, we focus on the programmability requirements.

If users are to compose and build complex analytical pipelines over Big Data, it is essential that they have appropriate high-level primitives to specify their needs in such flexible systems. The Map-Reduce framework has been tremendously valuable, but is only a first step. Even declarative languages that exploit it, such as Pig Latin, are at a rather low level when it comes to complex analysis tasks. Similar declarative specifications are required at higher levels to meet the programmability and composition needs of these analysis pipelines. Besides the basic technical need, there is a strong business imperative as well. Businesses typically will outsource Big Data processing, or many aspects of it. Declarative specifications are required to enable technically meaningful service level agreements,

since the point of the out-sourcing is to specify precisely what task will be performed without going into details of how to do it.

Declarative specification is needed not just for the pipeline composition, but also for the individual operations themselves. Each operation (cleaning, extraction, modeling etc.) potentially

runs on a very large data set. Furthermore, each operation itself is sufficiently complex that there are many choices and optimizations possible in how it is implemented. In databases, there is considerable work on optimizing individual operations, such as joins. It is well-known that there can be multiple orders of magnitude difference in the cost of two different ways to execute the same query. Fortunately, the user does not have to make this choice – the database system makes it for her. In the case of Big Data, these optimizations may be more complex because not all operations will be I/O intensive as in databases. Some operations may be, but others may be CPU intensive, or a mix. So standard database optimization techniques cannot directly be used. However, it should be possible to develop new techniques for Big Data operations inspired by database techniques.

The very fact that Big Data analysis typically involves multiple phases highlights a challenge that arises routinely in practice: production systems must run complex analytic pipelines, or workflows, at routine intervals, e.g., hourly or daily. New data must be incrementally accounted for, taking into account the results of prior analysis and pre-existing data. And of course, provenance must be preserved, and must include the phases in the analytic pipeline. Current systems offer little to no support for such Big Data pipelines, and this is in itself a challenging objective.

5. Conclusion

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

Bibliography

[CCC2011a] Advancing Discovery in Science and Engineering. Computing Community Consortium.
Spring 2011.

[CCC2011b] Advancing Personalized Education. Computing

Community Consortium. Spring 2011. [CCC2011c] Smart

Health and Wellbeing. Computing Community Consortium.

Spring 2011. [CCC2011d] A Sustainable Future.

Computing Community Consortium. Summer 2011.

[DF2011] Getting Beneath the Veil of Effective Schools: Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.

[Eco2011] Drowning in numbers ------ Digital data will flood the planet—and help us understand it better. *The Economist*, Nov 18, 2011. http://www.economist.com/blogs/dailychart/2011/11/big-data-0

[FJ+2011] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.

[Gar2011] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916

[Gon2008] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. *Nature* 453, 779-782 (5 June 2008)

[LP+2009] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer,Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. *Science* 6 February 2009: **323** (5915), 721-723.

[McK2011] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[MGI2011] Materials Genome Initiative for Global Competitiveness. National Science and Technology Council. June 2011.

[NPR2011a] Folowing the Breadcrumbs to Big Data Gold. Yuki Noguchi. *National Public Radio*, Nov. 29, 2011. http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data

[NPR2011b] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. *National Public Radio*, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data

[NYT2012] The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[PCAST2010] Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. PCAST Report, Dec. 2010. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf

[SDSS2008] SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. Jan. 2008. Available at http://www.sdss3.org/collaboration/description.pd

Chapter 3:

An Introduction to Big Data Concepts and Terminology

Introduction

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

Weird Definitions of Big Data Definitions: What's Yours?

Yesterday I got an email from UC Berkeley's Master of Information and Data Science program, asking me to respond to a survey of data science thought leaders, asking the question "What is big data"? I was especially delighted to be regarded as a "thought leader" by Berkeley's School of Information, whose previous dean, Hal Varian (now chief economist at Google GOOGL +0%), answered my challenge fourteen years ago and produced the first study to estimate the amount of new information created in the world annually, a study I consider to be a major milestone in the evolution of our understanding of big data.

The Berkeley researchers estimated that the world had produced about 1.5 billion gigabytes of information in 1999 and in a 2003 replication of the study found out that amount to have doubled in 3 years. Data was already getting bigger and bigger and around that time, in 2001, industry analyst Doug Laney described the "3Vs"—volume, variety, and velocity—as the key "data management challenges" for enterprises, the same "3Vs" that have been used in the last four years by just about anyone attempting to define or describe big data.

Source: Wikimedia Commons

The first documented use of the term “big data” appeared in a 1997 paper by scientists at NASA, describing the problem they had with visualization (i.e. computer graphics) which “provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.”

In 2008, a number of prominent American computer scientists popularized the term, predicting that “big-data computing” will “transform the activities of companies, scientific researchers, medical practitioners, and our nation’s defense and intelligence operations.” The term “big-data computing,” however, is never defined in the paper.

The traditional database of authoritative definitions is, of course, the *Oxford English Dictionary* (OED). Here's how the OED defines big data: (definition #1) "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges."

But this is 2014 and maybe the first place to look for definitions should be Wikipedia. Indeed, it looks like the OED followed its lead. Wikipedia defines big data (and it did it before the OED) as (#2) "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications."

While a variation of this definition is what is used by most commentators on big data, its similarity to the 1997 definition by the NASA researchers reveals its weakness. "Large" and "traditional" are relative and ambiguous (and potentially self-serving for IT vendors selling either "more resources" of the "traditional" variety or new, non-"traditional" technologies).

The widely-quoted 2011 big data study by McKinsey highlighted that definitional challenge. Defining big data as (#3) "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze," the McKinsey researchers acknowledged that "this definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data." As a result, all the quantitative insights of the study, including the updating of the UC Berkeley numbers by estimating how much new data is stored by enterprises and consumers annually, relate to digital data, rather than just big data, e.g., no attempt was made to estimate how much of the data (or "datasets") enterprises store is big data.

Another prominent source on big data is Viktor Mayer-Schönberger and Kenneth Cukier's book on the subject. Noting that "there is no rigorous definition of big data," they offer one that points to what can be done with the data and why its size matters:

(#4) “The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value” and “…things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value.”

In *Big Data@Work*, Tom Davenport concludes that because of “the problems with the definition” of big data, “I (and other experts I have consulted) predict a relatively short life span for this unfortunate term.” Still, Davenport offers this definition:

(#5) “The broad range of new and massive data types that have appeared over the last decade or so.”

Let me offer a few other possible definitions:

(#6) The new tools helping us find relevant data and analyze its implications.

(#7) The convergence of enterprise and consumer IT.

(#8) The shift (for enterprises) from processing internal data to mining external data.

(#9) The shift (for individuals) from consuming data to creating data.

(#10) The merger of Madame Olympia Maxima and Lieutenant Commander Data.

#(11) The belief that the more data you have the more insights and answers will rise automatically from the pool of ones and zeros.

#(12) A new attitude by businesses, non-profits, government agencies, and individuals that combining data from multiple sources could lead to better decisions.

I like the last two. #11 is a warning against blindly collecting more data for the sake of collecting more data (see NSA). #12 is an

acknowledgment that storing data in “data silos” has been the key obstacle to getting the data to work for us, to improve our work and lives. It’s all about attitude, not technologies or quantities.

What’s your definition of big data?

What Is Big Data?

An exact definition of “big data” is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, generally speaking, **big data** is:

- large datasets
- the category of computing strategies and technologies that are used to handle large datasets

In this context, “large dataset” means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

Why Are Big Data Systems Different?

The basic requirements for working with big data are the same as the requirements for working with datasets of any size. However, the massive scale, the speed of

ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions. The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.

In 2001, Gartner' s Doug Laney first presented what became known as the "three Vs of big data" to describe some of the characteristics that make big data different from other data processing:

Volume

The sheer scale of the information processed helps define big data systems. These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle.

Often, because the work requirements exceed the capabilities of a single computer, this becomes a challenge of pooling, allocating, and coordinating resources from groups of computers. Cluster management and algorithms capable of breaking tasks into smaller pieces become increasingly important.

Velocity

Another way in which big data differs significantly from other data systems is the speed that information moves through the system. Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.

This focus on near instant feedback has driven many big data practitioners away from a batch-oriented approach and closer to a real-time streaming system. Data is constantly being added, massaged, processed, and analyzed in order to keep up with the influx of new information and to surface valuable information early

when it is most relevant. These ideas require robust systems with highly available components to guard against failures along the data pipeline.

Variety

Big data problems are often unique because of the wide range of both the sources being processed and their relative quality.

Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers. Big data seeks to handle potentially useful data regardless of where it' s coming from by consolidating all information into a single system.

The formats and types of media can vary significantly as well. Rich media like images, video files, and audio recordings are ingested alongside text files, structured logs, etc. While more traditional data processing systems might expect data to enter the pipeline already labeled,

formatted, and organized, big data systems usually accept and store data closer to its raw state. Ideally, any transformations or changes to the raw data will happen in memory at the time of processing.

Other Characteristics

Various individuals and organizations have suggested expanding the original three Vs, though these proposals have tended to describe challenges rather than qualities of big data. Some common additions are:

- Veracity: The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis)
- Variability: Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.
- Value: The ultimate challenge of big data is delivering value. Sometimes, the systems and

processes in place are complex enough that using the data and extracting actual value can become difficult.

What Does a Big Data Life Cycle Look Like?

So how is data actually processed when dealing with a big data system? While approaches to implementation differ, there are some commonalities in the strategies and software that we can talk about generally. While the steps presented below might not be true in all cases, they are widely used.

The general categories of activities involved with big data processing are:

- Ingesting data into the system
- Persisting the data in storage
- Computing and Analyzing data
- Visualizing the results

Before we look at these four workflow categories in detail, we will take a moment to talk about clustered

computing, an important strategy employed by most big data solutions. Setting up a computing cluster is often the foundation for technology used in each of the life cycle stages.

Clustered Computing

Because of the qualities of big data, individual computers are often inadequate for handling the data at most stages. To better address the high storage and computational needs of big data, computer clusters are a better fit.

Big data clustering software combines the resources of many smaller machines, seeking to provide a number of benefits:

- Resource Pooling: Combining the available storage space to hold data is a clear benefit, but CPU and memory pooling is also extremely important. Processing large datasets requires large amounts of all three of these resources.

- High Availability: Clusters can provide varying levels of fault tolerance and availability guarantees to prevent hardware or software failures from affecting access to data and processing. This becomes increasingly important as we continue to emphasize the importance of real-time analytics.
- Easy Scalability: Clusters make it easy to scale horizontally by adding additional machines to the group. This means the system can react to changes in resource requirements without expanding the physical resources on a machine.

Using clusters requires a solution for managing cluster membership, coordinating resource sharing, and scheduling actual work on individual nodes. Cluster membership and resource allocation can be handled by software like Hadoop' s YARN (which stands for Yet Another Resource Negotiator) or Apache Mesos.

The assembled computing cluster often acts as a foundation which other software interfaces with to

process the data. The machines involved in the computing cluster are also typically involved with the management of a distributed storage system, which we will talk about when we discuss data persistence.

Ingesting Data into the System

Data ingestion is the process of taking raw data and adding it to the system. The complexity of this operation depends heavily on the format and quality of the data sources and how far the data is from the desired state prior to processing.

One way that data can be added to a big data system are dedicated ingestion tools. Technologies like Apache Sqoop can take existing data from relational databases and add it to a big data system. Similarly, Apache Flume and Apache Chukwa are projects designed to aggregate and import application and server logs. Queuing systems like Apache Kafka can also be used as an interface between various data generators and a big data system. Ingestion frameworks like Gobblin can help

to aggregate and normalize the output of these tools at the end of the ingestion pipeline.

During the ingestion process, some level of analysis, sorting, and labeling usually takes place. This process is sometimes called ETL, which stands for extract, transform, and load. While this term conventionally refers to legacy data warehousing processes, some of the same concepts apply to data entering the big data system. Typical operations might include modifying the incoming data to format it, categorizing and labeling data, filtering out unneeded or bad data, or potentially validating that it adheres to certain requirements.

With those capabilities in mind, ideally, the captured data should be kept as raw as possible for greater flexibility further on down the pipeline.

Persisting the Data in Storage

The ingestion processes typically hand the data off to the components that manage storage, so that it can be

reliably persisted to disk. While this seems like it would be a simple operation, the volume of incoming data, the requirements for availability, and the distributed computing layer make more complex storage systems necessary.

This usually means leveraging a distributed file system for raw data storage. Solutions like Apache Hadoop's HDFS file system allow large quantities of data to be written across multiple nodes in the cluster. This ensures that the data can be accessed by compute resources, can be loaded into the cluster's RAM for in-memory operations, and can gracefully handle component failures. Other distributed file systems can be used in place of HDFS including Ceph and GlusterFS.

Data can also be imported into other distributed systems for more structured access. Distributed databases, especially NoSQL databases, are well-suited for this role because they are often designed with the same fault tolerant considerations and can handle heterogeneous

data. There are many different types of distributed databases to choose from depending on how you want to organize and present the data. To learn more about some of the options and what purpose they best serve, read our NoSQL comparison guide.

Computing and Analyzing Data

Once the data is available, the system can begin processing the data to surface actual information. The computation layer is perhaps the most diverse part of the system as the requirements and best approach can vary significantly depending on what type of insights desired. Data is often processed repeatedly, either iteratively by a single tool or by using a number of tools to surface different types of insights.

Batch processing is one method of computing over a large dataset. The process involves breaking work up into smaller pieces, scheduling each piece on an individual machine, reshuffling the data based on the intermediate results, and then calculating and assembling the final

result. These steps are often referred to individually as splitting, mapping, shuffling, reducing, and assembling, or collectively as a distributed map reduce algorithm. This is the strategy used by Apache Hadoop's MapReduce. Batch processing is most useful when dealing with very large datasets that require quite a bit of computation.

While batch processing is a good fit for certain types of data and computation, other workloads require more real-time processing. Real-time processing demands that information be processed and made ready immediately and requires the system to react as new information becomes available. One way of achieving this is stream processing, which operates on a continuous stream of data composed of individual items. Another common characteristic of real-time processors is in-memory computing, which works with representations of the data in the cluster's memory to avoid having to write back to disk.

Apache Storm, Apache Flink, and Apache Spark provide different ways of achieving real-time or near real-time processing. There are trade-offs with each of these technologies, which can affect which approach is best for any individual problem. In general, real-time processing is best suited for analyzing smaller chunks of data that are changing or being added to the system rapidly.

The above examples represent computational frameworks. However, there are many other ways of computing over or analyzing data within a big data system. These tools frequently plug into the above frameworks and provide additional interfaces for interacting with the underlying layers. For instance, Apache Hive provides a data warehouse interface for Hadoop, Apache Pig provides a high level querying interface, while SQL-like interactions with data can be achieved with projects like Apache Drill, Apache Impala, Apache Spark SQL, and Presto. For machine learning, projects like Apache SystemML, Apache Mahout, and Apache Spark' s MLlib can be useful. For

straight analytics programming that has wide support in the big data ecosystem, both R and Python are popular choices.

Visualizing the Results

Due to the type of information being processed in big data systems, recognizing trends or changes in data over time is often more important than the values themselves. Visualizing data is one of the most useful ways to spot trends and make sense of a large number of data points.

Real-time processing is frequently used to visualize application and server metrics. The data changes frequently and large deltas in the metrics typically indicate significant impacts on the health of the systems or organization. In these cases, projects
like Prometheus can be useful for processing the data streams as a time-series database and visualizing that information.

One popular way of visualizing data is with the Elastic Stack, formerly known as the ELK stack. Composed of Log stash for data collection, Elastic search for indexing data, and Kibana for visualization, the Elastic stack can be used with big data systems to visually interface with the results of calculations or raw metrics. A similar stack can be achieved using Apache Solr for indexing and a Kibana fork called Banana for visualization. The stack created by these is called Silk.

Another visualization technology typically used for interactive data science work is a data "notebook". These projects allow for interactive exploration and visualization of the data in a format conducive to sharing, presenting, or collaborating. Popular examples of this type of visualization interface are Jupyter Notebook and Apache Zeppelin.

Big Data Glossary

While we've attempted to define concepts as we've used them throughout the guide, sometimes it's helpful to have specialized terminology available in a single place:

- Big data: Big data is an umbrella term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety. This term is also typically applied to technologies and strategies to work with this type of data.
- Batch processing: Batch processing is a computing strategy that involves processing data in large sets. This is typically ideal for non-time sensitive work that operates on very large sets of data. The process is started and at a later time, the results are returned by the system.
- Cluster computing: Clustered computing is the practice of pooling the resources of multiple

machines and managing their collective capabilities to complete tasks. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.

- Data lake: Data lake is a term for a large repository of collected data in a relatively raw state. This is frequently used to refer to the data collected in a big data system which might be unstructured and frequently changing. This differs in spirit to data warehouses (defined below).
- Data mining: Data mining is a broad term for the practice of trying to find patterns in large sets of data. It is the process of trying to refine a mass of data into a more understandable and cohesive set of information.
- Data warehouse: Data warehouses are large, ordered repositories of data that can be used for analysis and reporting. In contrast to a *data lake*, a data warehouse is composed of data that has been

cleaned, integrated with other sources, and is generally well-ordered. Data warehouses are often spoken about in relation to big data, but typically are components of more conventional systems.

- ETL: ETL stands for extract, transform, and load. It refers to the process of taking raw data and preparing it for the system's use. This is traditionally a process associated with data warehouses, but characteristics of this process are also found in the ingestion pipelines of big data systems.
- Hadoop: Hadoop is an Apache project that was the early open-source success in big data. It consists of a distributed file system called HDFS, with a cluster management and resource scheduler on top called YARN (Yet Another Resource Negotiator). Batch processing capabilities are provided by the MapReduce computation engine. Other computational and analysis systems can be run

alongside MapReduce in modern Hadoop deployments.

- In-memory computing: In-memory computing is a strategy that involves moving the working datasets entirely within a cluster’ s collective memory. Intermediate calculations are not written to disk and are instead held in memory. This gives in-memory computing systems like Apache Spark a huge advantage in speed over I/O bound systems like Hadoop’ s MapReduce.
- Machine learning: Machine learning is the study and practice of designing systems that can learn, adjust, and improve based on the data fed to them. This typically involves implementation of predictive and statistical algorithms that can continually zero in on “correct” behavior and insights as more data flows through the system.
- Map reduce (big data algorithm): Map reduce (the big data algorithm, not Hadoop’ s MapReduce computation engine) is an algorithm for scheduling

work on a computing cluster. The process involves splitting the problem set up (mapping it to different nodes) and computing over them to produce intermediate results, shuffling the results to align like sets, and then reducing the results by outputting a single value for each set.

- NoSQL: NoSQL is a broad term referring to databases designed outside of the traditional relational model. NoSQL databases have different trade-offs compared to relational databases, but are often well-suited for big data systems due to their flexibility and frequent distributed-first architecture.
- Stream processing: Stream processing is the practice of computing over individual data items as they move through a system. This allows for real-time analysis of the data being fed to the system and is useful for time-sensitive operations using high velocity metrics.

Conclusion

Big data is a broad, rapidly evolving topic. While it is not well-suited for all types of computing, many organizations are turning to big data for certain types of workloads and using it to supplement their existing analysis and business tools. Big data systems are uniquely suited for surfacing difficult-to-detect patterns and providing insight into behaviors that are impossible to find through conventional means. By correctly implement systems that deal with big data, organizations can gain incredible value from data that is already available.

The phrase Big Data has now been around for a while and we are at the stage where it is impacting more and more of us every day and it's a trend which is showing no signs of slowing down.

I have written hundreds of posts on big data, from what it is to how it is used in practice. To go alongside, I thought a post highlighting the meaning behind some of the jargon and buzzwords which have built up around the subject would be useful.

So here goes – these are topics everyone who want to know more about Big Data should have a general understanding of.

As-a-service infrastructure

Data-as-a-service, software-as-a-service, platform-as-a-service – all refer to the idea that rather than selling data, licenses to use data, or platforms for running Big Data technology, it can be provided “as a service”, rather than as a product. This reduces the upfront capital investment necessary for customers to begin putting their data, or platforms, to work for them, as the provider bears all of the costs of setting up and hosting the infrastructure. As a customer, as-a-service infrastructure can greatly reduce the initial cost and setup time of getting Big Data initiatives up and running.

Data science

Data science is the professional field that deals with turning data into value such as new insights or predictive models. It brings together expertise from fields including statistics, mathematics, computer science, communication as well as domain expertise such as business knowledge. Data scientist has recently been voted the No 1 job in the U.S., based on current demand and salary and career opportunities.

Data mining

Data mining is the process of discovering insights from data. In terms of Big Data, because it is so large, this is generally done by computational methods in an automated way using methods such as decision trees, clustering analysis and, most recently, machine learning. This can be thought of as using the brute mathematical power of computers to spot patterns in data which would not be visible to the human eye due to the complexity of the dataset.

Hadoop

Hadoop is a framework for Big Data computing which has been released into the public domain as open source software, and so can freely be used by anyone. It consists of a number of modules all tailored for a different vital step of the Big Data process – from file storage (Hadoop File System – HDFS) to database (HBase) to carrying out data operations (Hadoop MapReduce – see below). It has become so popular due to its power and flexibility that it has developed its own industry of retailers (selling tailored versions), support service providers and consultants.

Predictive modeling

At its simplest, this is predicting what will be next based on data about what has happened previously. In the Big Data age, because there is more data around than ever before,

predictions are becoming more and more accurate. Predictive modeling is a core component of most Big Data initiatives, which are formulated to help us choose the course of action which will lead to the most desirable outcome. The speed of modern computers and the volume of data available mean that predictions can be made based on a huge number of variables, allowing an ever-increasing number of variables to be assessed for the probability that it will lead to success.

MapReduce

MapReduce is a computing procedure for working with large datasets, which was devised due to difficulty of reading and analyzing really Big Data using conventional computing methodologies. As its name suggest, it consists of two procedures – mapping (sorting information into the format needed for analysis – i.e. sorting a list of people according to their age) and reducing (performing an operation, such checking the age of everyone in the dataset to see who is over 21).

NoSQL

NoSQL refers to a database format designed to hold more than data which is simply arranged into tables, rows, and columns, as is the case in a conventional relational database. This database format has proven very popular

in Big Data applications because Big Data is often messy, unstructured and does not easily fit into traditional database frameworks.

Python

Python is a programming language which has become very popular in the Big Data space due to its ability to work very well with large, unstructured datasets (see Part II for the difference between structured and unstructured data). It is considered to be easier to learn for a data science beginner than other languages such as R (see also Part II) and more flexible.

R

R is another programming language commonly used in Big Data, and can be thought of as more specialized than Python, being geared towards statistics. Its strength lies in its powerful handling of structured data. Like Python, it has an active community of users who are constantly expanding and adding to its capabilities by creating new libraries and extensions.

Recommendation engine

A recommendation engine is basically an algorithm, or collection of algorithms, designed to match an entity (for example, a customer) with something they are looking for.

Recommendation engines used by the likes of Netflix or Amazon heavily rely on Big Data technology to gain an overview of their customers and, using predictive modeling, match them with products to buy or content to consume. The economic incentive offered by recommendation engines has been a driving force behind a lot of commercial Big Data initiatives and developments over the last decade.

Real-time

Real-time means “as it happens” and in Big Data refers to a system or process which is able to give data-driven insights based on what is happening at the present moment. Recent years have seen a large push for the development of systems capable of processing and offering insights in real-time (or near-real-time), and advances in computing power as well as development of techniques such as machine learning have made it a reality in many applications today.

Reporting

The crucial “last step” of many Big Data initiative involves getting the right information to the people who need it to make decisions, at the right time. When this step is automated, analytics is applied to the insights themselves to ensure that they are communicated in a way that they

will be understood and easy to act on. This will usually involve creating multiple reports based on the same data or insights but each intended for a different audience (for example, in-depth technical analysis for engineers, and an overview of the impact on the bottom line for c-level executives).

Spark

Spark is another open source framework like Hadoop but more recently developed and more suited to handling cutting-edge Big Data tasks involving real time analytics and machine learning. Unlike Hadoop it does not include its own file system, though it is designed to work with Hadoop's HDFS or a number of other options. However, for certain data related processes it is able to calculate at over 100 times the speed of Hadoop, thanks to its in-memory processing capability. This means it is becoming an increasingly popular choice for projects involving deep learning, neural networks and other compute-intensive tasks.

Structured Data

Structured data is simply data that can be arranged neatly into charts and tables consisting of rows, columns or multi-dimensioned matrixes. This is traditionally the way that computers have stored data, and information in this format

can easily and simply be processed and mined for insights. Data gathered from machines is often a good example of structured data, where various data points – speed, temperature, rate of failure, RPM etc. – can be neatly recorded and tabulated for analysis.

Unstructured Data

Unstructured data is any data which cannot easily be put into conventional charts and tables. This can include video data, pictures, recorded sounds, text written in human languages and a great deal more. This data has traditionally been far harder to draw insight from using computers which were generally designed to read and analyse structured information. However, since it has become apparent that a huge amount of value can be locked away in this unstructured data, great efforts have been made to create applications which are capable of understanding unstructured data – for example visual recognition and natural language processing.

Visualization

Humans find it very hard to understand and draw insights from large amounts of text or numerical data – we can do it, but it takes time, and our concentration and attention is limited. For this reason effort has been made to develop computer applications capable of rendering information in

a visual form – charts and graphics which highlight the most important insights which have resulted from our Big Data projects. A subfield of reporting (see above), visualizing is now often an automated process, with visualizations customized by algorithm to be understandable to the people who need to act or take decisions based on them.

Big Data A to Z: A glossary of Big Data terminology

This is almost a complete glossary of Big Data terminology widely used today. Let us know if you would like to add any big data terminology missing in this list.

1.ACID test

A test applied to data for atomicity, consistency, isolation, and durability

2.**Aggregation**

A process of searching, gathering and presenting data

3.**Algorithm**

A mathematical formula placed in software that performs an analysis on a set of data.

4.**Anonymization**

The severing of links between people in a database and their records to prevent the discovery of the source of the records.

5.**Artificial Intelligence**

Developing intelligence machines and software that are capable of perceiving the environment and take corresponding action when required and even learn from those actions.

6. Automatic identification and capture (AIDC)

Any method of automatically identifying and collecting data on items, and then storing the data in a computer system. For example, a scanner might collect data about a product being shipped via an RFID chip.

7. Avro

Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing remote procedure calls

8. Behavioral analytics

Using data about people's behavior to understand intent and predict future actions.

9.Big Data Scientist

Someone who is able to develop the algorithms to make sense out of big data.

10. Business Intelligence (BI)

The general term used for the identification, extraction, and analysis of data.

11. Cascading

Cascading provides a higher level of abstraction for Hadoop, allowing developers to create complex jobs quickly, easily, and in several different languages that run in the JVM, including Ruby, Scala, and more. In effect, this has shattered the skills barrier, enabling Twitter to use Hadoop more broadly.

12. Call Detail Record (CDR) analysis

CDRs contain data that a telecommunications company collects about phone calls, such as time and length of call. This data can be used in any number of analytical applications.

13. Cassandra

Cassandra is a distributed and Open Source database. Designed to handle large amounts of distributed data across commodity servers while providing a highly available service. It is a NoSQL solution that was initially developed by Facebook. It is structured in the form of key-value.

14. Cell phone data

Cell phones generate a tremendous amount of data, and much of it is available for use with analytical applications.

15. Clickstream Analytics

The analysis of users' Web activity through the items they click on a page.

16. Classification analysis

A systematic process for obtaining important and relevant information about data, also meta data called; data about data.

17. Cloud computing

A distributed computing system over a network used for storing data off-premises

18. Clustering analysis

The process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.

19. Cold data storage

Storing old data that is hardly used on low-power servers. Retrieving the data will take longer

20. Comparative analysis

It ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.

21. Chukwa

Chukwa is a Hadoop subproject devoted to large-scale log collection and analysis. Chukwa is built on top of the Hadoop distributed filesystem (HDFS) and MapReduce framework and inherits Hadoop's scalability and robustness. Chukwa also includes a flexible and powerful toolkit for displaying monitoring and analyzing results, in order to make the best use of this collected data.

22. Clojure

Clojure is a dynamic programming language based on LISP that uses the Java Virtual Machine (JVM). It is well suited for parallel data processing.

23. Cloud

A broad term that refers to any Internet-based application or service that is hosted remotely.

24. Columnar database or column-oriented database

A database that stores data by column rather than by row. In a row-based database, a row might contain a name, address, and phone number. In a column-oriented database, all names are in one column, addresses in another, and so on. A key advantage of a columnar database is faster hard disk access.

25. Comparators

Two ways you may compare your keys is by implementing the interface or by implementing the RawComparator interface. In the former approach, you will compare (deserialized) objects, but in the latter approach, you will compare the keys using their corresponding raw bytes.

26. Complex event processing (CEP)

CEP is the process of monitoring and analyzing all events across an organization’s systems and acting on them when necessary in real time.

27. Confabulation

The act of making an intuition-based decision appear to be data-based.

28. Cross-channel analytics

Analysis that can attribute sales, show average order value, or the lifetime value.

29. Data access

The act or method of viewing or retrieving stored data.

30. Dashboard

A graphical representation of the analyses performed by the algorithms

31. Data aggregation

The act of collecting data from multiple sources for the purpose of reporting or analysis.

32. Data architecture and design

How enterprise data is structured. The actual structure or design varies depending on the eventual end result required. Data architecture has three stages or processes: conceptual representation of business entities. the logical representation of the relationships among those entities, and the physical construction of the system to support the functionality.

33. Database

A digital collection of data and the structure around which the data is organized. The data is typically entered into and accessed via a database management system (DBMS).

34. Database administrator (DBA)

A person, often certified, who is responsible for supporting and maintaining the integrity of the structure and content of a database.

35. Database as a service (DaaS)

A database hosted in the cloud and sold on a metered basis. Examples include Heroku Postgres and Amazon Relational Database Service.

36. Database management system (DBMS)

Software that collects and provides access to data in a structured format.

37. Data center

A physical facility that houses a large number of servers and data storage devices. Data centers might belong to a single organization or sell their services to many organizations.

38. Data cleansing

The act of reviewing and revising data to remove duplicate entries, correct misspellings, add missing data, and provide more consistency.

39. Data collection

Any process that captures any type of data.

40. Data custodian

A person responsible for the database structure and the technical environment, including the storage of data.

41. Data-directed decision making

Using data to support making crucial decisions.

42. Data exhaust

The data that a person creates as a byproduct of a common activity–for example, a cell call log or web search history.

43. Data feed

A means for a person to receive a stream of data. Examples of data feed mechanisms include RSS or Twitter.

44. Data governance

A set of processes or rules that ensure the integrity of the data and that data management best practices are met.

45. Data integration

The process of combining data from different sources and presenting it in a single view.

46. Data integrity

The measure of trust an organization has in the accuracy, completeness, timeliness, and validity of the data.

47. Data mart

The access layer of a data warehouse used to provide data to users.

48. Data migration

The process of moving data between different storage types or formats, or between different computer systems.

49. Data mining

The process of deriving patterns or knowledge from large data sets.

50. Data model, data modeling

A data model defines the structure of the data for the purpose of communicating between functional and technical people to show data needed for business processes, or for communicating a plan to develop how data is stored and accessed among application development team members.

51. Data point

An individual item on a graph or a chart.

52. Data profiling

The process of collecting statistics and information about data in an existing source.

53. Data quality

The measure of data to determine its worthiness for decision making, planning, or operations.

54. Data replication

The process of sharing information to ensure consistency between redundant sources.

55. Data repository

The location of permanently stored data.

56. Data science

A recent term that has multiple definitions, but generally accepted as a discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning, and database engineering to solve complex problems.

57. Data scientist

A practitioner of data science.

58. Data security

The practice of protecting data from destruction or unauthorized access.

59. Data set

A collection of data, typically in tabular form.

60. Data source

Any provider of data–for example, a database or a data stream.

61. Data steward

A person responsible for data stored in a data field.

62. Data structure

A specific way of storing and organizing data.

63. Data visualization

A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

64. Data warehouse

A place to store data for the purpose of reporting and analysis.

65. De-identification

The act of removing all data that links a person to a particular piece of information.

66. Demographic data

Data relating to the characteristics of a human population.

67. Deep Thunder

IBM’s weather prediction service that provides weather data to organizations such as utilities, which use the data to optimize energy distribution.

68. Distributed cache

A data cache that is spread across multiple systems but works as one. It is used to improve performance.

69. Distributed object

A software module designed to work with other distributed objects stored on other computers.

70. Distributed processing

The execution of a process across multiple computers connected by a computer network.

71. Distributed File System

Systems that offer simplified, highly available access to storing, analysing and processing data

72. Document Store Databases

A document-oriented database that is especially designed to store, manage and retrieve documents, also known as semi structured data.

73. Document management

The practice of tracking and storing electronic documents and scanned images of paper documents.

74. Drill

An open source distributed system for performing interactive analysis on large-scale datasets. It is similar to Google’s Dremel, and is managed by Apache.

Elasticsearch

An open source search engine built on Apache Lucene.

75. Event analytics

Shows the series of steps that led to an action.

76. Exabyte

One million terabytes, or 1 billion gigabytes of information.

77. External data

Data that exists outside of a system.

78. Extract, transform, and load (ETL)

A process used in data warehousing to prepare data for use in reporting or analytics.

79. Exploratory analysis

Finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.

80. Failover

The automatic switching to another computer or node should one fail.

81. Flume

Flume is a framework for populating Hadoop with data. Agents are populated throughout ones IT infrastructure – inside web servers, application servers and mobile devices, for example – to collect data and integrate it into Hadoop.

82. Grid computing

The performing of computing functions using resources from multiple distributed systems. Grid computing typically involves

large files and are most often used for multiple applications. The systems that comprise a grid computing network do not have to be similar in design or in the same geographic location.

83. Graph Databases

They use graph structures (a finite set of ordered pairs or certain entities), with edges, properties and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbor element.

84. Hadoop

An open source software library project administered by the Apache Software Foundation. Apache defines Hadoop as “a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model.”

Hama

Hama is a distributed computing framework based on Bulk Synchronous Parallel computing techniques for massive scientific computations e.g., matrix, graph and network algorithms. It's a Top Level Project under the Apache Software Foundation.

85. HANA

A software/hardware in-memory computing platform from SAP designed for high-volume transactions and real-time analytics.

86. HBase

HBase is a non-relational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. EBay and Facebook use HBase heavily

87. HCatalog

HCatalog is a centralized metadata management and sharing service for Apache Hadoop. It allows for a unified view of all data in Hadoop clusters and allows diverse tools, including Pig and Hive, to process any data elements without needing to know physically where in the cluster the data is stored.

88. HDFS (Hadoop Distributed File System)

HDFS (Hadoop Distributed File System) the storage layer of Hadoop, is a distributed, scalable, Java-based file system adept at storing large volumes of unstructured.

89. Hive

Hive is a Hadoop-based data warehousing-like framework originally developed by Facebook. It allows users to write queries in a SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it

easier to integrate with business intelligence and visualization tools such as Microstrategy, Tableau, Revolutions Analytics, etc.

90. **Hue**

Hue (Hadoop User Experience) is an open source web-based interface for making it easier to use Apache Hadoop. It features a file browser for HDFS, an Oozie Application for creating workflows and coordinators, a job designer/browser for MapReduce, a Hive and Impala UI, a Shell, a collection of Hadoop API and more.

91. **Impala**

Impala (By Cloudera) provides fast, interactive SQL queries directly on your Apache Hadoop data stored in HDFS or HBase using the same metadata, SQL syntax (Hive SQL), ODBC driver and user interface (Hue Beeswax) as Apache Hive. This provides a familiar and unified platform for batch-oriented or real-time queries.

92. In-database analytics

The integration of data analytics into the data warehouse.

93. In-memory database

Any database system that relies on memory for data storage.

94. In-memory data grid (IMDG)

The storage of data in memory across multiple servers for the purpose of greater scalability and faster access or analytics.

95. Internet of Things

Ordinary devices that are connected to the internet at any time any where via sensors

96. Kafka

Kafka (developed by LinkedIn) is a distributed publish-subscribe messaging system that offers a solution capable of

handling all data flow activity and processing these data on a consumer website. This type of data (page views, searches, and other user actions) are a key ingredient in the current social web.

97. Key Value Stores

Key value stores allow the application to store its data in a schema-less way. The data could be stored in a datatype of a programming language or an object. Because of this, there is no need for a fixed data model.

98. KeyValue Databases

They store data with a primary key, a uniquely identifiable record, which makes easy and fast to look up. The data stored in a KeyValue is normally some kind of primitive of the programming language.

99. Latency

Any delay in a response or delivery of data from one point to another.

100. **Linked data**

As described by World Wide Web inventor Time Berners-Lee, “Cherry-picking common attributes or languages to identify connections or relationships between disparate sources of data.”

101. **Load balancing**

The process of distributing workload across a computer network or computer cluster to optimize performance.

102. **Location analytics**

Location analytics brings mapping and map-driven analytics to enterprise business systems and data warehouses. It allows you to associate geospatial information with datasets.

103. **Location data**

Data that describes a geographic location.

104. **Log file**

A file that a computer, network, or application creates automatically to record events that occur during operation–for example, the time a file is accessed.

105. **Machine-generated data**

Any data that is automatically created from a computer process, application, or other non-human source.

106. **Machine2Machine data**

Two or more machines that are communicating with each other

107. **Machine learning**

The use of algorithms to allow a computer to analyze data for the purpose of "learning" what action to take when a specific pattern or event occurs.

108. **MapReduce**

MapReduce is a software framework that serves as the compute layer of Hadoop. MapReduce jobs are divided into two (obviously named) parts. The “Map” function divides a query into multiple parts and processes data at the node level. The “Reduce” function aggregates the results of the “Map” function to determine the “answer” to the query.

109. **Mashup**

The process of combining different datasets within a single application to enhance output–for example, combining demographic data with real estate listings.

110. **Mahout**

Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing and statistical modeling and implements them using the Map Reduce model.

111. Metadata

Data about data; gives information about what the data is about.

112. MongoDB

MongoDB is a NoSQL database oriented to documents, developed under the open source concept. It saves data structures in JSON documents with a dynamic scheme (called MongoDB BSON format), making the integration of the data in certain applications more easily and quickly.

113. MPP database

A database optimized to work in a massively parallel processing environment.

114. Multi-Dimensional Databases

A database optimized for data online analytical processing (OLAP) applications and for data warehousing.

115. MultiValue Databases

They are a type of NoSQL and multidimensional databases that understand 3 dimensional data directly. They are primarily giant strings that are perfect for manipulating HTML and XML strings directly

116. Network analysis

Viewing relationships among the nodes in terms of the network or graph theory, meaning analysing connections between nodes in a network and the strength of the ties.

117. NewSQL

An elegant, well-defined database system that is easier to learn and better than SQL. It is even newer than NoSQL

118. NoSQL

NoSQL (commonly interpreted as “not only SQL“) is a broad class of database management systems identified by non-

adherence to the widely used relational database management system model. NoSQL databases are not built primarily on tables, and generally do not use SQL for data manipulation.

119. Object Databases

They store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

120. Object-based Image Analysis

Analysing digital images can be performed with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.

121. Online analytical processing (OLAP)

The process of analyzing multidimensional data using three operations: consolidation (the aggregation of available), drill-

down (the ability for users to see the underlying details), and slice and dice (the ability for users to select subsets and view them from different perspectives).

122. **Online transactional processing (OLTP)**

The process of providing users with access to large amounts of transactional data in a way that they can derive meaning from it.

123. **OpenDremel**

The open source version of Google's Big Query java code. It is being integrated with Apache Drill.

124. **Open Data Center Alliance (ODCA)**

A consortium of global IT organizations whose goal is to speed the migration of cloud computing.

125. **Operational data store (ODS)**

A location to gather and store data from multiple sources so that more operations can be performed on it before sending to the data warehouse for reporting.

126. **Oozie**

Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages – such as Map Reduce, Pig and Hive — then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.

127. **Parallel data analysis**

Breaking up an analytical problem into smaller components and running algorithms on each of those components at the same time. Parallel data analysis can occur within the same system or across multiple systems.

128. Parallel method invocation (PMI)

Allows programming code to call multiple functions in parallel.

129. Parallel processing

The ability to execute multiple tasks at the same time.

130. Parallel query

A query that is executed over multiple system threads for faster performance.

131. Pattern recognition

The classification or labeling of an identified pattern in the machine learning process.

132. Pentaho

Pentaho offers a suite of open source Business Intelligence (BI) products called Pentaho Business Analytics providing data

integration, OLAP services, reporting, dashboarding, data mining and ETL capabilities

133. Petabyte

One million gigabytes or 1,024 terabytes.

134. Pig

Pig Latin is a Hadoop-based language developed by Yahoo. It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL).

135. Predictive analytics

Using statistical functions on one or more datasets to predict trends or future events.

136. Predictive modeling

The process of developing a model that will most likely predict a trend or outcome.

137. **Public data**

Public information or data sets that were created with public funding

138. **Query**

Asking for information to answer a certain question

139. **Query analysis**

The process of analyzing a search query for the purpose of optimizing it for the best possible result.

140. **R**

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

141. **Re-identification**

Combining several data sets to find a certain person within anonym zed data

142. **Real-time data**

Data that is created, processed, stored, analyzed and visualized within milliseconds

143. **Recommendation engine**

An algorithm that analyzes a customer’s purchases and actions on an e-commerce site and then uses that data to recommend complementary products.

144. **Reference data**

Data that describes an object and its properties. The object may be physical or virtual.

145. **Risk analysis**

The application of statistical methods on one or more datasets to determine the likely risk of a project, action, or decision.

146. **Root-cause analysis**

The process of determining the main cause of an event or problem.

147. **Routing analysis**

Finding the optimized routing using many different variables for a certain means of transport in order to decrease fuel costs and increase efficiency.

148. **Scalability**

The ability of a system or process to maintain acceptable performance levels as workload or scope increases.

149. Schema

The structure that defines the organization of data in a database system.

150. Search data

Aggregated data about search terms used over time.

151. Semi-structured data

Data that is not structured by a formal data model, but provides other means of describing the data and hierarchies.

152. Sentiment analysis

The application of statistical functions on comments people make on the web and through social networks to determine how they feel about a product or company.

153. **Server**

A physical or virtual computer that serves requests for a software application and delivers those requests over a network.

154. **Spatial analysis**

It refers to analysing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.

155. **SQL**

A programming language for retrieving data from a relational database

156. **Sqoop**

Sqoop is a connectivity tool for moving data from non-Hadoop data stores – such as relational databases and data warehouses – into Hadoop. It allows users to specify the target location inside

of Hadoop and instruct Sqoop to move data from Oracle, Teradata or other relational databases to the target.

157. **Storm**

Storm is a system of real-time distributed computing, open source and free, born into Twitter. Storm makes it easy to reliably process unstructured data flows in the field of real-time processing, which made Hadoop for batch processing.

158. **Software as a service (SaaS)**

Application software that is used over the web by a thin client or web browser. Salesforce is a well-known example of SaaS.

159. **Storage**

Any means of storing data persistently.

160. **Storm**

An open-source distributed computation system designed for processing multiple data streams in real time.

161. **Structured data**

Data that is organized by a predetermined structure.

162. **Structured Query Language (SQL)**

A programming language designed specifically to manage and retrieve data from a relational database system.

163. **Text analytics**

The application of statistical, linguistic, and machine learning techniques on text-based sources to derive meaning or insight.

164. **Transactional data**

Data that changes unpredictably. Examples include accounts payable and receivable data, or data about product shipments.

165. **Thrift**

“Thrift is a software framework for scalable cross-language services development. It combines a software stack with a code generation engine to build services that work efficiently and seamlessly between C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml.”

166. **Unstructured data**

Data that has no identifiable structure–for example, the text of email messages.

167. **Value**

All that available data will create a lot of value for organizations, societies and consumers. Big data means big business and every industry will reap the benefits from big data.

168. **Volume**

The amount of data, ranging from megabytes to brontobytes

169. **Visualization**

A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

170. **WebHDFS Apache Hadoop**

WebHDFS Apache Hadoop provides native libraries for accessing HDFS. However, users prefer to use HDFS remotely over the heavy client side native libraries. For example, some applications need to load data in and out of the cluster, or to externally interact with the HDFS data. WebHDFS addresses

these issues by providing a fully functional HTTP REST API to access HDFS.

171. **Weather data**

Real-time weather data is now widely available for organizations to use in a variety of ways. For example, a logistics company can monitor local weather conditions to optimize the transport of goods. A utility company can adjust energy distribution in real time.

172. **XML Databases**

XML Databases allow data to be stored in XML format. XML databases are often linked to document-oriented databases. The data stored in an XML database can be queried, exported and serialized into any format needed.

173. **ZooKeeper**

ZooKeeper is a software project of the Apache Software Foundation, a service that provides centralized configuration and open code name registration for large distributed systems. ZooKeeper is a subproject of Hadoop.

101 Big Data Terms You Should Know

With the huge amount of data that is generated every day, it becomes important to understand the complexity of Big Data. If you are planning to enter in the Big Data planet, you should be familiar with the Big Data terms. These terms will help you dive deep into the world of big data. So, let's start with the term Big Data itself –

An accurate definition of "Big Data" is quite hard to clinch as business professionals, projects, practitioners, and vendors take it in a different way. So, generally speaking, big data is:

- large/big data sets (large dataset refers to a dataset too large to store or process on a single computer) and,
- the classification of computing technologies and strategies which are used to confer large data sets.

Big Data is a wide spectrum and there is a lot to learn. Here are the big data terms, you should be familiar with.

Every field has its own terminology and thus, there are a number of Big Data terms to know while starting a career in Big Data. Once you will get familiar with these Big Data terms and definitions, you will be prepared to learn them in detail. In this article, we are going to define 101 Big Data terms that you should know to start a career in Big Data.

A

1. Algorithm

In computer science and mathematics, an algorithm is an effective categorical specification of how to solve a complex problem and how to perform data analysis. It consists of multiple steps to apply operations on data in order to solve a particular problem.

2. Artificial Intelligence (AI)

The popular Big Data term, Artificial Intelligence is the intelligence demonstrated by machines. AI is the development of computer systems to perform tasks normally having human intelligence such as speech recognition, visual perception, decision making and language translators etc.

3. Automatic Identification and Data Capture (AIDC)

Automatic identification and data capture (AIDC) is the big data term that refers to a method of automatically identifying and collecting data objects through computing algorithm and then storing them in the computer. For example, radio frequency identification, bar codes, biometrics, optical character recognition, magnetic strips all include algorithms for identification of data objects captured.

4. Avro

Avro is data serialization framework and a remote procedure call developed for Hadoop's project. It uses JSON to define protocols and data types and then serializes data in binary form. Avro provides both

- Serialization format for persistent data
- Wire format for communication between Hadoop nodes and from customer programs to Hadoop services.

B

5. Behavioural Analytics

Behavioural analytics is a recent advancement in business analytics that presents new insights into client's behaviour on e-commerce platforms,

web/mobile application, online games etc. It enables the marketers to make right offers to the right customers at right time.

6. Business Intelligence

Business Intelligence is a set of tools and methodologies that can analyze, manage, and deliver information which is relevant to the business. It includes reporting/query tools and dashboard same as found in analytics. BI technologies provide previous, current, and upcoming views of the business operations.

7. ***Big Data Scientist***

Big Data Scientist is a person who can take structured and unstructured data points and use his formidable skills in statistics, maths, and programming to organize them. He applies all his analytical power (contextual understanding, industry knowledge, and understanding of existing assumptions) to uncover the hidden solutions for the business development.

8. ***Biometrics***

Biometrics is the James Bondish technology linked with analytics to identify people by one or more physical traits. For example, biometrics technology is used in face recognition, fingerprint recognition, iris recognition etc.

C

9. Cascading

Cascading is the layer for the abstraction of software that provides the higher level abstraction for Apache Hadoop and Apache Flink. It is an open source framework that is available under Apache License. It is used to allow developers to perform processing of complex data easily and quickly in JVM based languages such as Java, Clojure, Scala, Rubi etc.

10. Call Detail Record (CDR) Analysis

CDR contains metadata i.e. data about data that a telecommunication company collects about phone calls such as length and time of the call. CDR analysis provides businesses the exact details about when, where, and how calls are made for billing and reporting purposes. CDR's metadata gives information about

- When the calls are made (date and time)
- How long the call lasted (in minutes)
- Who called whom (Contact number of source and destination)
- Type of call (Inbound, Outbound or Toll-free)
- How much the call costs (on the basis of per minute rate)

11. Cassandra

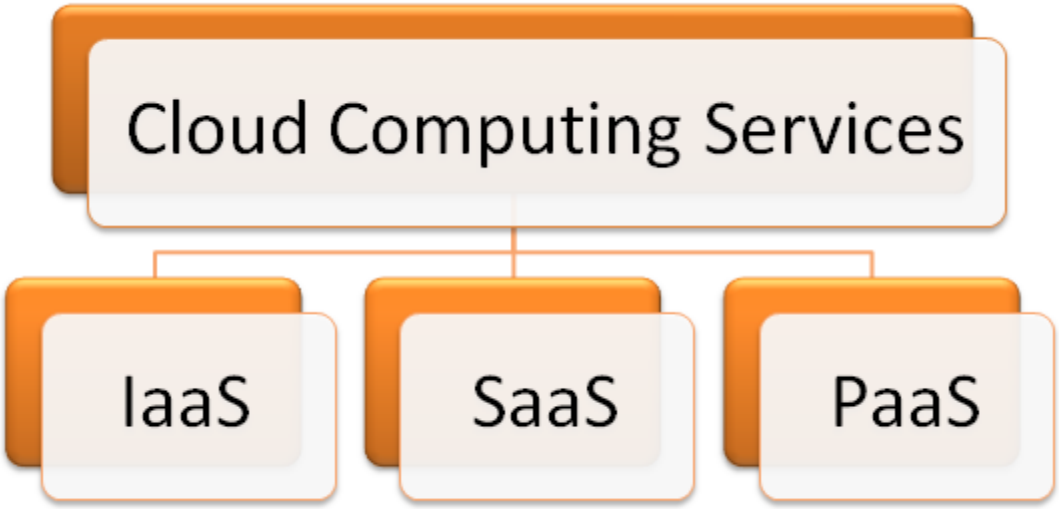
Cassandra is distributed and open source NoSQL database management system. It is schemed to manage a large amount of distributed data over commodity servers as it provides high availability of services with no point of failure. It was developed by Facebook initially and then structured in key-value form under Apache foundation.

12. Cell Phone Data

Cell phone data has surfaced as one of the big data sources as it generates a tremendous amount of data and much of it is available for use with analytical applications.

13. Cloud Computing

Cloud computing is one of the must-known big data terms. It is a new paradigm computing system which offers visualization of computing resources to run over the standard remote server for storing data and provides IaaS, PaaS, and SaaS. Cloud Computing provides IT resources such as Infrastructure, software, platform, database, storage and so on as services. Flexible scaling, rapid elasticity, resource pooling, on-demand self-service are some of its services.



14. Cluster Analysis

Cluster analysis is the big data term related to the process of the grouping of objects similar to each other in the common group (cluster). It is done to understand the similarities and differences between them. It is the important task of exploratory data mining, and common strategies to analyze statistical data in various fields such as image analysis, pattern recognition, machine learning, computer graphics, data compression and so on.

15. Chukwa

Apache Chukwa is an open source large-scale log collection system for monitoring large distributed systems. It is one of the common big data terms related to Hadoop. It is built on the top of Hadoop Distributed File System (HDFS) and Map/Reduce framework. It inherits Hadoop's robustness and scalability. Chukwa contains a powerful and flexible toolkit database for

monitoring, displaying, and analyzing results so that collected data can be used in the best possible ways.

16. Columnar Database / Column-Oriented Database

A database that stores data column by column instead of the row is known as the column-oriented database.

17. Comparative Analytic-oriented Database

Comparative analytic is a special type of data mining technology which compares large data sets, multiple processes or other objects using statistical strategies such as filtering, decision tree analytics, pattern analysis etc.

18. Complex Event Processing (CEP)

Complex event processing (CEP) is the process of analyzing and identifying data and then combining it to infer events that are able to suggest solutions to the complex circumstances. The main task of CEP is to identify/track meaningful events and react to them as soon as possible.

D

19. Data Analyst

The data analyst is responsible for collecting, processing, and performing statistical analysis of data. A data analyst discovers the ways how this data can be used to help the organization in making better business decisions. It is one of the big data terms that define a big data career. Data analyst works with end business users to define the types of the analytical report required in business.

20. Data Aggregation

Data aggregation refers to the collection of data from multiple sources to bring all the data together into a common athenaeum for the purpose of reporting and/or analysis.

The knowledge of one of the high-level programming languages is required to build a career in Big Data. Let's check out which are the Top 3 Big Data Programming Languages for You!

21. Dashboard

It is a graphical representation of analysis performed by the algorithms. This graphical report shows different color alerts to show the activity status. A green light is for the normal operations, a yellow light shows that there is some impact due to operation and a red light signifies that the operation has

been stopped. This alertness with different lights helps to track the status of operations and find out the details whenever required.

22. Data Scientist

Data Scientist is also a big data term that defines a big data career. A data scientist is a practitioner of data science. He is proficient in mathematics, statistics, computer science, and/or data visualization who establish data models and algorithms for complex problems to solve them.

23. Data Architecture and Design

In IT industry, Data architecture consists of models, policies standards or rules that control which data is aggregated, and how it is arranged, stored, integrated and brought to use in data systems. It has three phases

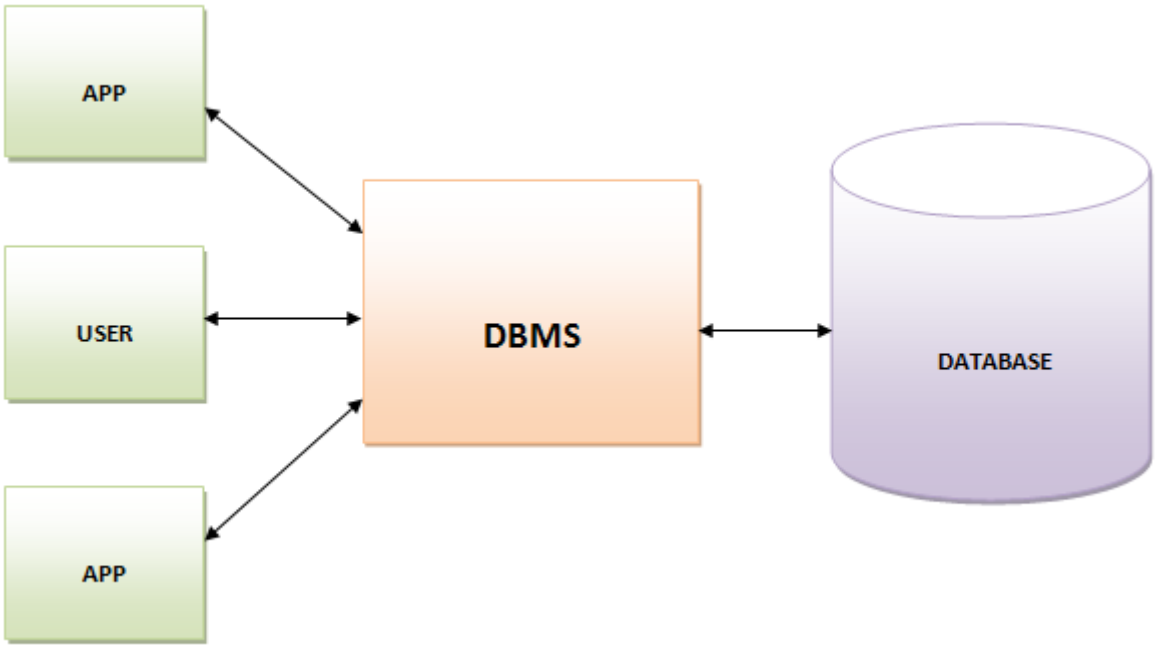
- Conceptual representation of business entities
- The logical representation of the relationships between business entities
- The physical construction of the system for functional support

24. Database administrator (DBA)

DBA is the big data term related to a role which includes capacity planning, configuration, database design, performance monitoring, migration, troubleshooting, security, backups and data recovery. DBA is responsible for maintaining and supporting the rectitude of content and structure of a database.

25. Database Management System (DBMS)

Database Management System is software that collects data and provides access to it in an organized layout. It creates and manages the database. DBMS provides programmers and users a well-organized process to create, update, retrieve, and manage data.



26. Data Model and Data Modelling

Data Model is a starting phase of a database designing and usually consists of attributes, entity types, integrity rules, relationships and definitions of objects.

Data modeling is the process of creating a data model for an information system by using certain formal techniques. Data modeling is used to define and analyze the requirement of data for supporting business processes.

Looking for big data tools to start a big data career? Here are the Top 10 Open Source Big Data Tools in 2018.

27. Data Cleansing

Data Cleansing/Scrubbing/Cleaning is a process of revising data to remove incorrect spellings, duplicate entries, adding missing data, and providing consistency. It is required as incorrect data can lead to bad analysis and wrong conclusions.

28. Document Management

Document management, often, referred to as Document management system is a software which is used to track, store, and manage electronic documents and an electronic image of paper through a scanner. It is one of the basic big data terms you should know to start a big data career.

29. Data Visualization

Data visualization is the presentation of data in a graphical or pictorial format designed for the purpose of communicating information or deriving meaning. It validates the users/decision makers to see analytics visually so that they would be able to understand the new concepts. This data helps –

- to derive insight and meaning from the data
- in the communication of data and information in a more effective manner

30. Data Warehouse

The data warehouse is a system of storing data for the purpose of analysis and reporting. It is believed to be the main component of business intelligence. Data stored in the warehouse is uploaded from the operational system like sales or marketing.

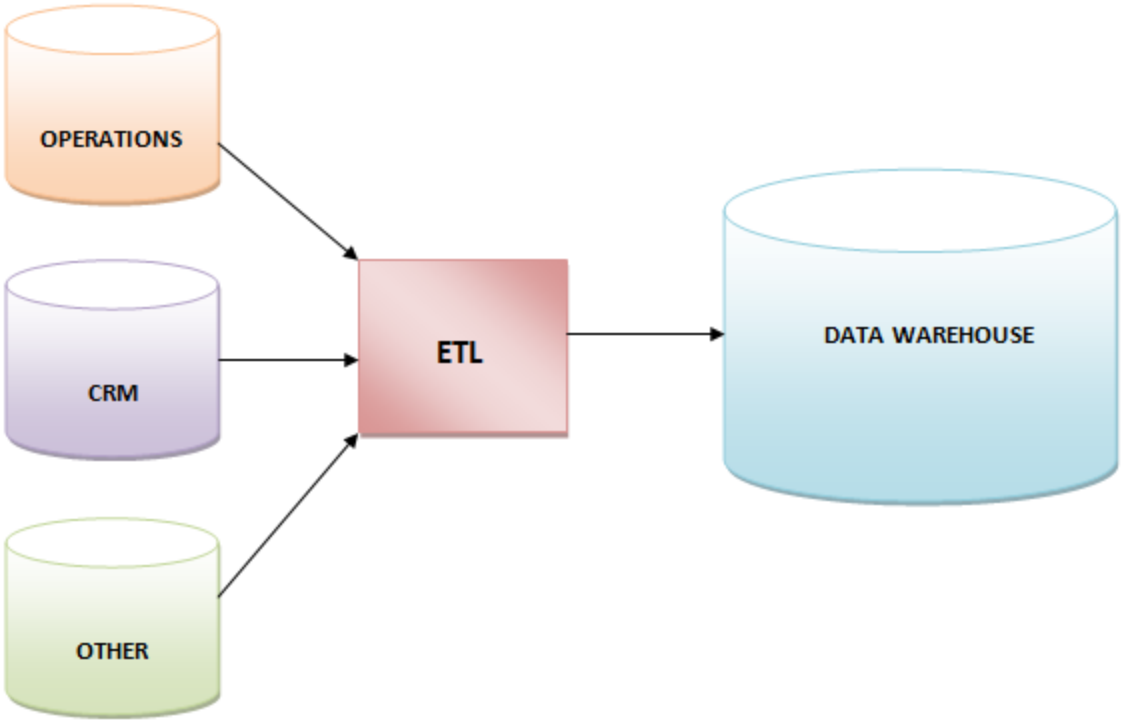
31. Drill

The drill is an open source, distributed, low latency SQL query engine for Hadoop. It is built for semi-structured or nested data and can handle fixed schemas. The drill is similar in some aspects to Google's Dremel and is handled by Apache.

E

32. Extract, Transform, and Load (ETL)

ETL is the short form of three database functions extract, transform and load. These three functions are combined together into one tool to place them from one to another database.



- **Extract**

It is the process of reading data from a database.

- **Transform**

It is the process of conversion of extracted data in the desired form so that it can be put into another database.

- **Load**

It is the process of writing data into the target database

F

33. Fuzzy Logic

Fuzzy logic is an approach to computing based on degrees of truth instead of usual true/false (1 or 0) Boolean algebra.

34. Flume

Flume is defined as a reliable, distributed, and available service for aggregating, collecting, and transferring huge amount of data in HDFS. It is robust in nature. Flume architecture is flexible in nature, based on data streaming.

G

35. Graph Database

A graph database is a group/collection of edges and nodes. A node typifies an entity i.e. business or individual whereas an edge typifies a relation or connection between nodes.

You must remember the statement given by graph database experts –

"If you can whiteboard it, you can graph it."

36. Grid Computing

Grid computing is a collection of computer resources for performing computing functions using resources from various domains or multiple distributed systems to reach a specific goal. A grid is designed to solve big problems to maintain the process flexibility. Grid computing is often used in scientific/marketing research, structural analysis, web services such as back-office infrastructures or ATM banking etc.

37. Gamification

Gamification refers to the principles used in designing the game to improve customer engagement in non-game businesses. Different companies use different gaming principles to enhance interest in a service or product or simply we can say gamification is used to deepen their client's relationship with the brand.

H

38. Hadoop User Experience (HUE)

Hadoop User Experience (HUE) is an open source interface which makes Apache Hadoop's use easier. It is a web-based application. It has a job designer for MapReduce, a file browser for HDFS, an Oozie application for making workflows and coordinators, an Impala, a shell, a Hive UI, and a group of Hadoop APIs.

39. High-Performance Analytical Application (HANA)

High-performance Analytical Application is a software/hardware scheme for large volume transactions and real-time data analytics in-memory computing platform from the SAP.

40. HAMA

Hama is basically a distributed computing framework for big data analytics based on Bulk Synchronous Parallel strategies for advanced and complex computations like graphs, network algorithms, and matrices. It is a Top-level Project of The Apache Software Foundation.

Big Data Analytics is the field with a number of career opportunities. Let's check out why is Big Data Analytics so important!

41. Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is primary data storage layer used by Hadoop applications. It employs DataNode and NameNode architecture to implement distributed and Java-based file system which supplies high-performance access to data with high scalable Hadoop Clusters. It is designed to be highly fault-tolerant.

42. HBase

Apache HBase is the Hadoop database which is an open source, scalable, versioned, distributed and big data store. Some features of HBase are

- Modular and linear scalability
- Easy to use Java APIs
- Configurable and automatic sharing of tables
- Extensible JIRB shell

43. Hive

Hive is an open source Hadoop-based data warehouse software project for providing data summarization, analysis, and query. Users can write queries in the SQL-like language known as HiveQL. Hadoop is a framework which handles large datasets in the distributed computing environment.

I

44. Impala

Impala is an open source MPP (massively parallel processing) SQL query engine which is used in computer cluster for running Apache Hadoop. Impala provides parallel database strategy to Hadoop so that user will be able to apply low-latency SQL queries on the data that is stored in Apache HBase and HDFS without any data transformation.

K

45. Key Value Stores / Key Value Databases

Key value store or key-value database is a paradigm of data storage which is schemed for storing, managing, and retrieving a data structure. Records are stored in a data type of a programming language with a key attribute which identifies the record uniquely. That's why there is no requirement of a fixed data model.

L

46. Load balancing

Load balancing is a tool which distributes the amount of workload between two or more computers over a computer network so that work gets completed in small time as all users desire to be served faster. It is the main reason for computer server clustering and it can be applied with software or hardware or with the combination of both.

47. Linked Data

Linked data refers to the collection of interconnected datasets that can be shared or published on the web and collaborated with machines and users. It is highly structured, unlike big data. It is used in building Semantic Web in which a large amount of data is available in the standard format on the web.

48. Location Analytics

Location analytics is the process of gaining insights from geographic component or location of business data. It is the visual effect of analyzing and interpreting the information which is portrayed by data and allows the user to connect location-related information with the dataset.

49. Log File

A log file is the special type of file that allows users keeping the record of events occurred or the operating system or conversation between the users or any running software.

M

50. Metadata

Metadata is data about data. It is administrative, descriptive, and structural data that identifies the assets.

51. MongoDB

MongoDB is an open source and NoSQL document-oriented database program. It uses JSON documents to save data structures with an agile scheme known a MongoDB BSON format. It integrates data in applications very quickly and easily.

52. Multi-Dimensional Database (MDB)

A multidimensional database (MDB) is a kind of database which is optimized for OLAP (Online Analytical Processing) applications and data warehousing. MDB can be easily created by using the input of relational database. MDB is the ability of processing data in the database so that results can be developed quickly.

53. Multi-Value Database

Multi-Value Database is a kind of multi-dimensional and NoSQL database which is able to understand three-dimensional data. These databases are enough for manipulating XML and HTML strings directly.

Some examples of Commercial Multi-value Databases are OpenQM, Rocket D3 Database Management System, jBASE, Intersystem Cache, OpenInsight, and InfinityDB.

54. Machine-Generated Data

Machine generated data is the information generated by machines (computer, application, process or another inhuman mechanism). Machine generated data is known as amorphous data as humans can rarely modify/change this data.

55. Machine Learning

Machine learning is a computer science field that makes use of statistical strategies to provide the facility to “learn” with data on the computer. Machine learning is used for exploiting the opportunities hidden in big data.

56. MapReduce

MapReduce is a processing technique to process large datasets with the parallel distributed algorithm on the cluster. MapReduce jobs are of two types. “Map” function is used to divide the query into multiple parts and then process the data at the node level. “Reduce’ function collects the result of “Map” function and then find the answer to the query. MapReduce is used to handle big data when coupled with HDFS. This coupling of HDFS and MapReduce is referred to as Hadoop.

57. Mahout

Apache Mahout is an open source data mining library. It uses data mining algorithms for regression testing, performing, clustering, statistical modeling, and then implementing them using MapReduce model.

N

58. Network Analysis

Network analysis is the application of graph/chart theory that is used to categorize, understand, and viewing relationships between the nodes in

network terms. It is an effective way of analyzing connections and to check their capabilities in any field such as prediction, marketing analysis, and healthcare etc.

59. NewSQL

NewSQL is a class of modern relational database management system which provide the scalable performance same as NoSQL systems for OLTP read/write workloads. It is well-defined database system which is easy to learn.

Want to extend your knowledge of Big Data? Here is the complete list of Big Data Blogs, just read and become a Big Data expert!

60. NoSQL

Widely known as 'Not only SQL', it is a system for the management of databases. This database management system is independent of the relational database management system. A NoSQL database is not built on tables, and it doesn't use SQL for the manipulation of data.

O

61. Object Databases

The database that stores data in the form of objects is known as the object database. These objects are used in the same manner as that of the objects used in OOP. An object database is different from the graph and relational

databases. These databases provide a query language most of the time that helps to find the object with a declaration.

62. Object-based Image Analysis

It is the analysis of object-based images that is performed with data taken by selected related pixels, known as image objects or simply objects. It is different from the digital analysis that is done using data from individual pixels.

63. Online Analytical Processing (OLAP)

It is the process by which analysis of multidimensional data is done by using three operators – drill-down, consolidation, and slice and dice.

- **Drill-down** is the capability provided to users to view underlying details
- **Consolidation** is the aggregate of available
- **Slice and dice** is the capability provided to users for selecting subsets and viewing them from various contexts

64. Online transactional processing (OLTP)

It is the big data term used for the process that provides users an access to the large set of transactional data. It is done in such a manner that users are able to derive meaning from the accessed data.

65. Open Data Center Alliance (ODCA)

OCDA is the combination of IT organizations over the globe. The main goal of this consortium is to increase the movement of cloud computing.

66. Operational Data Store (ODS)

It is defined as a location to collect and store data retrieved from various sources. It allows users to perform many additional operations on the data before it is sent for reporting to the data warehouse.

67. Oozie

It is the big data term used for a processing system that allows users to define a set of jobs. These jobs are written in different languages such as Pig, MapReduce, and Hive. Oozie allows users to link those jobs to one another.

P

68. Parallel Data Analysis

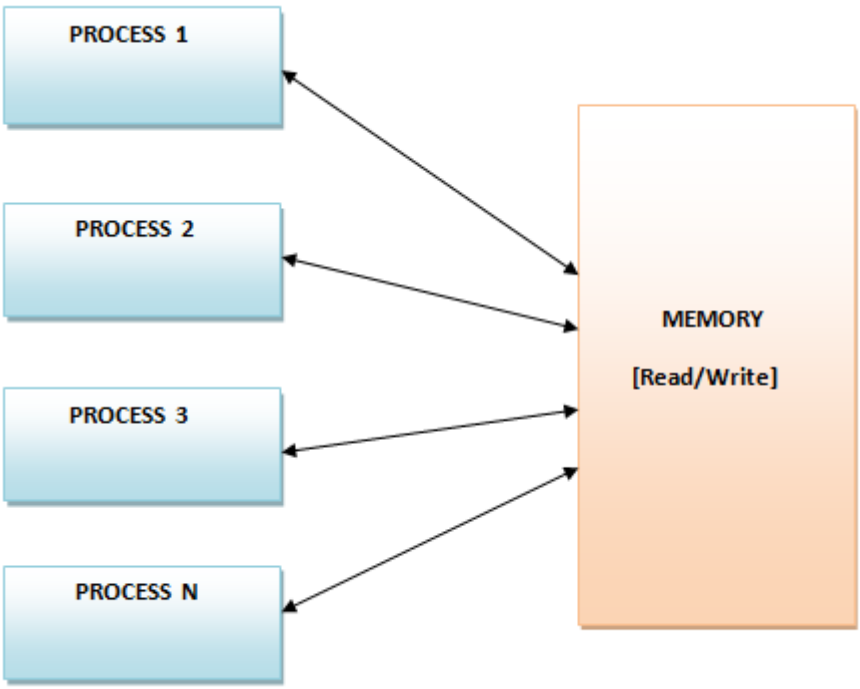
The process of breaking an analytical problem into small partitions and then running analysis algorithms on each of the partitions simultaneously is known as parallel data analysis. This type of data analysis can be run either on the different systems or on the same system.

69. Parallel Method Invocation (PMI)

It is the system that allows program code to call or invoke multiple methods/functions simultaneously at the same time.

70. Parallel Processing

It is the capability of a system to perform the execution of multiple tasks simultaneously.



71. Parallel Query

A parallel query can be defined as a query that can be executed over multiple system threads in order to improve the performance.

72. Pattern Recognition

A process to classify or label the identified pattern in the process of machine learning is known as pattern recognition.

73. Pentaho

Pentaho, a software organization, provides open source Business Intelligence products those are known as Pentaho Business Analytics. Pentaho offers OLAP services, data integration, dashboarding, reporting, ETL, and data mining capabilities.

74. Petabyte

The data measurement unit equals to 1,024 terabytes or 1 million gigabytes is known as petabyte.

Q

75. Query

A query is a method to get some sort of information in order to derive an answer to the question.

Big Data world is steadily evolving with the time. Let’s have a look at the upcoming Big Data Trends in 2018.

76. Query Analysis

The process to perform the analysis of search query is called query analysis. The query analysis is done to optimize the query to get the best possible results.

R

77. R

It is a programming language and an environment for the graphics and statistical computing. It is very extensible language that provides a number of graphical and statistical techniques such as nonlinear and linear modeling, time-series analysis, classical statistical tests, clustering, classification etc.

78. Re-identification

The data re-identification is a process that matches anonymous data with the available auxiliary data or information. This practice is helpful to find out the individual whom this data belongs to.

79. Real-time Data

The data that can be created, stored, processed, analyzed, and visualized instantly i.e. in milliseconds, is known as real-time data.

80. Reference Data

It is the big data term that defines the data used to describe an object along with its properties. The object described by reference data may be virtual or physical in nature.

81. Recommendation Engine

It is an algorithm that performs the analysis of various actions and purchases made by a customer on an e-commerce website. This analyzed data is then used to recommend some complementary products to the customer.

82. Risk Analysis

It is a process or procedure to track the risks of an action, project or decision. The risk analysis is done by applying different statistical techniques on the datasets.

83. Routing Analysis

It is a process or procedure to find the optimized routing. It is done with the use of various variables for transport to improve efficiency and reduce costs of the fuel.

S

84. SaaS

It is the big data term used for Software-as-a-Service. It allows vendors to host an application and then make this application available over the internet. The SaaS services are provided in the cloud by SaaS providers.

85. Semi-Structured Data

The data, not represented in the traditional manner with the application of regular methods is known as semi-structured data. This data is neither totally structured nor unstructured but contains some tags, data tables, and structural elements. Few examples of semi-structured data are XML documents, emails, tables, and graphs.

86. Server

The server is a virtual or physical computer that receives requests related to the software application and thus sends these requests over a network. It is the common big data term used almost in all the big data technologies.

87. Spatial Analysis

The analysis of spatial data i.e. topological and geographic data is known as spatial analysis. This analysis helps to identify and understand everything about a particular area or position.

88. Structured Query Language (SQL)

SQL is a standard programming language that is used to retrieve and manage data in a relational database. This language is very useful to create and query relational databases.

89. Sqoop

It is a connectivity tool that is used to move data from non-Hadoop data stores to Hadoop data stores. This tool instructs Sqoop to retrieve data from Teradata, Oracle or any other relational database and to specify target destination in Hadoop to move that retrieved data.

90. Storm

Apache Storm is a distributed, open source, and real-time computation system used for data processing. It is one of the must-known big data terms, responsible to process unstructured data reliably in real-time.

A big data certification validates your Big Data skills and helps you stand out of the crowd. Here is the list of best Big Data Certifications in 2018.

T

91. Text Analytics

The text analytics is basically the process of the application of linguistic, machine learning, and statistical techniques on the text-based sources. The text analytics is used to derive an insight or meaning from the text data by application of these techniques.

92. Thrift

It is a software framework that is used for the development of the ascendable cross-language services. It integrates code generation engine with the software stack to develop services that can work seamlessly and efficiently between different programming languages such as Ruby, Java, PHP, C++, Python, C# and others.

U

93. Unstructured Data

The data for which structure can’t be defined is known as unstructured data. It becomes difficult to process and manage unstructured data. The common examples of unstructured data are the text entered in email messages and data sources with texts, images, and videos.

V

94. Value

This big data term basically defines the value of the available data. The collected and stored data may be valuable for the societies, customers, and organizations. It is one of the important big data terms as big data is meant for big businesses and the businesses will get some value i.e. benefits from the big data.

95. Volume

This big data term is related to the total available amount of the data. The data may range from megabytes to brontobytes.

W

96. WebHDFS Apache Hadoop

WebHDFS is a protocol to access HDFS to make the use of industry RESTful mechanism. It contains native libraries and thus allows to have an access of the HDFS. It helps users to connect to the HDFS from outside by taking advantage of Hadoop cluster parallelism. It also offers the access of web services strategically to all Hadoop components.

97. Weather Data

The data trends and patterns that help to track the atmosphere is known as the weather data. This data basically consists of numbers and factors. Now, real-time data is available that can be used by the organizations in a different manner. Such as a logistics company uses weather data in order to optimize goods transportation.

X

98. XML Databases

The databases that support the storage of data in XML format is known as XML database. These databases are generally connected with the document-specific databases. One can export, serial, and put a query on the data of XML database.

Y

99. Yottabyte

It is the big data term related to the measurement of data. One yottabyte is equal to 1000 zettabytes or the data stored in 250 trillion DVDs.

Z

100. ZooKeeper

It is an Apache software project and Hadoop subproject which provides open code name generation for the distributed systems. It also supports consolidated organization of the large-sized distributed systems.

101. Zettabyte

It is the big data term related to the measurement of data. One zettabyte is equal to 1 billion terabytes or 1000 exabytes.

Bottom Line

Big data is not only a buzz word but the broad term that has a lot to learn. So, we have enlisted and described these Big Data terms that will be helpful in your big data career. Not to mention, it is important to validate your big data skills and knowledge for the bright career. And big data certifications are meant to demonstrate your big data skills to the employers.

Whizlabs, the pioneer in Big Data Certifications Training, is aimed to help you learn and get certified in big data technologies. Whether you are a Hadoop or Spark professional, Whizlabs Hadoop Admin (HDPCA), Spark Developer (HDPCD), and CCA Administrator certification online training will prepare you for a bright future!

An Extensive Glossary Of Big Data Terminology

A

Aggregation – a process of searching, gathering and presenting data
Algorithms – a mathematical formula that can perform certain analyses on data
Analytics – the discovery of insights in data
Anomaly detection – the search for data items in a dataset that do not match a projected pattern or expected behaviour. Anomalies are also called outliers, exceptions, surprises or contaminants and they often provide critical and actionable information.
Anonymization – making data anonymous; removing all data points that could lead to identify a person
Application – computer software that enables a computer to perform a certain task
Artificial Intelligence – developing intelligence machines and software that are capable of perceiving the environment and take corresponding action when required and even learn from those actions.

B

Behavioural Analytics – analytics that informs about the how, why and what instead of just the who and when. It looks at humanized patterns in the data

Big Data Scientist – someone who is able to develop the algorithms to make sense out of big data
Big data startup – a young company that has developed new big data technology
Biometrics – the identification of humans by their characteristics
Brontobytes– approximately 1000 Yottabytes and the size of the digital universe tomorrow. A Brontobyte contains 27 zeros
Business Intelligence – the theories, methodologies and processes to make data understandable

C

Classification analysis - a systematic process for obtaining important and relevant information about data, also meta data called; data about data.
Cloud computing– a distributed computing system over a network used for storing data off-premises
Clustering analysis– the process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.
Cold data storage– storing old data that is hardly used on low-power servers. Retrieving the data will take longer
Comparative analysis– it ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.
Complex structured data– data that are composed of two or more complex, complicated, and interrelated parts that cannot be easily interpreted by structured query

languages and tools.
Computer generated data– data generated by computers such as log files
Concurrency – performing and executing multiple tasks and processes at the same time
Correlation analysis– the analysis of data to determine a relationship between variables and whether that relationship is negative (- 1.00) or positive (+1.00).
Customer Relationship Management – managing the sales and business processes, big data will affect CRM strategies

D

Dashboard – a graphical representation of the analyses performed by the algorithms
Data aggregation tools - the process of transforming scattered data from numerous sources into a single new one.
Data analyst – someone analysing, modelling, cleaning or processing data
Database – a digital collection of data stored via a certain technique
Database-as-a-Service – a database hosted in the cloud on a pay per use basis, for example Amazon Web Services
Database Management System– collecting, storing and providing access of data
Data centre – a physical location that houses the servers for storing data
Data cleansing– the process of reviewing and revising

data in order to delete duplicates, correct errors and provide consistency
Data custodian– someone who is responsible for the technical environment necessary for data storage
Data ethical guidelines – guidelines that help organizations being transparent with their data, ensuring simplicity, security and privacy
Data feed – a stream of data such as a Twitter feed or RSS
Data marketplace – an online environment to buy and sell data sets
Data mining – the process of finding certain patterns or information from data sets
Data modelling – the analysis of data objects using data modelling techniques to create insights from the data
Data set – a collection of data
Data virtualization – a data integration process in order to gain more insights. Usually it involves databases, applications, file systems, websites, big data techniques, etc.)
De-identification – same as anonymization; ensuring a person cannot be identified through the data
Discriminant analysis - cataloguing of the data; distributing data into groups, classes or categories. A statistical analysis used where certain groups or clusters in data are known upfront and that uses that information to derive the classification rule.
Distributed File System – systems that offer simplified, highly available access to storing, analysing and processing data
Document Store Databases– a document-

oriented database that is especially designed to store, manage and retrieve documents, also known as semi structured data.

E

Exploratory analysis– finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.
Exabytes– approximately 1000 petabytes or 1 billion gigabytes. Today we create one Exabyte of new information globally on a daily basis.
Extract, Transform and Load (ETL) – a process in a database and data warehousing meaning extracting the data from various sources, transforming it to fit operational needs and loading it into the database

F

Failover – switching automatically to a different server or node should one fail
Fault-tolerant design – a system designed to continue working even if certain parts fail

G

Gamification– using game elements in a non game context; very useful to create data therefore coined as the friendly scout of big data
Graph Databases– they use graph structures (a finite set of ordered pairs or certain entities), with edges, properties

and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbour element.
Grid computing– connecting different computer systems from various location, often via the cloud, to reach a common goal

H

Hadoop – an open-source framework that is built to enable the process and storage of big data across a distributed file system
HBase – an open source, non-relational, distributed database running in conjunction with Hadoop
HDFS – Hadoop Distributed File System; a distributed file system designed to run on commodity hardware
High-Performance-Computing (HPC) – using supercomputers to solve highly complex and advanced computing problems

I

In-memory – a database management system stores data on the main memory instead of the disk, resulting is very fast processing, storing and loading of the data
Internet of Things – ordinary devices that are connected to the internet at any time any where via sensors

J

Juridical data compliance – relevant when you use cloud solutions and where the data is stored in a different

country or continent. Be aware that data stored in a different country has to oblige to the law in that country.

K

KeyValue Databases– they store data with a primary key, a uniquely identifiable record, which makes easy and fast to look up. The data stored in a KeyValue is normally some kind of primitive of the programming language.

L

Latency – a measure of time delayed in a system
Legacy system– an old system, technology or computer system that is not supported any more
Load balancing – distributing workload across multiple computers or servers in order to achieve optimal results and utilization of the system
Location data – GPS data describing a geographical location
Log file – a file automatically created by a computer to record events that occur while operational

M

Machine2Machine data – two or more machines that are communicating with each other
Machine data – data created by machines via sensors or algorithms
Machine learning – part of artificial intelligence where machines learn from what they are doing and become better over time

MapReduce – a software framework for processing vast amounts of data
Massively Parallel Processing (MPP) – using many different processors (or computers) to perform certain computational tasks at the same time
Metadata – data about data; gives information about what the data is about.
MongoDB – an open-source NoSQL database
Multi-Dimensional Databases – a database optimized for data online analytical processing (OLAP) applications and for data warehousing.
MultiValue Databases– they are a type of NoSQL and multidimensional databases that understand 3 dimensional data directly. They are primarily giant strings that are perfect for manipulating HTML and XML strings directly

N

Natural Language Processing– a field of computer science involved with interactions between computers and human languages
Network analysis– viewing relationships among the nodes in terms of the network or graph theory, meaning analysing connections between nodes in a network and the strength of the ties.
NewSQL– an elegant, well-defined database system that is easier to learn and better than SQL. It is even newer than NoSQL
NoSQL – sometimes referred to as ‘Not only SQL’ as it is a database that doesn’t adhere to traditional relational

database structures. It is more consistent and can achieve higher availability and horizontal scaling.

O

Object Databases – they store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.
Object-based Image Analysis – analysing digital images can be performed with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.
Operational Databases– they carry out regular operations of an organisation and are generally very important to a business. They generally use online transaction processing that allows them to enter, collect and retrieve specific information about the company.
Optimization analysis - the process of optimization during the design cycle of products done by algorithms. It allows companies to virtually design many different variations of a product and to test that product against pre-set variables.
Ontology– ontology represents knowledge as a set of concepts within a domain and the relationships between those concepts
Outlier detection – an outlier is an object that deviates significantly from the general average within a dataset or a combination of data. It is numerically distant from the rest of the data and therefore, the outlier indicates that

something is going on and generally therefore requires additional analysis.

P

Pattern Recognition – identifying patterns in data via algorithms to make predictions of new data coming from the same source.
Petabytes - approximately 1000 terabytes or 1 million gigabytes. The CERN Large Hydron Collider generates approximately 1 petabyte per second
Platform-as-a-Service – a services providing all the necessary infrastructure for cloud computing solutions
Predictive analysis – the most valuable analysis within big data as they help predict what someone is likely to buy, visit, do or how someone will behave in the (near) future. It uses a variety of different data sets such as historical, transactional, social or customer profile data to identify risks and opportunities.
Privacy – to seclude certain data / information about oneself that is deemed personal
Public data – public information or data sets that were created with public funding

Q

Quantified Self – a movement to use application to track ones every move during the day in order to gain a better understanding about ones behaviour
Query – asking for information to answer a certain question

R

Re-identification – combining several data sets to find a certain person within anonymized data
Regression analysis – to define the dependency between variables. It assumes a one-way causal effect from one variable to the response of another variable.
RFID– Radio Frequency Identification; a type of sensor using wireless non-contact radio-frequency electromagnetic fields to transfer data
Real-time data– datathat is created, processed, stored, analysed and visualized within milliseconds
Recommendation engine – an algorithm that suggests certain products based on previous buying behaviour or buying behaviour of others
Routing analysis– finding the optimized routing using many different variables for a certain means of transport in order to decrease fuel costs and increase efficiency.

S

Semi-structured data - a form a structured data that does not have a formal structure like structured data. It does however have tags or other markers to enforce hierarchy of records.
Sentiment Analysis– using algorithms to find out how people feel about certain topics
Signal analysis– it refers to the analysis of measurement of time varying or spatially varying physical quantities to analyse the performance of a product. Especially used with sensor data.
Similarity searches – finding the closest object to a query

in a database, where the data object can be of any type of data.
Simulation analysis – a simulation is the imitation of the operation of a real-world process or system. A simulation analysis helps to ensure optimal product performance taking into account many different variables.
Smart grid– refers to using sensors within an energy grid to monitor what is going on in real-time helping to increase efficiency
Software-as-a-Service – a software tool that is used of the web via a browser
Spatial analysis– refers to analysing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.
SQL – a programming language for retrieving data from a relational database
Structured data – data that is identifiable as it is organized in structure like rows and columns. The data resides in fixed fields within a record or file or the data is tagged correctly and can be accurately identified.

T

Terabytes – approximately 1000 gigabytes. A terabyte can store up to 300 hours of high-definition video
Time series analysis - analysing well-defined data obtained through repeated measurements of time. The data has to be well defined and measured at successive points in time spaced at identical time intervals.
Topological Data Analysis– focusing on the shape of

complex data and identifying clusters and any statistical significance that is present within that data.
Transactional data– dynamic data that changes over time
Transparency – consumers want to know what happens with their data and organizations have to be transparent about that

U

Un-structured data - unstructured data is regarded as data that is in general text heavy, but may also contain dates, numbers and facts.

V

Value– all that available data will create a lot of value for organizations, societies and consumers. Big data means big business and every industry will reap the benefits from big data.
Variability– it means that the meaning of the data can change (rapidly). In (almost) the same tweets for example a word can have a totally different meaning
Variety – data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data
Velocity – the speed at which the data is created, stored, analysed and visualized
Veracity – organizations need to ensure that the data is correct as well as the analyses performed on the data are correct. Veracity refers to the correctness of the data
Visualization – with the right visualizations, raw data can be put to use. Visualizations of course do not mean

ordinary graphs or pie-charts. They mean complex graphs that can include many variables of data while still remaining understandable and readable
Volume– the amount of data, ranging from megabytes to brontobytes

W

Weather data – an important open public data source that can provide organisations with a lot of insights if combined with other sources

X

XML Databases – XML Databases allow data to be stored in XML format. XML databases are often linked to document-oriented databases. The data stored in an XML database can be queried, exported and serialized into any format needed.

Y

Yottabytes– approximately 1000 Zettabytes, or 250 trillion DVD's. The entire digital universe today is 1 Yottabyte and this will double every 18 months.

Z

Zettabytes – approximately 1000 Exabytes or 1 billion terabytes. Expected is that in 2016 over 1 zettabyte will cross our networks globally on a daily basis.

IMPORTANT BIG DATA TERMS EVERYONE SHOULD KNOW

These Big Data Terms; It'll impress anyone you'll meet, even if you're new to this field or you just steped to swim in the Big Data's Ocean!

Let's get started:

1. **Algorithm:** A mathematical formula or statistical process used to perform an analysis of data. How is 'Algorithm' related to Big Data? Even though algorithm is a generic term, Big Data analytics made the term contemporary and more popular.

2. **Analytics:** Most likely, your credit card company sent you year-end statements with all your transactions for the entire year. What if you dug into it to see what % you spent on food, clothing, entertainment etc? You are doing 'analytics'. You are drawing insights from your raw data which can help you make decisions regarding spending for the upcoming year. What if you did the same exercise on tweets or facebook posts by an entire city's population? Now we are talking Big Data analytics. It is about making inferences and story-telling with large sets of data. There are 3 different types of analytics, so let's discuss them while we are on this topic.

3. **Descriptive Analytics:** If you just told me that you spent 25% on food, 35% on clothing, 20% on entertainment and the rest on miscellaneous items last year using your credit card, that is descriptive analytics. Of course, you can go into lot more detail as well.

4. **Predictive Analytics:** If you analyzed your credit card history for the past 5 years and the split is somewhat consistent, you can safely forecast with high probability that next year will be similar to past years. The fine print here is that this is not about 'predicting the future' rather 'forecasting with probabilities' of what might happen. In Big Data predictive analytics, data scientists may use advanced techniques like machine learning, and advanced statistical processes (we'll discuss all these terms later) to forecast the weather, economic changes, etc.

5. **Prescriptive Analytics:** Still using the credit card transactions example, you may want to find out which spending to target (i.e. food, entertainment, clothing etc.) to make a huge impact on your overall spending. Prescriptive analytics builds on predictive analytics by including 'actions' (i.e. reduce food or clothing or entertainment) and analyzing the resulting outcomes to 'prescribe' the best category to target to reduce your overall spend. You can extend this to Big Data and imagine how executives can make data-driven decisions by looking at the impacts of various actions in front of them.

6. **Batch processing:** Even though Batch data processing has been around since mainframe days, it gained additional significance with Big Data given the large data sets that it deals with. Batch data processing is an efficient way of processing high volumes of data where a group of transactions is collected over a period of time. Hadoop, which I'll describe later, is focused on batch data processing.

7. **Cassandra** is a popular open source database management system managed by The Apache Software Foundation. Apache can be credited with many big data technologies and Cassandra was designed to handle large volumes of data across distributed servers.

8. **Cloud computing:** Well, cloud computing has become ubiquitous so it may not be needed here but I included just for completeness sake. It's essentially software and/or data hosted and running on remote servers and accessible from anywhere on the internet.

9. **Cluster computing:** It's a fancy term for computing using a 'cluster' of pooled resources of multiple servers. Getting more technical, we might be talking about nodes, cluster management layer, load balancing, and parallel processing, etc.

10. **Dark Data:** This, in my opinion, is coined to scare the living daylights out of senior management. Basically, this refers to all the data that is gathered and processed by enterprises not used for any meaningful purposes and hence it is 'dark' and may never be analyzed. It could be social network feeds, call center logs, meeting notes and what have you. There are many estimates that anywhere from 60-90% of all enterprise data may be 'dark data' but who really knows.

11. **Data lake:** When I first heard of this, I really thought someone was pulling an April fool's joke. But it's a real term! A Data Lake is a large repository of enterprise-wide data in raw format. While we are here, let's talk about **Data** warehouses, which are similar in concept in that they, too, are repositories for enterprise-wide data – but in a **structured format after cleaning and integrating** with other sources. **Data warehouses** are typically used for conventional data (but not exclusively). Supposedly, a data lake makes it easy to access enterprise-wide data you really need to know what you are looking for and how to process it and make intelligent use of it.

12. **Data mining:** Data mining is about finding meaningful patterns and deriving insights in large sets of data using sophisticated pattern recognition techniques. It is closely related the term Analytics that we discussed earlier in that you mine the data to do analytics. To derive meaningful patterns, data miners use statistics (yup, good old math), machine learning algorithms, and artificial intelligence.

13. **Data Scientist:** Talk about a career that is HOT! It is someone who can make sense of big data by extracting raw data (did you say from data lake?), massage it, and come up with insights. Some of the skills required for data scientists are what a superman/woman would have: analytics, statistics, computer science, creativity, story-telling and understand business context. No wonder they are so highly paid.

14. **Distributed File System:** As big data is too large to store on a single system, Distributed File System is a data storage system meant to store large volumes of data across multiple storage devices and will help decrease the cost and complexity of storing large amounts of data.

15. **ETL:** ETL stands for extract, transform, and load. It refers to the process of 'extracting' raw data, 'transforming' by cleaning/enriching the data for 'fit for use' and 'loading' into the appropriate repository for the system's use. Even though it originated with data warehouses, ETL processes are used while 'ingesting i.e. taking/absorbing data from external sources in big data systems.

16. **Hadoop:** When people think of big data, they immediately think about Hadoop. Hadoop (with its cute elephant logo) is an open source software framework that consists of what is called a Hadoop Distributed File System (HDFS) and allows for storage, retrieval, and analysis of very large data sets using distributed hardware. If you really want to impress someone, talk about YARN (Yet Another Resource Scheduler) which, as the name says it, is a resource scheduler. I am really impressed by the folks who come up with these names. Apache foundation, which came up with Hadoop, is also responsible for Pig, Hive, and Spark (yup, they are all names of various software pieces). Aren't you impressed with these names?

17. **In-memory computing:** In general, any computing that can be done without accessing I/O is expected to be faster. In-memory computing is a technique to moving the working datasets entirely within a cluster's collective memory and avoid writing intermediate calculations to disk. Apache Spark is is an in-memory computing system and it has huge advantage in speed over I/O bound systems like Hadoop's MapReduce.

18. **IoT:** The latest buzzword is Internet of Things or IOT. IOT is the interconnection of computing devices in embedded objects (sensors, wearables, cars, fridges etc.) via internet and they enable sending / receiving data. IOT generates huge amounts of data presenting many big data analytics opportunities.

19. **Machine learning:** Machine learning is a method of designing systems that can learn, adjust, and improve based on the data fed to them. Using predictive and statistical algorithms that are fed to these machines, they learn and continually zero in on "correct" behavior and insights and they keep improving as more data flows through the system. Fraud detection, online recommendations based.

20. **MapReduce:** MapReduce could be little bit confusing but let me give it a try. MapReduce is a programming model and the best way to understand this is to note that Map and Reduce are two separate items. In this, the programming model first breaks up the big data dataset into pieces (in technical terms into 'tuples' but let's not get too technical here) so it can be distributed across different computers in different locations (i.e. cluster computing described earlier) which is essentially the Map part. Then the model collects the results and 'reduces' them into one report. MapReduce's data processing model goes hand-in-hand with hadoop's distributed file system.

21. **NoSQL:** It almost sounds like a protest against 'SQL (Structured Query Language) which is the bread-and-butter for traditional Relational Database Management Systems (RDBMS) but NOSQL actually stands for Not ONLY SQL :-). NoSQL actually refers to database management systems that are designed to handle large volumes of data that does not have a structure or what's technically called a 'schema' (like relational databases have). NoSQL databases are often well-suited for big data systems because of their flexibility and distributed-first architecture needed for large unstructured databases.

22. **R:** Can anyone think of any worse name for a programming language? Yes, 'R' is a programming language that works very well with statistical computing. You ain't a data scientist if you don;'t know 'R'. (Please don't send me nasty grams if you don't know 'R'). It is just that 'R' is one of the most popular languages in data science.

23. **Spark (Apache Spark):** Apache Spark is a fast, in-memory data processing engine to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. Spark is generally a lot faster than MapReduce that we discussed earlier.

24. **Stream processing:** Stream processing is designed to act on real-time and streaming data with "continuous" queries. Combined with streaming analytics i.e. the ability to continuously calculate mathematical or statistical analytics on the fly within the stream, stream processing solutions are designed to handle very high volume in real time.

25. **Structured v Unstructured Data:** This is one of the 'V's of Big Data i.e.Variety. Structured data is basically anything than can be put into relational databases and organized in such a way that it relates to other data via tables. Unstructured data is everything that can't – email messages, social media posts and recorded human speech etc.

Big Data & Analytics Terms

Big Data stands true to its name. With billion bytes of data being collected, understanding the intricacies of Big Data has become more than a necessity.

However, Big Data often turns out to be complicated because of its highly technical lingo. Not everyone is a highly technical big data practitioner or corporate executive. The common user is often lost over these complex terminologies. Let's take a dive into Big Data Glossary.

We have come up with a list of Big Data terms and definitions, that would serve as a guide for beginners. Our list comprises of extensive terminologies, from the basics to the advanced, would help you get a clear understanding of Big Data terms.

Big Data and Analytics Glossary from A-D

A

- **Analytics**: Analytics refers to the course of depicting conclusions based on the raw

data. it helps us sort out meaningful data from data mass.

- **Automatic Identification and Capture (AIDC)**: Automatic Identification and Data Capture (AIDC) refers to a broad set of technologies used to glean data from an object, image, or sound without manual intervention.
- **Algorithm**: The algorithm refers to a mathematical formula placed in software that performs an analysis on a set of data.
- **Artificial Intelligence**: Artificial Intelligence refers to the process of developing intelligence machines and software that can perceive the environment and take the corresponding action as and when required and even learn from those actions.
- **ACID Test**: ACID Test stands for atomicity, consistency, isolation, and durability (ACID) test of data. These four attributes are the

benchmarks for ensuring the validity of a data transaction.

- **Apache Avro**: Apache Avro is a row-oriented object container storage format for Hadoop as well as a remote procedure call and data serialization framework. Avro is optimized for write operations and includes a wire format for communication between nodes. Avro ensures simpler translation between different nodes by way of the data definition and serialized permanent data. Avro uses JavaScript object notation to define protocols and data types, as well as serializes data into a compact binary format.

B

- **Batch Processing**: Batch processing is a standard computing strategy that involves processing data in large sets. This practice becomes imperative for non-time sensitive work that operates on very large datasets.

The process is scheduled and at a later time, the results are retrieved by the system.

- **Big Data**: Big Data is an umbrella term used for huge volumes of heterogeneous datasets that cannot be processed by traditional computers or tools due to their varying volume, velocity, and variety.
- **Biometrics**: Biometrics implies using analytics and technology in identifying people by one or many of their physical characteristics, such as fingerprint recognition, facial recognition, iris recognition, etc. It is most commonly used in modern smartphones.
- **Business Intelligence**: Business Intelligence is the general term used for the identification, extraction, and analysis of data.

- **Call Detail Record (CDR) analysis**: CDRs include data that a telecommunications company collects about phone calls. This may include call duration and time when the call was made or received. This data is used in any number of analytical applications.
- **Cascading**: Cascading refers to s a higher level of abstraction for Hadoop, that allows developers to create complex jobs quickly, easily, and in several different languages that run in the JVM, including Ruby, Scala, and more.
- **Cassandra**: Cassandra is a popular open-source database management system managed by The Apache Software Foundation. Cassandra was designed to handle large volumes of data across distributed servers.

- **Clickstream Analytics**: It refers to the analysis of users' online activity based on the items that users click on a web page.
- **Clojure**: Clojure is a functional programming language constructed in LISP which uses the JVM (Java Virtual Machine). Clojure is particularly suitable for parallel data processing.
- **Cluster computing**: Clustered Computing is the practice of segmenting the resources of multiple machines and managing their collective capabilities to complete tasks in a more simplified manner. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.

D

- **Database-as-a-Service:** Database-as-a-Service refers to a database hosted in the

cloud on a pay per use basis. For example, Amazon Web Services.

- **Data cleansing**: It refers to the process of reviewing and revising data in order to delete duplicates, correct errors and provide consistency.
- **Dark Data**: Dark Data refers to all the data that is gathered and processed by enterprises not used for any meaningful purposes. It is called dark because it is unused and unexplored. This includes social network feeds, call center logs, meeting notes, etc.

- **Data Lake**: The term Data Lake refers to a storage repository that can hold a huge amount of raw data in its original format. Data Lake uses a flat architecture to store data, unlike a hierarchical data warehouse, which stores data in files or folders. Each data element in a Data Lake is assigned a unique identifier and tagged with a set of

extended metadata tags. The Data Lake can be queried for relevant data, and a smaller set of data be analyzed to answer any relevant business question.

- **Data Mining**: Data mining is a broad term used for finding rational patterns in large data sets. The purpose is to refine a humongous data into a more comprehensible and cohesive set of information.
- **Data Modelling:** Data Modelling is defined as the analysis of data objects using data modeling techniques to create insights from the data.
- **Data Science**: Data Science is a broad subject that incorporates statistics, data visualization, computer programming, data mining, machine learning, and database engineering to solve complex problems.

- **Data Virtualization:** It refers to a data integration process to gain more insights. Usually, it involves databases, applications, file systems, websites, big data techniques, etc.

Big Data and Analytics Glossary from E-H

E

- **Econometrics**: Econometrics is the application of statistical and mathematical theories in economics for testing hypotheses and forecasting future trends. Econometrics makes use of economic models, tests them through statistical trials and then compare the results against real-life examples. It can be subdivided into two major categories: theoretical and applied.
- **ETL:** ETL is the acronym for extract, transform, and load. It refers to the process of 'extracting' raw data, 'transforming' by cleaning/enriching the data for 'fit for use'

and ‘loading’ into the appropriate repository for future use.

- **Exabytes**: Exabyte (abbreviated as EB) is a large unit of computer data storage, two to the sixtieth power bytes. The prefix *ex-* means one billion, or one quintillion, which is a decimal term. One exabyte is equal to 1,000 petabytes and precedes the zettabyte unit of measurement. Exabytes are slightly smaller than exbibytes, which contain 1,152,921,504,606,846,976 (260) bytes.
- **Exploratory Analysis**: Exploratory Data Analysis (EDA) is an approach to analyzing data. EPA is often the first step in data analysis, implemented before any formal statistical techniques are applied. Exploratory data analysis is a complement to inferential statistics, which tends to be rigid with rules and formulas. EDA involves the analyst trying to get a “feel” for the data set, often using their own judgment to

determine what the most important elements in the data set are.

F

- **Failover:** Failover is defined as the constant capability to automatically and seamlessly switch to a highly reliable backup. This can be operated in a redundant manner or in a standby operational mode upon the failure of a primary server, application, system or another primary system component. The main purpose of Failover is to eliminate, or at least reduce, the impact on users when a system failure occurs.
- **Fault-tolerant design:** It is a system designed to continue working even if certain parts fail.
- **Feature:** Feature is the machine learning expression for a piece of measurable information about something. For example, height, length, and breadth of a solid object. Other terms like property, attribute or

characteristic are also used instead of a feature.

- **Feature/Engineering:** Feature Engineering is the process of creating new input features for machine learning. It is one of the most effective ways to improve predictive models. Feature Engineering allows you to isolate key information, highlight patterns, and bring in domain expertise.
- **Feature Reduction:** Feature Reduction is the process of reducing the number of features to work on a computation-intensive task without losing much information. principal component analysis (PCA) is one of the most popular Feature Reduction techniques.
- **Feature Selection:** Feature Selection is the process of selecting relevant features for explaining the predictive power of a statistical model.

- **Flume:** Apache Flume or Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It has a simple and flexible architecture based on streaming data flows, and is robust and fault-tolerant with tunable reliability mechanisms for failover and recovery.
- **Frequentist Statistics:** Frequentist Statistics is a procedure to test the probability or improbability of an event (hypothesis). It calculates the probability of an event overall (i.e. the experiment is repeated under the same conditions to obtain the outcome). Here, the sampling distributions of fixed size are taken. Then, the experiment is theoretically repeated an infinite number of times but practically done with a stopping intention. For example, a person may decide

to stop testing an application, after 125 test cases are performed.

- **F-Score:** F-score evaluation metric combines both precision and recall as a measure of the effectiveness of classification. It is calculated in terms of the ratio of weighted importance on either recall or precision as determined by a β.

F measure = 2 x (Recall × Precision) / (β^2 × Recall + Precision)

G

- **Gamification:** Gamification is the application of game elements and digital game design techniques to non-game problems, such as business and social impact challenges. It is the process of Gamification takes the data-driven techniques that game designers use to

engage players and applies them to non-game experiences to motivate actions that add value to your business.

- **Graph Analytics:** It is a way to organize and visualize relationships between different data points in a set.
- **Grid Computing:** Grid Computing refers to performing computing functions with resources from several distributed systems. It usually involves large files and is often used for various applications. Systems containing a grid computing network are not required to be alike in design or be from the same location.

H

- **HANA:** HANA refers to a hardware/software in-memory computing tool from SAP which is intended for high-volume transactions and analytics in real-time.

- **Hadoop**: Hadoop is an open-source, a Java-based programming framework that supports the processing and storage of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. With Hadoop, you may run applications on systems with several commodity hardware nodes, and to handle thousands of terabytes of data. Its distributed file system ensures rapid data transfer rates among nodes and allows the system to continue operating in case of a node failure. This approach lowers the risk of catastrophic system failure and unexpected data loss, even if a significant number of nodes become inoperative. Hadoop was created by Doug Cutting and Mike Cafarella in 2006 to support distribution for the Nutch search engine.
- **Hive**: Hive is a Hadoop-based data warehousing-like framework originally

developed by Facebook. It allows users to write queries in a SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools such as MicroStrategy, Tableau, Revolutions Analytics, etc.

Big Data and Analytics Glossary from I-L

I

- **Imputation:** Imputation is the technique used for handling missing values in the data. It is performed by statistical metrics like mean/mode imputation or by machine learning techniques like kNN imputation.
- **In-memory:** In-memory refers to a database management system stores data on the main memory instead of the disk, for

faster processing, storage, and loading of the data.

- **In-memory Computing:** In-memory computing is the storage of information in the main random-access memory (RAM) of dedicated servers rather than in complicated relational databases operating on comparatively slow disk drives. In-memory computing helps business customers, including retailers, banks, and utilities, to quickly detect patterns, analyze massive data volumes on the fly, and perform their operations quickly. The drop-in memory prices in the present market have led to increased popularity of in-memory computing technology. This has made in-memory computing economical among a wide variety of applications.
- **In-memory data grid (IMDG):** In-memory data grid (IMDG) is a data structure that resides entirely in RAM (random access memory) and is distributed among multiple

servers. Recent advances in 64-bit and multi-core systems have made it possible to store terabytes of data completely in RAM, obviating the need for electromechanical mass storage media such as hard disks.

- **Independent Variable**: An Independent Variable is a variable that is manipulated to determine the value of a dependent variable s. The dependent variable is what is being measured in an experiment or evaluated in a mathematical equation and the independent variables are the inputs to that measurement.

- **Inferential Statistics**: Inferential Statistics refers to mathematical methods that employ probability theory for deducing (inferring) the properties of a population from the analysis of the properties of a data sample drawn from it. Inferential Statistics also deals with the precision and reliability of the inferences it helps to draw.

- **IoT:** The Internet of Things (IoT) is a system of interrelated computing devices, mechanical and digital machines, objects, animals, or people that are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.
- **IQR:** The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

J

- **Juridical Data Compliance:** In legal terminology, the word “Compliance” refers to the act of adherence to the law of the land. In business terms, in case of any

organization, compliance implies strict adherence to the laws, regulations, guidelines, and specifications that are relevant to the life cycle of a business entity. Juridical data compliance is commonly used in the context of cloud-based solutions, where the data is stored in a different country or continent. Data storage in a server or data center located in a foreign country must abide by the data security laws of the nation.

K

- **Kafka:** Kafka, a LinkedIn product, is a dispersed publish-subscribe system for messaging. It provides a solution that is proficient in conducting all activities related to data flow and processing this data over a consumer website. It is an essential element of the current social web.

- **KeyValue Databases**: A key-value database, also known as a key-value store, is the most flexible type of NoSQL database. Key-value databases have emerged as an alternative to many of the limitations of traditional relational databases, where data is structured in tables and the schema must be predefined. In a key-value store, there is no schema and the value of the data is opaque. Values are identified and accessed via a key, and stored values can be numbers, strings, counters, JSON, XML, HTML, binaries, images, short videos, and more. It is the most flexible NoSQL model because the application has complete control over what is stored in the value.
- **K-Means**: K-means clustering is a type of unsupervised learning, used for segregating unlabeled data (i.e., data without defined categories or groups). The purpose is to find groups in the data, with the number of

groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.

- **K-nearest neighbors**: K nearest neighbors (also known as kNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already at the beginning of the 1970's as a non-parametric technique.

L

- **Latency:** Latency refers to delays in transmitting or processing data. Latency is f two types, network latency, and disk latency.
- **Legacy system:** Legacy system refers to outdated computer systems, programming languages or application software that are used instead of available upgraded versions. Legacy systems are also associated with

terminology or processes that are no longer applicable to current contexts or content, thus creating confusion.

- **Linear Regression:** Linear Regression refers to a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model of apparently random data, showing trends in data, such as in cancer diagnoses or in stock prices.
- **Load Balancing:** Load balancing refers to the efficient distribution of incoming network traffic across a group of backend servers (also known as a server farm or server pool) to get more work done in the same amount of time. Load balancing can be implemented with hardware, software, or a combination of both. Typically, load balancing is the main reason for computer server clustering.

- **Location Data**: Location data refers to the information collected by a network or service about where a user's phone or device is located.
- **Logfile:** A log file is defined as a file that maintains a registry of events, processes, messages, and communication between various communicating software applications and the operating system. Log files are present in the executable software, operating systems, and programs whereby all the messages and process details are recorded.
- **Logarithm:** Logarithm refers to an exponent used in mathematical calculations to depict the perceived levels of variable quantities such as visible light energy, electromagnetic field strength, and sound intensity.
- **Logistic Regression:** Logistic regression refers to a statistical method for analyzing a dataset in which there are one or more

independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

- **Machine learning**: Machine Learning is the study and practice of designing systems that can learn, adjust, and improve based on the data fed. This typically involves the implementation of predictive and statistical algorithms that can continually zero in on "correct" behavior and insights as more data flows through the system.
- **MongoDB**: MongoDB refer to a cross-platform, open-source database that uses a document-oriented data model, rather than a traditional table-based relational database structure. It is designed to make the integration of structured and unstructured

data in certain types of applications easier and faster.

- **MapReduce**: MapReduce refers to specific tools that support distributed computing on large datasets.
- **Mashup**: Mashup refers to the method of merging different datasets into a single application to improve output. For instance, combining job listings with demographic data.
- **Mahout**: Mahout refers to a library for data mining. It uses the best prevalent data mining algorithms in performing regression testing, clustering, modeling, and implementing them with the use of the MapReduce model.
- **Metadata**: Metadata may be defined as the data that serves to provide context or additional information about other data. For example, information about the title,

subject, author, typeface, enhancements, and size of the data file of a document constitute metadata about that document. It may also describe the conditions under which the data stored in a database was acquired, its accuracy, date, time, a method of compilation and processing, etc.

- **Munging**: Munging refers to the process of manually converting or mapping data from one raw form into another format for more convenient consumption.

N

- **Natural Language Processing**: Natural Language Processing refers to the standard methods for gleaning facts from the human-created
- **Normalization**: Normalization is the process of reorganizing data in a database so that it meets two basic requirements: (1) There is no redundancy of data (all data is

stored in only one place), and (2) data dependencies are logical (all related data items are stored together). Normalization is important for many reasons, but chiefly because it allows databases to take up as little disk space as possible, resulting in increased performance. Normalization is also known as data normalization.

- **NoSQL**: NoSQL is a broad term that refers to databases designed outside of the traditional relational model. Unlike relational databases, NoSQL databases have different trade-offs. However, NoSQL databases are well-suited for big data systems because of their flexibility and frequent distributed-first architecture.

o

- **Object Databases**: Object Databases store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases

and most of them offer a query language that allows the object to be found with a declarative programming approach.

- **Object-based Image Analysis**: It refers to analyzing digital images with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.
- **Operational Databases**: Operational Databases carry out regular operations of an organization and are generally very important to a business. They generally use online transaction processing that allows them to enter, collect and retrieve specific information about the company
- **Optimization Analysis**: It refers to the process of optimization during the design cycle of products done by algorithms. It allows companies to virtually design many different variations of a product and to test that product against pre-set variables.

- **Oozie:** Oozie refers to a workflow processing system which allows its users to define a series of jobs which can be written in several languages like Pig, MapReduce, and Hive. It then intelligently links them to each other. It permits users to state, for instance, that a query is to be started only after defined previous jobs on which it depends on for data are completed.

P

- **Parse**: Parse refers to the division of data, such as a string, into smaller parts for analysis.
- **Pattern Recognition:** Pattern recognition takes place when an algorithm locates recurrences or regularities within large data sets or across disparate data sets. Pattern recognition is like machine learning and data mining.

- **Persistent storage**: Persistent storage refers to a non-changing place, such as a disk, where data is saved after the process that created it has ended.
- **Python**: Python is a general-purpose programming language that emphasizes code readability to allow programmers to use fewer lines of code to express their concepts.

Big Data and Analytics Glossary from Q-T

Q

- **Quad-Core Processor:** A quad-core processor is defined as a multiprocessor architecture that is designed to provide faster processing power. It is a successor to the dual-core processor, which has two processor cores. Quad-core processors integrate two dual-core processors into a single processor. The two separate dual cores communicate with each other using processor cache. A quad-core processor can

execute multiple instructions simultaneously, meaning that each core can be dedicated to separate instruction.

- **Query:** A query is a request for data or information from a database table or combination of tables. This data may be generated as results returned by Structured Query Language (SQL) or as pictorials, graphs, or complex results, e.g., trend analyses from data-mining tools. SQL is the most well-known and widely-used query language.
- **Quick Response Code**: A quick response code (QR code) is defined as a type of two-dimensional barcode that consists of square black modules on a white background. QR codes are designed to be read by smartphones. Because they can carry information both vertically and horizontally, they can provide a vast amount of

information, including links, text, or other data.

- **Query Analysis**: Query Analysis is a process used in databases which make use of SQL to determine how to further optimize queries for performance.
- **Quantum Bit (Qubit)**: A quantum bit (qubit) is defined as the smallest unit of quantum information, which is the quantum analog of the regular computer bit, used in the field of quantum computing. A quantum bit can exist in superposition, which means that it can exist in multiple states at once. Compared to a regular bit, which can exist in one of two states, 1 or 0, the quantum bit can exist as a 1, 0 or 1 and 0 at the same time. This allows for very fast computing and the ability to do multitudes of calculations at once, theoretically.
- **Quantum Computing**: Quantum computing is defined as a theoretical

computing model that uses a very different form of data handling to perform calculations. The emergence of quantum computing is based on a new kind of data unit that could be called non-binary, as it has more than two possible values.

R

- **R**: R is a language and environment for statistical computing and graphics. It is a GNU project which is like the S language. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible.
- **Readme**: Readme refers to a file that is attached to a software program that contains critical or important information about that program. The readme file is, in a

sense, a manual or instructional resource for that program.

- **R-Commerce**: R-Commerce or Relationship e-commerce is a form of electronic commerce that focuses upon business-to-consumer(B2C) and peer-to-peer(P2P) interaction. In r-commerce, the focus is shifted from product sales toward customer relationship building. The concept of r-commerce stems from the relationships a consumer has with other consumers and with a business by purchasing its online merchandise or services. The customer's positive experience becomes evident in relaying product information to other customers pointing out its good qualities or benefits and urges further purchases.
- **Refactoring**: Refactoring refers to the process of altering an application's source code without changing its external behavior. Code refactoring improves the

nonfunctional properties of the code, such as readability, complexity, maintainability, and extensibility.

- **Relational Data Model**: A Relational Data Model involves the use of data tables that collect groups of elements into relations. These models work are based on the idea that each table setup will include a primary key or identifier. Other tables use that identifier to provide "relational" data links and results. Database administrators use Structured Query Language (SQL) to retrieve data elements from a relational database.

- **Regression Testing**: It refers to a type of software testing used to determine whether new problems are the result of software changes. A program is tested prior to a change. The same program is re-tested after the change is implemented to determine whether any new bugs or issues are created. Regression Testing is also required to check

if the actual change achieved its intended purpose.

- **Reseller Hosting**: Reseller hosting is a business development model provided by a Web hosting service/provider. One or more organizations make use of Reseller hosting to lease out Webspace that is packaged, rebranded, and sold under their brand name. Depending on the primary hosting service provider, each reseller host may be provided with a completely white/private labeled control panel to manage their leased space and their customers.

- S

- **Spark**: Apache Spark is a fast, in-memory data processing engine to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. Spark is generally a lot faster than MapReduce.

- **Spatial Analysis**: It refers to analyzing spatial data such as geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.
- **Software as a service (SaaS)**: SaaS refers to an application software that is used over the web by a thin client or web browser. For example
- **Serialization**: Serialization refers to the standard procedures for converting data structure or object state into standard storage formats.
- **Sqoop**: Sqoop is a connectivity tool for moving data from non-Hadoop data stores – such as relational databases and data warehouses – into Hadoop. It allows users to specify the target location inside of Hadoop. Accordingly, instructions are sent to Sqoop to move data from Oracle,

Teradata or other relational databases to the target.

- **Storm**: Storm is a free and open-source real-time distributed computing system. Storm makes processing of unstructured data easier and faster with instantaneous processing. It uses Hadoop for batch processing.
- **Stream Processing**: Stream Processing is the standard practice of computing over individual data items as they move through a system. This allows for real-time analysis of the data being fed into the system. It is useful for time-sensitive operations using high-velocity
- **Structured Query Language (SQL):** SQL refers to a programming language designed specifically to manage and retrieve data from a relational database system.

- **Tag**: A tag is a piece of information that describes the data or content that it is assigned to. Tags are nonhierarchical keywords used for Internet bookmarks, digital images, videos, files and so on.
- **Talking Trojan**: Talking Trojan is a kind of Trojan virus introduced in 2007 that replays an audio message while it deletes the contents of a hard drive or otherwise attacks a system. This is a type of Trojan program, a virus that looks legitimate but attacks the user system when it is run.
- **Taxonomy**: Taxonomy refers to the classification of data according to a pre-determined system with the resulting catalog. It provides a conceptual framework for easy access and retrieval.
- **TensorFlow**: TensorFlow is a free software library focused on machine learning created

by Google. Initially released as part of the Apache 2.0 open-source license, TensorFlow was originally developed by engineers and researchers of the Google Brain Team, mainly for internal use.

- **Transactional Data:** This refers to data that changes unpredictably. Examples include accounts payable and receivable data or data about product shipments.
- **Thrift:** Thrift is a software structure for accessible cross-language services development. Thrift pools a stack of software with a code generation engine to form services which work proficiently and flawlessly between Java, C++, Python, Ruby, PHP, Erlang, Haskell, Perl, and C#.

Big Data and Analytics Glossary from U-Z

U

- **Unique Constraint**: A unique constraint refers to a type of column restriction within

a table, which dictates that all values in that column must be unique though may be null. To ensure that a column is UNIQUE and cannot contain null values, the column must be specified as NOT NULL. Interestingly, these are a primary key's two main attributes. Defining both attributes in a newly-created column should be given serious consideration for the primary key designation.

- **Unique Visitor**: Unique visitor refers to a person who visits a site at least once during the reporting period. A unique visitor is also known as a unique user.
- **Unstructured data:** Data that either does not have a pre-defined data model or is not organized in a pre-defined manner.
- **Unstructured Data Analysis**: Unstructured data analysis refers to the process of analyzing data objects that don't follow a

predefined data model /architecture and/or is unorganized. It refers to the analysis of any data that is stored over time within an organizational data repository without any intent for its orchestration, pattern, or categorization.

- **Unstructured Data Mining**: Unstructured data mining refers to the practice of looking at relatively unstructured data and trying to get more refined data sets out of it. It often consists of extracting data from sources not traditionally used for data mining activities.

v

- **Vector Graphics Rendering:** Vector graphics rendering is the process of generating models from geometrical primitives such as lines, points, curves, and shapes to represent images in computer graphics.

- **Vector Markup Language (VML)**: Vector Markup Language (VML) refers to an application of XML 1.0 that defines the encoding of vector graphics in HTML. It was submitted to the W3C in 1998 but never gained traction. Instead, a working group at the W3C created Scalable Vector Graphics (SVG), which became a W3C Recommendation in 2001.
- **Vertical Scalability**: Vertical scalability is defined as the addition of resources to a single system node, such as a single computer or network station, which often results in additional CPUs or memory. Vertical scalability provides more shared resources for the operating system and applications.
- **View-Based Conversions**: View-based conversions are the result of a tracking method employed by Google AdWords.

- **Virtual Address Extension (VAX)**: A Virtual Address Extension (VAX) was a midrange server computer developed in the late 1970s by Digital Equipment Corporation (DEC). The VAX was introduced as mainframe computers were being developed. The VAX computer had a 32-bit processor and a virtual memory setup
- **Visualization**: Visualization refers to applications that are used for graphical representation of meaningful data.

W

- **WannaCry**: WannaCry is a kind of ransomware attack that developed in the spring of 2017 and brought the idea of ransomware threats further into the mainstream. This global attack disabled many systems, including public-service systems such as those supporting hospitals and law-enforcement officers. Experts classified WannaCry as a cryptoworm.

- **Web Application Security**: Web application security refers to the process of securing confidential data stored online from unauthorized access and modification.
- **Wearable Robot**: It refers to a specific type of wearable device that is used to enhance a person's motion and/or physical abilities. Wearable robots are also known as bionic robots or exoskeletons.
- **Whiteboarding**: Whiteboarding refers to the manipulation of digital files on a visual digital whiteboard. It is used for different kinds of collaborative projects and represents a useful form of data visualization in which a sequential series of files can be shown on a screen as a visual object-based model.
- **Wireless Access Point:** A wireless access point (WAP) refers to a hardware device or configured node on a local area network (LAN) that allows wireless capable devices

and wired networks to connect through a wireless standard, including Wi-Fi or Bluetooth. WAPs feature radio transmitters and antennae, which facilitate connectivity between devices and the Internet or a network.

- **Xanadu**: Xanadu refers to a hypertext/hypermedia project first conceptualized by Ted Nelson.
- **X Client:** X client refers to the application program that is displayed on an X server, although this application program is otherwise separate from that server. All application programs that run in a GUI delivered by the X Window System, which is virtually any GUI employed on Linux as well as other Unix-like operating systems, is considered an X client. Therefore Apache, OpenOffice, gFTP, gedit, GIMP, Xpdf, and

rCalc are typically X clients if employed on such operating systems.

- **XML Databases**: XML Databases allow data to be stored in XML format. XML databases are often linked to document-oriented databases. The data stored in an XML database can be queried, exported, and serialized into any format needed.
- **XML-Query Language**: XML Query Language or XQuery refers to a specific query and programming language for processing XML documents and data. XML data and other databases that store data in a format analogous to HTML can be processed with XQuery.
- **X Server**: X server refers to a server program that connects X terminals running on the X Window System, whether locally or in a distributed network. The X server is installed with the X Window System, which is a cross-platform and complete client-

server system for managing graphical user interfaces on a single computer or networked ones.

- **X Terminal**: An X terminal is an input terminal with a display, keyboard, mouse, and touchpad that uses X server software to render images. Used as an open-source windowing system known as the X Window System, the X terminal does not perform application processing – this is handled by the network server.

Y

- **YMODEM**: YMODEM refers to an asynchronous communication protocol for modems developed by Chuck Forsberg as a successor to Xmodem and Modem7. It supports batch file transfers and increases transfer block size, enabling the transmission of a whole list or batch of files at one time.

- **Yoda Condition**: Yoda condition refers to a scenario when a piece of computer syntax is inverted or swapped around, for example, where instead of declaring a variable equal to a constant, the programmer declares a constant equal to a variable. A key characteristic of Yoda conditions is that they do not impair the function of the code in any way.
- **Yoyo Mode**: Yoyo mode refers to a situation wherein a computer or a similar device seems stuck in a loop — turning on briefly, then turning off again. The idea is that the rapid restart and shut off patterns can be compared to the down and up cycles of a yo-yo.
- **Yottabytes:** Yottabytes is a data unit. It is equivalent to approximately 1000 Zettabytes, or 250 trillion DVDs. The entire digital universe is estimated to be 1

Yottabyte and this is expected to be double every 18 months.

Z

- **Zachman Framework**: Zachman Framework refers to a visual aid for organizing ideas about enterprise technology. It is attributed to IBM professional John Zachman, as presented in the article “A Framework for Information Systems Architecture” published in the IBM Systems Journal in 1987.
- **Zend Optimizer**: Zend Optimizer refers to an open-source runtime application used with file scripts encoded by Zend Encoder and Zend Safeguard to boost the overall PHP application runtime speed.
- **Zettabytes**: Zettabytes is another data unit, that equals to approximately 1000 Exabytes or 1 billion terabytes.

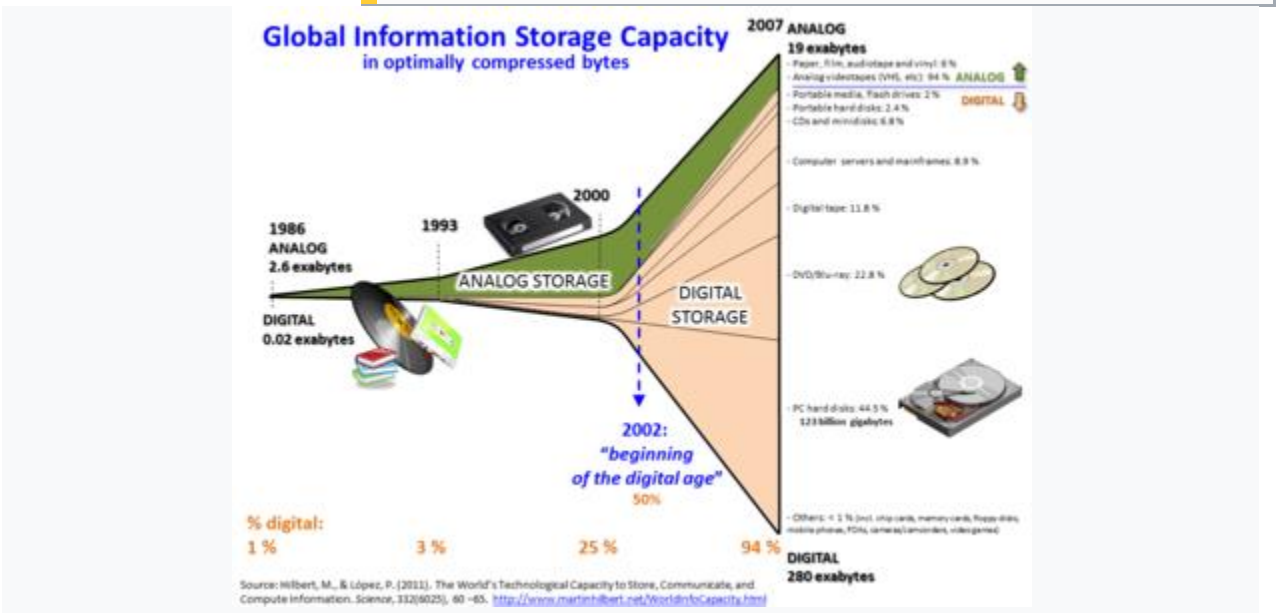
- **ZooKeeper:** ZooKeeper is a software project of the Apache Software Foundation. It is a service that provides centralized configuration and open code name registration for large distributed systems. ZooKeeper is a sub-project of Hadoop.

Big data

From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see Big Data (band). For buying and selling of personal and consumer data, see Surveillance capitalism.

This article **may contain an excessive number of citations**. Please consider removing references to unnecessary or disreputable sources, merging citations where possible, or, if necessary, flagging the content for deletion. *In particular many references are "spammed" here for promotional purposes. These need to be removed.. (November 2019) (Learn how and when to remove this template message)*



Growth of and digitization of global information-storage capacity[1]

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.[2] Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*.[3] When we handle big data, we may not sample but simply observe and track what happens. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time[4] and *value.* [5]

Current usage of the term *big data* tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."[6] Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."[7] Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet searches, fintech, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics,[8] connectomics, complex physics simulations, biology and environmental research.[9]

Data sets grow rapidly, in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[10][11] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[12] as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data are generated.[13] Based on an IDC report prediction, the global data volume will grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of

data.[14] One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[15]

Relational database management systems, desktop statistics[*clarification needed*] and software packages used to visualize data often have difficulty handling big data. The work may require "massively parallel software running on tens, hundreds, or even thousands of servers".[16] What qualifies as being "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."[17]

We Make It Easy For You?

Big Data Key Terms, Explained

Are you tired and exhausted of Big Data Terms? Don’t worry, Here’s the Big Data Terms used by experts with brief explanation.

1. Big Data

There are all sorts of popular and academic articles available defining Big Data, and the definitions vary considerably. In the interest of capturing the term's essence, we like this definition: "Data is big when data size becomes part of the problem." Big Data is a moving target, and this definition provides the flexibility required to capture its central characteristic.

Big Data is often characterized by the (originally) 3 Vs, which has grown to 4, 5, 6, or more, depending on where you look. I believe the following 6 Vs are enough to explain Big Data at a very high level.

Trying to freshly define the Vs of Big Data seems futile, and so we turn to the ever-authoritative Wikipedia for definitions:

2. **Big Data Volume**

Volume refers to the quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can actually be considered big data or not.

3. **Big Data Velocity**

Velocity is the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

4. **Big Data Variety**

   Variety refers to the type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

5. **Big Data Veracity**

   Veracity is the quality of captured data, which can vary greatly, affecting accurate analysis.

6. **Big Data Variability**

   Variability is the inconsistency of the data set, which can hamper processes to handle and manage it.

7. **Big Data Value**

   Value is tossed around as an important Big Data V from time to time, and I agree with its consideration, especially from a business standpoint. Value refers to what insight can be leveraged from the patterns, processing, and other Big Data-related tasks on the data in question.

8. Cloud Computing

Cloud computing, or what is simply referred to as the cloud, can be defined as an Internet-based computing model that largely offers on-demand access to computing resources. These resources comprise of many things, such as application software, computing resources, servers and data centers etc. Cloud service providers usually adopt a ‘pay-as-you-go’ model, something that allows companies to scale their costs as per need. It allows businesses to bypass infrastructural setup costs, which was inevitable prior to the advent of the cloud.

(From Kaushik Pal's Cloud Computing Key Terms, Explained)

9. Predictive Analytics

Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.

(From Eric Siegel's Predictive Analytics Introductory Key Terms, Explained)

Predictive analytics employ **predictive models**:

A mechanism that predicts a behavior of an individual, such as click, buy, lie, or die. It takes characteristics (variables) of the individual as input and provides a predictive score as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior.

(From Eric Siegel's Predictive Analytics Introductory Key Terms, Explained)

10. Descriptive Analytics

This form of analytics is descriptive in nature, as its name clearly hints at. Descriptive analytics summarizes data, focusing less on the precise details of every piece of data, and instead focusing on an overall narrative.

11. Prescriptive Analytics

Prescriptive analytics generally follows prediction, in that actions can be prescribed based on what has been gleaned from predictive modeling.

12. Database

Data needs to be curated, coddled, and cared for. It needs to be stored and processed, so that it may be transformed into information, and further refined into knowledge. The mechanism for storing data, subsequently facilitating these transformations, is, clearly, the database.

13. Data Warehouse

Data warehouse is another potentially elusive term. Han, Kamber & Pei define a data warehouse as data storage architectures which allow for "business executives to systematically organize, understand, and use their data to make strategic decisions." Vague, to be sure, but generally speaking, a data warehouse exhibits these characteristics:

- maintained separately from an organization's operational and transactional databases, which are characterized by frequent access and are used by day-to-day organizational operations
- allow for the integration of various disparate application systems
- house and allow access to consolidated historical data for processing and analysis

Bill Inmon, the Godfather of the Data Warehouse, gave this original and lasting definition, with which we will conclude:

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

14. ETL

ETL stands for **Extract, Transform and Load**. This is the process of extracting data from source systems, such as transactional databases, and placing it into data warehouses. If you are familiar with **online transactional processing (OLTP)** and **online analytical processing (OLAP)**, ETL can be thought of as the bridge between these 2 system types.

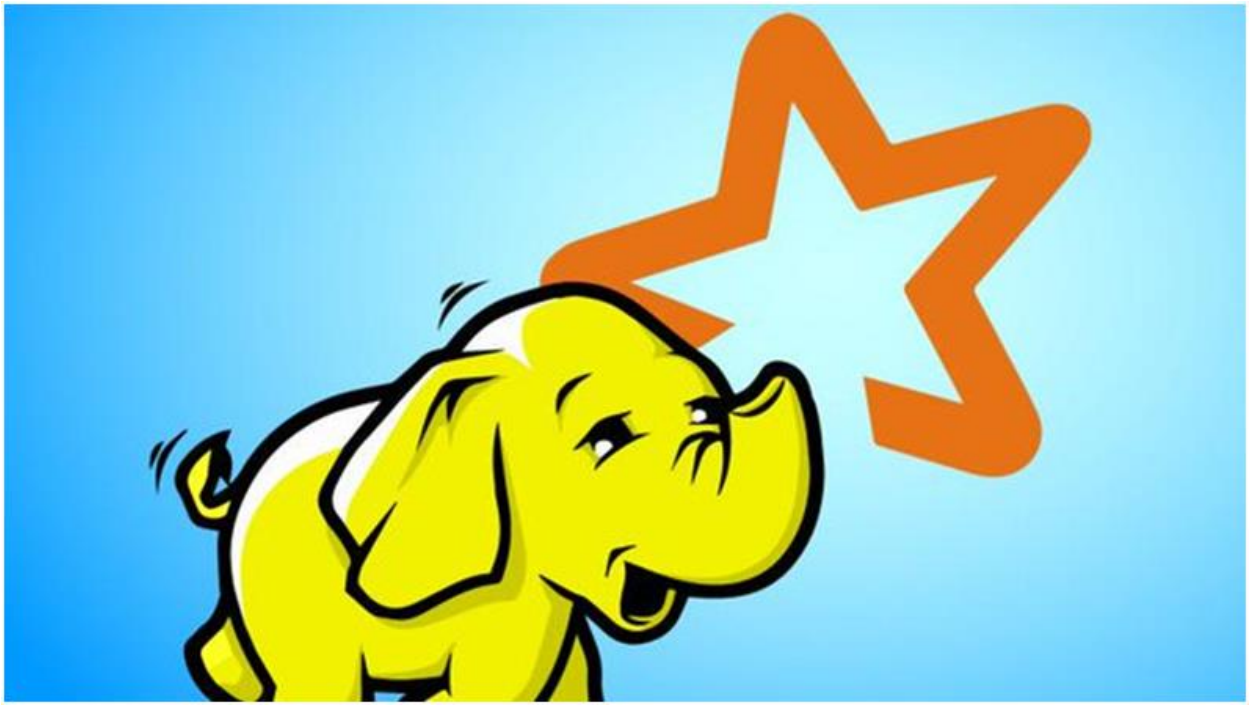
15. Business Intelligence

And perhaps the most ambiguous term of all (an incredible feat in a set of Big Data terminology definitions) is business intelligence (BI). BI is an unstable, ill-defined set of tools, technologies, and concepts which support business by providing historical, current, and predictive views on its operations. The relationship between BI and data mining, in particular, is a curious one, with various definitions proposing that: BI is a subset of data mining; data mining is a subset of BI, BI is driven by data mining; BI and data mining are separate and mutually exclusive. So, that settles that.

In the age of data science and Big Data, BI is generally thought to include OLAP, competitive intelligence, benchmarking, reporting, and other business management approaches (all of which tend toward ambiguity in definition as well), and is heavily influenced by the dashboard culture.

16. Apache Hadoop

Apache's Hadoop could almost single-handedly be responsible for the rise of the Big Data Revolution, at least from a software point of view.



Apache Hadoop is an open-source framework for processing large volume of data in a clustered environment. It uses simple MapReduce programming model for reliable, scalable and distributed computing. The storage and computation both are distributed in this framework.

(From Kaushik Pal's Hadoop Key Terms, Explained)

17. Apache Spark

Apache Spark is a powerful open-source processing engine built around speed, ease of use, and sophisticated analytics, with APIs in Java, Scala, Python, R, and SQL. Spark runs programs up to

100x faster than Apache Hadoop MapReduce in memory, or 10x faster on disk. It can be used to build data applications as a library, or to perform ad-hoc data analysis interactively. Spark powers a stack of libraries including SQL, DataFrames, and Datasets, MLlib for machine learning, GraphX for graph processing, and Spark Streaming. You can combine these libraries seamlessly in the same application. As well, Spark runs on a laptop, Apache Hadoop, Apache Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Apache Cassandra, Apache HBase, and S3.

(From Denny Lee and Jules Damji's Apache Spark Key Term's, Explained)

18. Internet of Things

The Internet of Things (IoT) is a growing source of Big Data moving forward. IoT is:

The concept to allow internet based communications to happen between physical objects, sensors, and controllers.

(From Geethika Bhavya Peddibhotla's Internet of Things Key Terms, Explained)

19. Machine Learning

Machine learning can be employed for predictive analysis and pattern recognition in Big Data. According to Mitchell, machine learning is "concerned with the question of how to construct computer programs that automatically improve with experience." Machine learning is interdisciplinary in nature, and

employs techniques from the fields of computer science, statistics, and artificial intelligence, among others. The main artefacts of machine learning research are algorithms which facilitate this automatic improvement from experience, algorithms which can be applied in such diverse fields as computer vision, artificial intelligence, and data mining.

20. Data Mining

Fayyad, Piatetsky-Shapiro & Smyth define data mining as "the application of specific algorithms for extracting patterns from data." This demonstrates that, in data mining, the emphasis is on the application of algorithms, as opposed to on the algorithms themselves. We can define the relationship between machine learning and data mining as follows: data mining is a process, during which machine learning algorithms are utilized as tools to extract potentially-valuable patterns held within datasets.

Big Data Terms Every Expert Should Know?

With billions of bytes of data being collected daily, it's more important than ever to understand the intricacies of big data. In an effort to help bring clarity to this field, we created a compiled list from our recent big data guides of what we feel are the most important related terms and definitions you need to know. Any terms you think we should add? Let us know in the comments!

A

Algorithm: A set of rules given to an AI, neural network, or other machines to help it learn on its own; classification, clustering, recommendation, and regression are four of the most popular types.

Apache Flink: An open-source streaming data processing framework. It is written in Java and Scala and is used as a distributed streaming dataflow engine.

Apache Hadoop: An open-source tool to process and store large distributed data sets across machines by using MapReduce.

Apache Kafka: A distributed streaming platform that improves upon traditional message brokers through improved throughput, built-in partitioning, replication, latency, and reliability.

Apache NiFi: An open-source Java server that enables the automation of data flows between systems in an extensible, pluggable, open manner. NiFi was open-sourced by the NSA.

Apache Spark: An open-source big data processing engine that runs on top of Apache Hadoop, Mesos, or the cloud.

Artificial intelligence: A machine's ability to make decisions and perform tasks that simulate human intelligence and behavior.

B

Big data: A common term for large amounts of data. To be qualified as big data, data must be coming into the system at a high velocity, with large variation, or at high volumes.

Blob storage: An Azure service that stores unstructured data in the cloud as a blob or an object.

Business intelligence: The process of visualizing and analyzing business data for the purpose of making actionable and informed decisions.

C

Cluster: A subset of data that share particular characteristics. Can also refer to several machines that work together to solve a single problem.

COAP: Constrained Application Protocol is an Internet Application protocol for limited resource devices that can be translated to HTTP if needed.

D

Data engineering: The collection, storage, and processing of data so that it can be queried by a data scientist.

Data flow management: The specialized process of ingesting raw device data, while managing the flow of thousands of producers and consumers. Then performing basic data enrichment, analysis in stream, aggregation, splitting, schema translation, format conversion, and other initial steps to prepare the data for further business processing.

Data governance: The process of managing the availability, usability, integrity, and security of data within a data lake.

Data integration: The process of combining data from different sources and providing a unified view for the user.

Data lake: A storage repository that holds raw data in its native format.

Data mining: A practice to generate new information through the process of examining and analyzing large databases.

Data operationalization: The process of strictly defining variables into measurable factors.

Data preparation: The process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis.

Data processing: The process of retrieving, transforming, analyzing, or classifying information by a machine.

Data science: A field that explores repeatable processes and methods to derive insights from data.

Data swamp: What a data lake becomes without proper governance.

Data validation: The act of examining data sets to ensure that all data is clean, correct, and useful before it is processed.

Data warehouse: A large collection of data from various sources used to help companies make informed decisions.

Device layer: The entire range of sensors, actuators, smartphones, gateways, and industrial equipment that send data streams corresponding to their environment and performance characteristics.

G

GPU-accelerated databases: Databases which are required to ingest streaming data.

Graph analytics: A way to organize and visualize relationships between different data points in a set.

H

Hadoop: A programming framework for processing and storing big data, particularly in distributed computing environments.

I

Ingestion: The intake of streaming data from any number of different sources.

M

MapReduce: A data processing model that filters and sorts data in the Map stage, then performs a function on that data and returns an output in the Reduce stage.

Munging: The process of manually converting or mapping data from one raw form into another format for more convenient consumption.

N

Normal distribution: A common graph representing the probability of a large number of random variables, where those variables approach normalcy as the data set increases in size. Also called a Gaussian distribution or bell curve.

Normalizing: The process of organizing data into tables so that the results of using the database are always unambiguous and as intended.

P

Parse: To divide data, such as a string, into smaller parts for analysis.

Persistent storage: A non-changing place, such as a disk, where data is saved after the process that created it has ended.

Python: A general-purpose programming language that emphasizes code readability in order to allow programmers to use fewer lines of code to express their concepts.

R

R: An open-source language primarily used for data visualization and predictive analytics.

Real-time stream processing: A model for analyzing sequences of data by using machines in parallel, though with reduced functionality.

Relational database management system (RDBMS): A system that manages, captures, and analyzes data that is grouped based on shared attributes called relations.

Resilient distributed dataset: The primary way that Apache Spark abstracts data, where data is stored across multiple machines in a fault-tolerant way.

S

Shard: An individual partition of a database.

Smart data: Digital information that is formatted so it can be acted upon at the collection point before being sent to a downstream analytics platform for further data consolidation and analytics.

Stream processing: The real-time processing of data. The data is processed continuously, concurrently, and record-by-record.

Structured data: Information with a high degree of organization.

T

Taxonomy: The classification of data according to a pre-determined system with the resulting catalog used to provide a conceptual framework for easy access and retrieval.

Telemetry: The remote acquisition of information about an object (for example, from an automobile, smartphone, medical device, or IoT device).

Transformation: The conversion of data from one format to another.

U

Unstructured data: Data that either does not have a pre-defined data model or is not organized in a pre-defined manner.

V

Visualization: The process of analyzing data and expressing it in a readable, graphical format, such as a chart or graph.

Z

Zones: Distinct areas within a data lake that serve specific, well-defined purposes.

Famous Big Data Analytics Glossary Only For Experts?

ACID Test: A test applied to data for atomicity, consistency, isolation and durability.

Aggregation: A process of searching, gathering and presenting data.

Algorithm: A mathematical formula or statistical process used to perform analysis of data.

Alpha Risk: The maximum probability of making a Type I error. This probability is established by the experimenter and often set at 5%.

Alternative Hypothesis (Ha): Statement of a change or difference; assumed to be true if the null hypothesis is rejected.

Anomaly Detection: The process of identifying rare or unexpected items or events in a dataset that do not conform to other items in the dataset and do not match a projected pattern or expected behavior. Anomalies are also called outliers,

exceptions, surprises or contaminants and they often provide critical and actionable information.

Anonymization: Making data anonymous; severing of links between people in a database and their records to prevent the discovery of the source of the records.

ANOVA: One-way ANOVA is a generalization of the 2-sample t-test, used to compare the means of more than two samples to each other.

ANOVA Table: The ANOVA table is the standard method of organizing the many calculations necessary for conducting an analysis of variance.

API (Application Program Interface): A set of programming standards and instructions for accessing or building web-based software applications.

Application: Software that enables a computer to perform a certain task.

Artificial Intelligence: The apparent ability of a machine to apply information gained from previous experience accurately to new situations in a way that a human would.

Batch Processing: Batch data processing is an efficient way of processing high volumes of data where a group of transactions is collected over a period of time. Hadoop is focused on batch data processing.

Bayes Theorem: A theorem based on conditional probabilities. It uses relevant evidence, also known as conditional probability, to determine the probability of an event, based on prior knowledge of conditions that might be related to the event.

Beta Risk: The risk or probability of making a Type II error.

Big Data: Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.

Business Intelligence: The general term used for the identification, extraction and analysis of data.

Cassandra: A popular open source database management system managed by The Apache Software Foundation. Designed to handle large amounts of distributed data across commodity servers while providing a highly available service. It is a NoSQL solution that was initially developed by Facebook.

Classification Analysis: A systematic process for obtaining important and relevant information about data (metadata) and assigning data to a particular group or class.

Clickstream Analytics: The analysis of users' web activity through the items they click on a page.

Cloud: A broad term that refers to any internet-based application or service that is hosted remotely.

Cloud Computing: A distributed computing system hosted and running on remote servers and accessible from anywhere on the internet.

Cluster Computing: Computing using a ‘cluster’ of pooled resources of multiple servers. Getting more technical, we might be talking about nodes, cluster management layer, load balancing and parallel processing etc.

Clustering Analysis: The process of identifying objects that are similar to each other and clustering them in order to understand the differences as well as the similarities within the data.

Coefficient of Variation: Standard deviation normalized by the mean: ?/?.

Columnar Database or Column-oriented Database: A database that stores data by column rather than by row. In a row-based database, a row might contain a name, address and phone number. In a column-oriented database, all names are in one column, addresses in another and so on. A key advantage of a columnar database is faster hard disk access.

Comparative Analysis: Data analysis that compares two or more data sets or processes to detect patterns within very large data sets.

Confidence Interval: A range of values which is likely to contain the population parameter of interest with a given level of confidence.

Continuous Data: Data from a measurement scale that can be divided into finer and finer increments (e.g. temperature, time, pressure). Also known as variable data.

Correlation Analysis: A means to determine a statistical relationship between variables, often for the purpose of identifying predictive factors among the variables. A technique for quantifying the strength of the linear relationship between two variables.

Dashboard: A graphical representation of analyses performed by algorithms.

Dark Data: All the data that is gathered and processed by enterprises but not used for any meaningful purposes.

Data: A quantitative or qualitative value. Common types of data include sales figures, marketing research results, readings from monitoring equipment, user actions on a website, market growth projections, demographic information and customer lists.

Data Aggregation: The process of collecting data from multiple sources for the purpose of reporting or analysis.

Data Analyst: A person responsible for the tasks of modelling, preparing and cleaning data for the purpose of deriving actionable information from it.

Data Analytics: The process of examining large data sets to uncover hidden patterns, unknown correlations, trends, customer preferences and other useful business insights. The end result might be a report, an indication of status or an action taken automatically based on the information received. Businesses typically use the following types of analytics:

- *Behavioral Analytics*: Using data about people's behavior to understand intent and predict future actions.
- *Descriptive Analytics*: Condensing big numbers into smaller pieces of information. This is similar to summarizing the data story. Rather than listing every single number and detail, there is a general thrust and narrative.
- *Diagnostic Analytics*: Reviewing past performance to determine what happened and why. Businesses use this type of analytics to complete root cause analysis.
- *Predictive Analytics*: Using statistical functions on one or more data sets to predict trends or future events. In big data predictive analytics, data scientists may use advanced

techniques like data mining, machine learning and advanced statistical processes to study recent and historical data to make predictions about the future. It can be used to forecast weather, predict what people are likely to buy, visit, do or how they may behave in the near future.

- *Prescriptive Analytics*: Prescriptive analytics builds on predictive analytics by including actions and make data-driven decisions by looking at the impacts of various actions.

Data Architecture and Design: How enterprise data is structured. The actual structure or design varies depending on the eventual end result required. Data architecture has three stages or processes: (1) conceptual representation of business entities, (2) the logical representation of the relationships among those entities and (3) the physical construction of the system to support the functionality.

Data as a Service (DaaS): Treat data as a product. DaaS providers use cloud solutions to give on-demand access of data to customers.

Data Center: A physical facility that houses a large number of servers and data storage devices. Data centers might belong to a single organization or sell their services to many organizations.

Data Cleansing: The process of reviewing and revising data to delete duplicate entries, correct misspelling and other errors, add missing data and provide consistency.

Data Ethical Guidelines: Guidelines that help organizations be transparent with the data, ensuring simplicity, security and privacy.

Data Feed: A means for a person to receive a stream of data such as a Twitter feed or RSS.

Data Governance: A set of processes or rules that ensure data integrity and that data management best practices are met.

Data Integration: The process of combining data from different sources and presenting it in a single view.

Data Integrity: The measure of trust an organization has in the accuracy, completeness, timeliness and validity of the data.

Data Lake: A large repository of enterprise-wide data in raw format. Supposedly data lakes make it easy to access enterprise-wide data. However, you really need to know what you are looking for and how to process it and make intelligent use of it.

Data Mart: The access layer of a data warehouse used to provide data to users.

Data Warehouse: A repository for enterprise-wide data but in a structured format after cleaning and integrating with other sources. Data warehouses are typically used for conventional data (but not exclusively).

Data Mining: Finding meaningful patterns and deriving insights in large sets of data using sophisticated pattern recognition techniques. To derive meaningful patterns, data miners use statistics, machine learning algorithms and artificial intelligence.

Data Modelling: A data model defines the structure of the data for the purpose of communicating between functional and technical people to show data needed for business processes, or for communicating a plan to develop how data is stored and accessed among application development team members.

Data Scientist: Someone who can make sense of big data by extracting raw data, massaging it and come up with insights. Skills needed are statistics, computer science, creativity, story-telling and understanding of business context.

Data Science: A discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning and database engineering to solve complex problems.

Data Set: A collection of data, very often in tabular form.

Database: A digital collection of data and the structure around which the data is organized. The data is typically entered into and accessed via a database management system.

Database Management System (DBMS): Software that collects and provides access to data in a structured format.

Demographic Data: Data relating to the characteristics of a human population.

Discrete Data: Data which is not measured on a continuous scale. Examples are binomial (pass/fail), Counts per unit, Ordinal (small/medium/large) and Nominal (red/green/blue). Also known as attribute or categorical data.

Discriminant Analysis: A statistical analysis technique used to predict cluster membership from labelled data.

Distributed File System: A data storage system meant to store large volumes of data across multiple storage devices and will help decrease the cost and complexity of storing large amounts of data.

Empirical Model: An equation derived from the data that expresses a relationship between the inputs and an output (Y=f(x)).

ETL (Extract, Transform and Load): The process of extracting raw data, transforming by cleaning/enriching the data to make it fit operational needs and loading into the appropriate repository for the system's use. Even though it originated with data warehouses, ETL processes are used while taking/absorbing data from external sources in big data systems.

Event: A set of outcomes of an experiment (a subset of the sample space) to which a probability is assigned.

Exploratory Analysis: An approach to data analysis focused on identifying general patterns in data, including outliers and features of the data that are not anticipated by the experimenter's current knowledge or preconceptions. EDA aims to uncover underlying structure, test assumptions, detect mistakes and understand relationships between variables.

External Data: Data that exists outside of a system.

F-test: A hypothesis test for comparing variances.

Fit: The average outcome predicted by a model.

Grid Computing: Connecting different computer systems from various locations, often via the cloud, to reach a common goal.

Hadoop: An open source software framework administered by Apache that allows for storage, retrieval and analysis of very large data sets across clusters of computers.

High Performance Computing: Using supercomputers to solve highly complex and advanced computing problems.

Histograms: Representation of frequency of values by intervals.

In-database Analytics: The integration of data analytics into the data storage layer.

In-memory Computing: A technique of moving the working datasets entirely within a cluster's collective memory and avoid writing intermediate calculations to disk. This results in very fast processing, storing and loading of data.

IoT (Internet of Things): The network of physical objects or "things" embedded with electronics, software, sensors and connectivity to enable it to achieve greater value and service by exchanging data with the manufacturer, operator and/or other connected devices. Each thing is uniquely identifiable through its embedded computing system but is able to interoperate within the existing Internet infrastructure.

Juridical Data Compliance: Use of data stored in a country must follow the laws of that country. Relevant when using cloud solutions with data stored in difference countries or continents.

Key Value Databases: Storing data with a primary key, a uniquely identifiable record, which makes it easy and computationally efficient to look up.

Latency: Any delay in a response or delivery of data from one point to another.

Load Balancing: The process of distributing workload across a computer network or computer cluster to optimize performance.

Location Analytics: Location analytics brings mapping and map-driven analytics to enterprise business systems and data warehouses. It allows you to associate geospatial information with datasets.

Location Data: GPS data describing a geographical location.

Log File: A file that a computer, network or application creates automatically to record events that occur during operation. For example, the time a file is accessed.

Logistic Regression: Investigates the relationship between response (Y's) and one or more predictors (X's) where Y's are categorical, not continuous and X's can be either continuous or categorical. Types of logistic regression are:

- *Binary Logistic Regression*: Y variable takes on one of two outcomes (levels), e.g. pass/fail, agree/disagree.
- *Ordinal Logistic Regression*: Y variable can have more than two levels. Levels are rank ordered, e.g. Low/Medium/High, 1-5 preference scale.
- *Nominal Logistic Regression*: Y variable can have more than two levels. There is no implied order to the levels, e.g. Blue/Yellow/Green, Company A/B/C/D.

Machine-generated Data: Data automatically created by machines via sensors or algorithms or any other non-human source.

Machine Learning: A method of designing systems that can learn, adjust and improve based on the data fed to them. Using predictive and statistical algorithms that are fed to these machines, they learn and continually zero in on "correct"

behavior and insights and they keep improving as more data flows through the system.

MapReduce: A programming model for processing and generating large data sets. This model does two distinct things. First, the “Map” includes turning one dataset into another, more useful and broken down dataset made of parts called tuples. Tuples may typically be processed independently from each other across multiple processors. Second, “Reduce” takes all of the broken down, processed tuples and combines their output into a usable result. The result is a practical breakdown of processing.

Massively Parallel Processing (MPP): Using many different processors (or computers) to perform certain computational tasks at the same time.

Mean: The weighted average of data. The population mean is denoted by ? (Greek letter mu) and the sample mean is denoted by x?.

Median: The middle value of a data set when arranged in order of magnitude.

Metadata: Data about data; it gives information about what the data is about. For example, where data points were collected.

Mode: The measurement that occurs most often in a data set.

Multi-dimensional Databases: A database optimized for data online analytical processing (OLAP) applications and for data warehousing.

Naïve Bayes: A classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, it assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Network Analysis: Analyzing connections between nodes in a network and the strength of their ties.

Neural Network: Models inspired by the real-life biology of the brain. These are used to estimate mathematical functions and facilitate different kinds of learning algorithms. Deep Learning is a similar term and is generally seen as a modern buzzword, rebranding the Neural Network paradigm for the modern day.

Normal Distribution: The most important continuous probability distribution in statistics is the normal distribution (a.k.a. Gaussian distribution). The normal distribution is the familiar bell curve. Once m and s are specified, the entire curve is determined.

NoSQL (Not ONLY SQL): A broad class of database management systems identified by non-adherence to the widely used relational database management system model. NoSQL databases are not built primarily on tables and generally do not use SQL for data manipulation. Database management systems that are designed to handle large volumes of data and are often well-suited for big data systems because of their flexibility and distributed-first architecture needed for large unstructured databases.

Null Hypothesis (H0): Statement of no change or difference; assumed to be true until sufficient evidence is presented to reject it.

One Sample T-test: Statistical test to compare the mean of one sample of data to a target. Uses the t-distribution

Operational Databases: Databases that carry out regular operations of an organization that are generally very important to the business. They typically use online transaction processing that allows them to enter, collect and retrieve specific information about the organization.

Optimization Analysis: The process of finding optimal problem parameters subject to constraints. Optimization algorithms heuristically test a large number of parameter configurations in order to find an optimal result, determined by a characteristic function (also called a fitness function).

Outlier Detection: An object that deviates significantly from the general average within a dataset or a combination of data. It is numerically distant from the rest of the data and therefore indicates that something unusual and generally requires additional analysis.

Paired T-test: A test used to compare the average difference between two samples of data that are linked in pairs. Special case of the 1-sample t-test. Uses the t-distribution.

Pattern Recognition: Identifying patterns in data via algorithms to make predictions about new data coming from the same source.

Pig: A data flow language and execution framework for parallel computation.

Population: A dataset that consists of all the members of some group. Descriptive parameters (such as ?, ?) are used to describe the population.

Power (1-beta): The ability of a statistical test to detect a real difference when there is one; the probability of correctly rejecting the null hypothesis. Determined by alpha and sample size.

Predictive Modelling: The process of developing a model that will most likely predict a trend or outcome.

Probability: The likelihood of a given event's occurrence, which is expressed as a number between 1 and 0.

Probability Distribution: A statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. Probability distributions may be discrete or continuous.

Public Data: Public information or data sets that were created with public funding.

Query: Asking for information to answer a certain question.

R: An open source programming language used for statistical computing and graphics. It is a GNU project which is similar to the S language. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques and is highly extensible. It is one of the most popular languages in data science.

Range: Difference between the largest and smallest measurement in a data set.

Real Time Data: Data that is created, processed, stored, analyzed and visualized within milliseconds.

Regression Analysis: A modelling technique used to define the association between variables. It assumes a one-way causal effect from predictor variables (independent variables) to a response of another variable (dependent variable). Regression can be used to explain the past and predict future events.

Residual: The difference between reality (an actual measurement) and the fit (model output).

Sample: A data set which consists of only a portion of the members from some population. Sample statistics are used to draw inferences about the entire population from the measurements of a sample.

Scalability: The ability of a system or process to maintain acceptable performance levels as workload or scope increases.

Semi-structured Data: Data that is not structured by a formal data model, but provides other means of describing the data hierarchies (tags or other markers).

Sentiment Analysis: The application of statistical functions and probability theory to comments people make on the web or social networks to determine how they feel about a product, service or company.

Significant Difference: The term used to describe the results of a statistical hypothesis test where a difference is too large to be reasonably attributed to chance.

Single-variance Test (Chi-square Test): Compares the variance of one sample of data to a target. Uses the Chi-square distribution.

Software as a Service (SaaS): Enables vendors to host an application and make it available via the internet (cloud servicing). SaaS providers provide services over the cloud rather than hard copies.

Spark (Apache Spark): A fast, in-memory open source data processing engine to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. Spark is generally a lot faster than MapReduce.

Spatial Analysis: Analyzing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in a geographic space.

SQL (Structured Query Language): A programming language for retrieving data from a relational database.

Standard Deviation: The positive square root of the variance:

- Population: ?
- Sample: s

Stream Processing: Stream processing is designed to act on real-time and streaming data with “continuous” queries. Combined with streaming analytics (i.e. the ability to continuously calculate mathematical or statistical analytics on the fly within the stream), stream processing solutions are designed to handle high volumes in real time.

Structured Data: Data that is organized according to a predetermined structure.

Sum of Squares: In ANOVA, the total sum of squares helps express the total variation that can be attributed to various factors. From the ANOVA table, %SS is the Sum of Squares of the Factor divided by the Sum of Squares Total. Similar to R2 in Regression.

Terabyte: 1024 gigabytes. A terabyte can store approximately 300 hours of high-definition video.

Test for Equal Variance (F-test): Compares the variance of two samples of data against each other. Uses the F distribution.

Test Statistic: A standardized value (Z, t, F, etc.) which represents the likelihood of H0 and is distributed in a known manner such that the probability for this value can be determined.

Text Analytics: The application of statistical, linguistic and machine learning techniques on text-based data-sources to derive meaning or insight.

Time Series Analysis: Analysis of well-defined data measured at repeated measures of time to identify time based patterns.

Topological Data Analysis: Analysis techniques focusing on the theoretical shape of complex data with the intent of identifying clusters and other statistically significance trends that may be present.

Transactional Data: Data that relates to the conducting of business, such as accounts payable and receivable data or product shipments data.

Two Sample t-test: A statistical test to compare the means of two samples of data against each other. Uses the t-distribution.

Type I Error: The error that occurs when the null hypothesis is rejected when, in fact, it is true.

Type II Error: The error that occurs when the null hypothesis is not rejected when it is, in fact, false.

Unstructured Data: Data that has no identifiable structure, such as email message text, social media posts, audio files (recorded human speech, music), etc.

Variance: The average squared deviation for all values from the mean:

- Population: 2
- Sample: s2

Variety: The different types of data available to collect and analyze in addition to the structured data found in a typical database. Categories include machine generated data, computer log data, textual social media information, multimedia social and other information.

Velocity: The speed at which data is acquired and used. Not only are companies and organizations collecting more and more data at a faster rate, they want to derive meaning from that data as soon as possible, often in real time.

Veracity: Ensuring that data used in analytics is correct and precise.

Visualization: A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively. Visuals created are usually complex, but understandable in order to convey the message of data.

References

1. Augur, H. (2016, May). A Beginner's Guide to Big Data Terminology. Retrieved from http://dataconomy.com.
2. Dontha, R. (2017, January). 25 Big Data Terms You Must Know to Impress Your Date (or whomever you want). Retrieved from http://www.datasciencecentral.com.
3. NT, B. (2014, July). Big Data A to Z: A glossary of Big Data terminology. Retrieved from http://bigdata-madesimple.com.
4. Analytics and Big Data Glossary for the Enterprise. (2017, March). Retrieved from http://data-informed.com.
5. An Extensive Glossary of Big Data Terminology. Retrieved from https://datafloq.com.

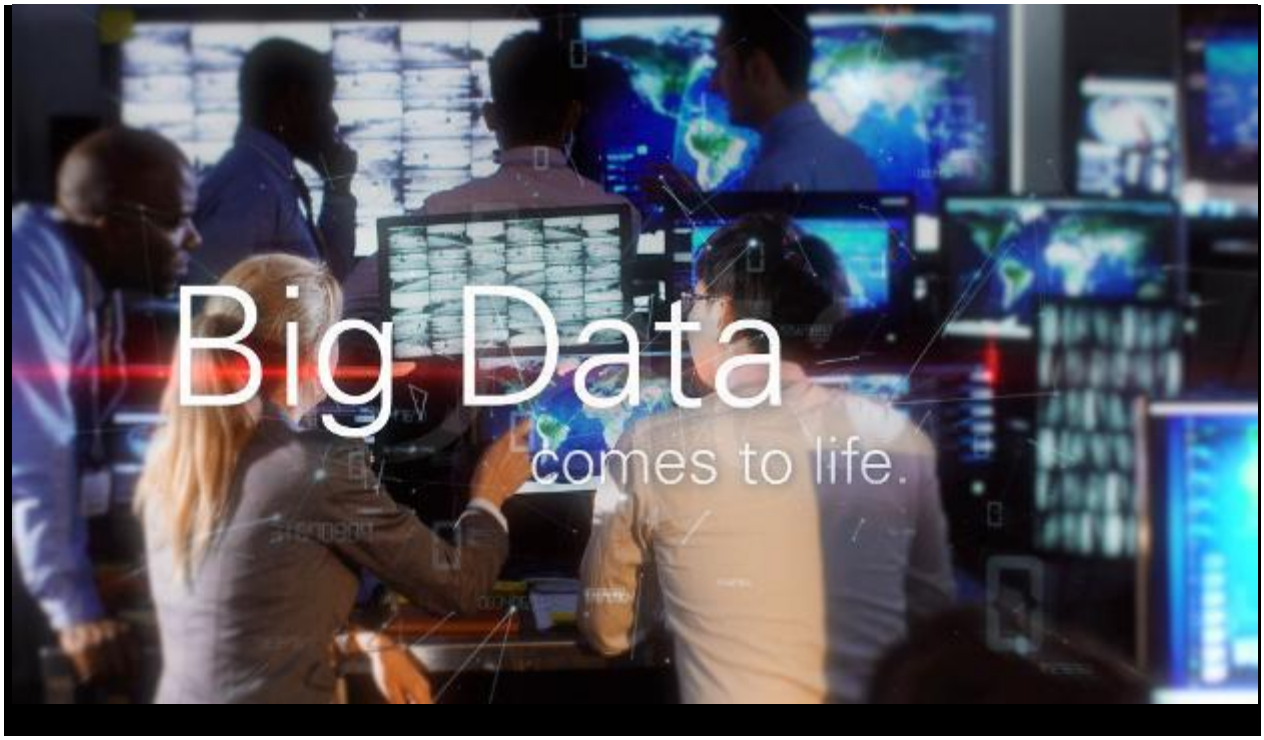
Again What Is Big Data?

The Definition of Big Data

What exactly is big data?

To really understand big data, it's helpful to have some historical background. Here is Gartner's definition, circa 2001 (which is still the go-to definition): Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity. This is known as the three Vs.

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before.



Why Oracle Cloud for Big Data?

Oracle Cloud Platform for Big Data unlocks the value of data by quickly and seamlessly integrating new insights into every aspect of production, across all departments, available to all employees in languages they can understand.

The Three Vs of Big Data

Volume: The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a webpage or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

Velocity: Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.

Variety: Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

The Value—and Truth—of Big Data

Two more Vs have emerged over the past few years: value and veracity.

Data has intrinsic value. But it's of no use until that value is discovered. Equally important: How truthful is your data—and how much can you rely on it?

Today, big data has become capital. Think of some of the world's biggest tech companies. A large part of the value they offer comes from their data, which they're constantly analyzing to produce more efficiency and develop new products.

Recent technological breakthroughs have exponentially reduced the cost of data storage and compute, making it easier and less expensive to store more data than ever before. With an increased volume of big data now cheaper and more accessible, you can make more accurate and precise business decisions.

Finding value in big data isn't only about analyzing it (which is a whole other benefit). It's an entire discovery process that requires insightful analysts, business users, and executives who ask the right questions, recognize patterns, make informed assumptions, and predict behavior.

But how did we get here?

Going Big Data? You need a cloud strategy.

The History of Big Data

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers

truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

Benefits of Big Data and Data Analytics:

Big data makes it possible for you to gain more complete answers because you have more information.

More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Big Data Use Cases

Big data can help you address a range of business activities, from customer experience to analytics. Here are just a few. (More use cases can be found at Oracle Big Data Solutions.)

Product Development: Companies like Netflix and Procter & Gamble use big data to anticipate customer demand. They build predictive models for new products and services by classifying key attributes of past and current products or services and modeling the relationship between those attributes and the commercial success of the offerings. In addition, P&G uses data and analytics from focus groups, social media, test markets, and early store rollouts to plan, produce, and launch new products.

Predictive Maintenance: Factors that can predict mechanical failures may be deeply buried in structured data, such as the year, make, and model of equipment, as well as in unstructured data that covers millions of log entries, sensor data, error messages, and engine temperature. By analyzing these indications of potential issues before the problems happen, organizations can deploy maintenance more cost effectively and maximize parts and equipment uptime.

Customer Experience: The race for customers is on. A clearer view of customer experience is more possible now than ever before. Big data enables you to gather data from social media, web visits, call logs, and other sources to improve the interaction experience and maximize the value delivered. Start delivering personalized offers, reduce customer churn, and handle issues proactively.

Fraud and Compliance: When it comes to security, it's not just a few rogue hackers—you're up against entire expert teams. Security landscapes and compliance requirements are constantly evolving. Big data helps you identify patterns in data that indicate fraud and aggregate large volumes of information to make regulatory reporting much faster.

Machine Learning: Machine learning is a hot topic right now. And data—specifically big data—is one of the reasons why. We are now able to teach machines instead of program them. The availability of big data to train machine learning models makes that possible.

Operational Efficiency: Operational efficiency may not always make the news, but it's an area in which big data is having the most impact. With big data, you can analyze and assess production, customer feedback and returns, and

other factors to reduce outages and anticipate future demands. Big data can also be used to improve decision-making in line with current market demand.

Drive Innovation: Big data can help you innovate by studying interdependencies among humans, institutions, entities, and process and then determining new ways to use those insights. Use data insights to improve decisions about financial and planning considerations. Examine trends and what customers want to deliver new products and services. Implement dynamic pricing. There are endless possibilities.

Big Data Challenges

While big data holds a lot of promise, it is not without its challenges.

First, big data is…big. Although new technologies have been developed for data storage, data volumes are doubling in size about every two years. Organizations still struggle to keep pace with their data and find ways to effectively store it.

But it's not enough to just store the data. Data must be used to be valuable and that depends on curation. Clean data, or data that's relevant to the client and organized in a way that enables meaningful analysis, requires a lot of work. Data scientists spend 50 to 80 percent of their time curating and preparing data before it can actually be used.

Finally, big data technology is changing at a rapid pace. A few years ago, Apache Hadoop was the popular technology used to handle big data. Then Apache Spark was introduced in 2014. Today, a combination of the two frameworks appears to be the best approach. Keeping up with big data technology is an ongoing challenge.

How Big Data Works?

Big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

1. Integrate

Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as ETL (extract, transform, and load) generally aren't up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.

During integration, you need to bring in the data, process it, and make sure it's formatted and available in a form that your business analysts can get started with.

2. Manage

Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis. Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed.

3. Analyze

Your investment in big data pays off when you analyze and act on your data. Get new clarity with a visual analysis of your varied data sets. Explore the data further to make new discoveries. Share your findings with others. Build data models with machine learning and artificial intelligence. Put your data to work.

Big Data Best Practices

To help you on your big data journey, we've put together some key best practices for you to keep in mind. Here are our guidelines for building a successful big data foundation.

Align Big Data with Specific Business Goals: More extensive data sets enable you to make new discoveries. To that end, it is important to base new investments in skills, organization, or infrastructure with a strong business-driven context to guarantee ongoing project investments and funding. To determine if you are on the right track, ask how big data supports and enables your top business and IT priorities. Examples include understanding how to filter web logs to understand ecommerce behavior, deriving sentiment from social media and customer support interactions, and understanding statistical correlation methods and their relevance for customer, product, manufacturing, and engineering data.

Ease Skills Shortage with Standards and Governance: One of the biggest obstacles to benefiting from your investment in big data is a skills shortage. You can mitigate this risk by ensuring that big data technologies,

considerations, and decisions are added to your IT governance program. Standardizing your approach will allow you to manage costs and leverage resources. Organizations implementing big data solutions and strategies should assess their skill requirements early and often and should proactively identify any potential skill gaps. These can be addressed by training/cross-training existing resources, hiring new resources, and leveraging consulting firms.

Optimize Knowledge Transfer with a Center of Excellence: Use a center of excellence approach to share knowledge, control oversight, and manage project communications. Whether big data is a new or expanding investment, the soft and hard costs can be shared across the enterprise. Leveraging this approach can help increase big data capabilities and overall information architecture maturity in a more structured and systematic way.

Top Payoff Is Aligning Unstructured with Structured Data: It is certainly valuable to analyze big data on its own. But you can bring even greater business insights by connecting and integrating low density big data with the structured data you are already using today.

Whether you are capturing customer, product, equipment, or environmental big data, the goal is to add more relevant data points to your core master and analytical summaries, leading to better conclusions. For example, there is a difference in distinguishing all customer sentiment from that of only your best customers. Which is why many see big data as an integral extension of their existing business intelligence capabilities, data warehousing platform, and information architecture.

Keep in mind that the big data analytical processes and models can be both human- and machine-based. Big data analytical capabilities include statistics, spatial analysis, semantics, interactive discovery, and visualization. Using analytical models, you can correlate different types and sources of data to make associations and meaningful discoveries.

Plan Your Discovery Lab for Performance: Discovering meaning in your data is not always straightforward. Sometimes we don't even know what we're looking for. That's expected. Management and IT needs to support this "lack of direction" or "lack of clear requirement."

At the same time, it's important for analysts and data scientists to work closely with the business to understand key business knowledge gaps and requirements. To accommodate the interactive exploration of data and the experimentation of statistical algorithms, you need high-performance work areas. Be sure that sandbox environments have the support they need—and are properly governed.

Align with the Cloud Operating Model: Big data processes and users require access to a broad array of resources for both iterative experimentation and running production jobs. A big data solution includes all data realms including transactions, master data, reference data, and summarized data. Analytical sandboxes should be created on demand. Resource management is critical to ensure control of the entire data flow including pre- and post-processing, integration, in-database summarization, and analytical modeling. A well-planned private and public cloud provisioning and security strategy plays an integral role in supporting these changing requirements.

Big Data

Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

Intelligent Decisions

Big Data has the potential to help companies improve operations and make faster, more intelligent decisions. The data is collected from a number of sources including emails, mobile devices, applications, databases, servers and other means. This data, when captured, formatted, manipulated, stored and then analyzed, can help a company to gain useful insight to increase revenues, get or retain customers and improve operations.

Is Big Data a Volume or a Technology?

While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes the tools and processes), that an organization requires to handle the large amounts of data and storage facilities. The term is believed to have originated with web search companies who needed to query very large distributed aggregations of loosely-structured data.

An Example

An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

Business Datasets

When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Big Data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets.

Glossary of Big Data Terms

Learning how to work with Big Data comes with a lot a new terminology (and jargon!)

A
Aggregation – a process of searching, gathering and presenting data
Algorithms – a mathematical formula that can perform certain analyses on data
Analytics – the discovery of insights in data
Anonymization – making data anonymous; removing all data points that could lead to identify a person

API - an application programming interface is a set of subroutine definitions, protocols, and tools for building application software
Application – computer software that enables a computer to perform a certain task
Artificial Intelligence – developing intelligence machines and software that are capable of perceiving the environment and take corresponding action when required and even learn from those actions.

B
Behavioral Analytics – analytics that informs about the how, why and what instead of just the who and when. It looks at humanized patterns in the da
Big Data - Big data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data-processing applications.
Biometrics - The use of technology to identify people by one or more of their physical traits.
Brontobyte - A brontobyte is a measure of memory or data storage

that is equal to 10 to the 27th power of bytes. There are approximately 1,024 yottabytes in a brontobyte. Approximately 1,024 brontobytes make up a geopbyte.

C

Classification analysis - a systematic process for obtaining important and relevant information about data, also meta data called; data about data.

Cloud computing – a distributed computing system over a network used for storing data off-premises

Clustering analysis – the process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.

Comparative analysis – it ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.

Complex structured data – data that are composed of two or more complex, complicated, and interrelated parts that cannot be easily interpreted by structured query languages and tools.

Computer generated data – data generated by computers such as log files

Concurrency – performing and executing multiple tasks and processes at the same time

Correlation analysis – the analysis of data to determine a relationship between variables and whether that relationship is negative (- 1.00) or positive (+1.00).

D

Dashboard – a graphical representation of the analyses performed by the algorithms

Data aggregation tools - the process of transforming scattered data from numerous sources into a single new one.

Data analyst – someone analyzing, modeling, cleaning or processing data

Database – a digital collection of data stored via a certain technique

Database Management System– collecting, storing and providing access of data
Data centre – a physical location that houses the servers for storing data
Data cleansing – the process of reviewing and revising data in order to delete duplicates, correct errors and provide consistency
Data custodian– someone who is responsible for the technical environment necessary for data storage
Data feed – a stream of data such as a Twitter feed or RSS
Data mining – the process of finding certain patterns or information from data sets
Data modeling – the analysis of data objects using data modeling techniques to create insights from the data
Data set – a collection of data
Data virtualization – a data integration process in order to gain more insights. Usually it involves databases, applications, file systems, websites, big data techniques, etc.)
De-identification – same as anonymization; ensuring a person cannot be identified through the data

E
Exploratory analysis – finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.
Exabytes – approximately 1000 petabytes or 1 billion gigabytes.
Extract, Transform and Load (ETL) – a process in a database and data warehousing meaning extracting the data from various sources, transforming it to fit operational needs and loading it into the database

F
Failover – switching automatically to a different server or node should one fail
Fault-tolerant design – a system designed to continue working even if certain parts fail **Feature -** a piece of measurable information about

something, for example features you might store about a set of people, are age, gender and income.

G

Graph Databases – they use graph structures (a finite set of ordered pairs or certain entities), with edges, properties and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbor element.

Grid computing – connecting different computer systems from various location, often via the cloud, to reach a common goal

H

Hadoop – an open-source framework that is built to enable the process and storage of big data across a distributed file system

HBase – an open source, non-relational, distributed database running in conjunction with Hadoop

HDFS – Hadoop Distributed File System; a distributed file system designed to run on commodity hardware

High-Performance-Computing (HPC) – using supercomputers to solve highly complex and advanced computing problems

Histogram - A graphical representation of the distribution of a set of numeric data, usually a vertical bar graph

I

In-memory – a database management system stores data on the main memory instead of the disk, resulting is very fast processing, storing and loading of the data

Internet of Things – ordinary devices that are connected to the internet at any time anywhere via sensors

J

JavaScript - a scripting language designed in the mid-1990s for embedding logic in web pages, but which later evolved into a more general-purpose development language.

Juridical data compliance – relevant when you use cloud solutions

and where the data is stored in a different country or continent. Be aware that data stored in a different country has to oblige to the law in that country.

K

KeyValue Databases – they store data with a primary key, a uniquely identifiable record, which makes easy and fast to look up. The data stored in a KeyValue is normally some kind of primitive of the programming language.

L

Latency – a measure of time delayed in a system
Load balancing – distributing workload across multiple computers or servers in order to achieve optimal results and utilization of the system
Location data – GPS data describing a geographical location
Log file – a file automatically created by a computer to record events that occur while operational

M

Machine data – data created by machines via sensors or algorithms
Machine learning – part of artificial intelligence where machines learn from what they are doing and become better over time
Metadata – data about data; gives information about what the data is about.
Multi-Dimensional Databases – a database optimized for data online analytical processing (OLAP) applications and for data warehousing.
MultiValue Databases– they are a type of NoSQL and multidimensional databases that understand 3 dimensional data directly. They are primarily giant strings that are perfect for manipulating HTML and XML strings directly

N

Natural Language Processing– a field of computer science involved with interactions between computers and human languages
Network analysis– viewing relationships among the nodes in terms of

the network or graph theory, meaning analyzing connections between nodes in a network and the strength of the ties.

O

Object Databases – they store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

Object-based Image Analysis – analyzing digital images can be performed with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.

Operational Databases – they carry out regular operations of an organization and are generally very important to a business. They generally use online transaction processing that allows them to enter, collect and retrieve specific information about the company.

Optimization analysis - the process of optimization during the design cycle of products done by algorithms. It allows companies to virtually design many different variations of a product and to test that product against pre-set variables.

P

Pattern Recognition – identifying patterns in data via algorithms to make predictions of new data coming from the same source.

Petabytes - approximately 1000 terabytes or 1 million gigabytes. The CERN Large Hydron Collider generates approximately 1 petabyte per second

Predictive analysis – analysis within big data to help predict how someone will behave in the (near) future. It uses a variety of different data sets such as historical, transactional, or social profile data to identify risks and opportunities.

Privacy – to seclude certain data / information about oneself that is deemed personal

Public data – public information or data sets that were created with public funding

Q

Quantified Self – a movement to use application to track ones every move during the day in order to gain a better understanding about ones behavior

Query – asking for information to answer a certain question

R

Re-identification – combining several data sets to find a certain person within anonym zed data

Regression analysis – to define the dependency between variables. It assumes a one-way causal effect from one variable to the response of another variable.

RFID – Radio Frequency Identification; a type of sensor using wireless non-contact radio-frequency electromagnetic fields to transfer data

Real-time data – data that is created, processed, stored, analyzed and visualized within milliseconds

S

Scripting - the use of a computer language where your program, or script, can be run directly with no need to first compile it to binary code. **Semi-structured data** - a form a structured data that does not have a formal structure like structured data. It does however have tags or other markers to enforce hierarchy of records.

Sentiment Analysis – using algorithms to find out how people feel about certain topics

Similarity searches – finding the closest object to a query in a database, where the data object can be of any type of data.

Simulation analysis – a simulation is the imitation of the operation of a real-world process or system. A simulation analysis helps to ensure optimal product performance taking into account many different variables.

Smart grid – refers to using sensors within an energy grid to monitor what is going on in real-time helping to increase efficiency

Spatial analysis – refers to analysing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.
SQL – a programming language for retrieving data from a relational database
Structured data – data that is identifiable as it is organized in structure like rows and columns.

T
Tableau - a commercial data visualization package often used in data science projects. **Terabytes** – approximately 1000 gigabytes.
Time series analysis - analyzing well-defined data obtained through repeated measurements of time. The data has to be well defined and measured at successive points in time spaced at identical time intervals.
Topological Data Analysis – focusing on the shape of complex data and identifying clusters and any statistical significance that is present within that data.

U
Un-structured data - unstructured data is regarded as data that is in general text heavy, but may also contain dates, numbers and facts.

V
Variability – it means that the meaning of the data can change (rapidly). In (almost) the same tweets for example a word can have a totally different meaning
Variety – data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data
Velocity – the speed at which the data is created, stored, analyzed and visualized
Veracity – ensuring that the data is correct as well as the analyses performed on the data are correct.
Visualization – complex graphs that can include many variables of

data while still remaining understandable and readable
Volume – the amount of data, ranging from megabytes to brontobytes

X
XML Databases – XML Databases allow data to be stored in XML format. The data stored in an XML database can be queried, exported and serialized into any format needed.

Y
Yottabytes – approximately 1000 Zettabytes, or 250 trillion DVD's. The entire digital universe today is 1 Yottabyte and this will double every 18 months.

Z
Zettabytes – approximately 1000 Exabytes or 1 billion terabytes.

Big Data

What is Big Data?

The term “big data” refers to the massive amount of data available to organizations that — because of its volume and complexity — is not easily managed or analyzed by many business intelligence tools. Basics of big data include the volume of the data collected, the speed at which that data becomes available to an organization for analysis, and the complexity (or varieties) of that data.

How big is big data?

According to Forbes, there are 2.5 quintillion bytes of data created every day. Because big data is so big, new terminology is required to define the size of that data. Big data consists of petabytes (more than 1 million gigabytes) and exabytes (more than 1 billion gigabytes), as opposed to the gigabytes common for personal devices.

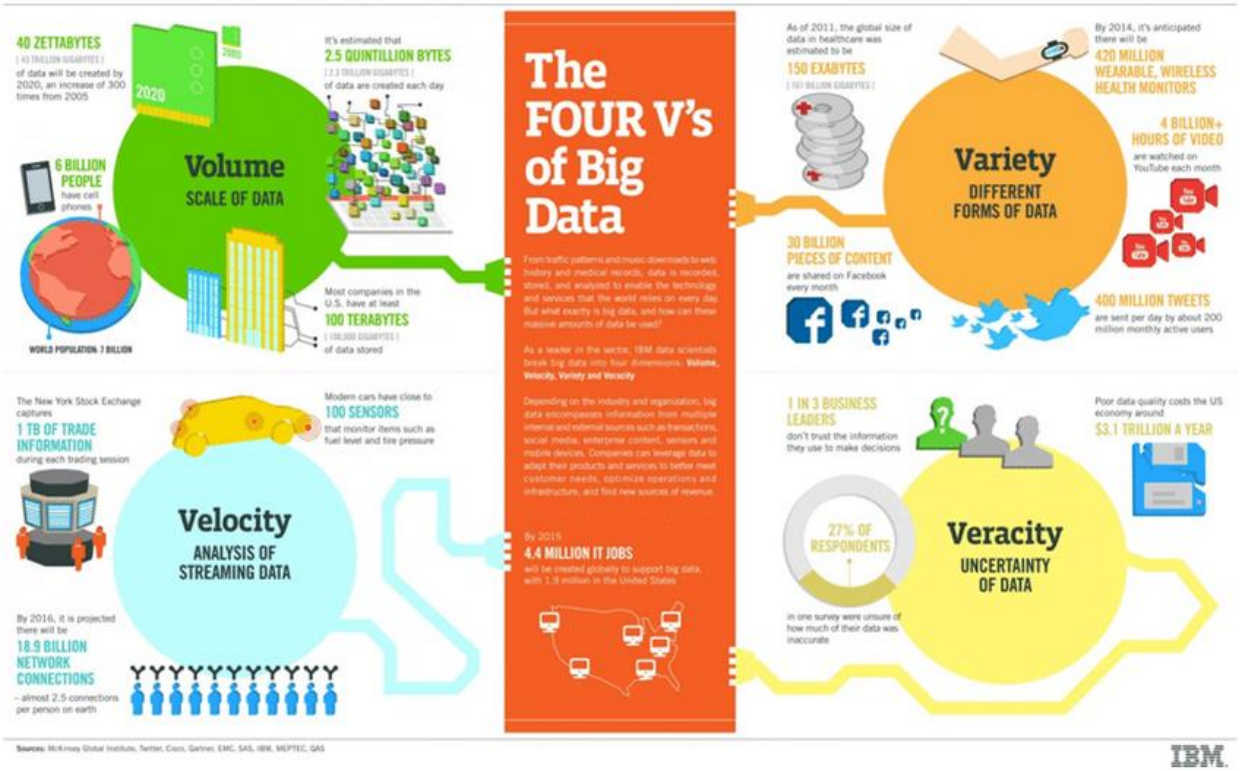
How can you access big data?

As big data emerged, so did computing models with the ability to store and manage it. Centralized or distributed computing systems provide access to big data. Centralized computing means the data is stored on a central computer and processed by computing platforms like BigQuery, Snowflake, Azure, or AWS.
Distributed computing means big data is stored and processed on different computers, which communicate over a network. A software framework like Hadoop makes it possible to store the data and run applications to process it.

There are benefits to using centralized computing and analyzing big data where it lives, rather than extracting it for analysis from a distributed system. Insights are accessible to every user in your company — and integrated into daily workflows — when big data is housed in one place and analyzed by one platform.

Characteristics of big data

Big data is different from typical data assets because of its volume complexity and need for advanced business intelligence tools to process and analyze it. The attributes that define big data are volume, variety, velocity, and variability (commonly referred to as the four v's).



Volume

The key characteristic of big data is its scale — the volume of data that is available for collection by your enterprise from a variety of devices and sources.

Variety

Variety refers to the formats that data comes in, such as email messages, audio files, videos, sensor data, and more. Classifications of big data variety include structured, semi-structured, and unstructured data.

- Structured data usually refers to data that adheres to a defined structure or model, which makes it easier to analyze. Examples of structured data can include spreadsheets or a list of customer addresses.
- Semi-structured data does not conform to a defined data model, but the data have semantic tags that make data organization and search easier. An example of semi-structured data is HTML code.
- Unstructured data is data that is not organized by a predefined model. Examples of unstructured data can include emails and satellite images.

Velocity

Big data velocity refers to the speed at which large data sets are acquired, processed, and accessed.

Variability

Big data variability means the meaning of the data constantly changes. Therefore, before big data can be analyzed, the context and meaning of the data sets must be properly understood.

Examples and applications of big data

The varied and high-volume, high-velocity big data your enterprise manages is a vital asset, one that can drive enhanced decision-making for improved business outcomes. Harnessing

big data through effective data analytics provides many competitive advantages. Applications of big data include:

Hyper-personalization in retail with big data

Insight from big data can help retail enterprises better understand their customers' preferences and behaviors. With that understanding, a retailer can hyper-personalize marketing initiatives and shopping experiences that redefine the customer experience.

Streamlining process in finance with big data

Applications of big data can help firms make the most of their financial data to improve operational efficiencies by streamlining the time and processes to actionable insights. This streamlining minimizes bottlenecks and allows more time for identifying new revenue opportunities.

Scaling expansion with big data

There are insights hidden in big data. Those insights help companies enhance performance, boost competitiveness, and effectively adjust the business model for successful expansion into new markets.

Why is big data important?

Data can be a company’s most valuable asset. Using big data to reveal insights can help understand the areas that affect your business — from market conditions and customer purchasing behaviors to your business processes. These understandings help drive impactful decision-making.

Chapter 4:

DEPTH DIGING INTENSIVE Overview of Big Data Tools and Techniques, Discoveries and Pitfalls

Overview of Big Data

Tools and Techniques, Discoveries and Pitfalls

What Does "Big Data" Mean?

(1) Collecting large amounts of data

Via computers, sensors, people, events …

(2) Doing something with it

Making decisions, confirming hypotheses, gaining insights, predicting future …

"Data Science" = Going from (1) to (2)

Big Data is Here to Stay

- Ability to collect data will only increase
- Ability to analyze data will only improve

This Overview

- Promises of Big Data
  - Applications and services
- Big Data tools and techniques
  - Database management systems
  - Data mining and machine learning
- Pitfalls of Big Data
  - Correlation and causation
  - Underfitting and overfitting
  - Privacy and a few others
- Big Data systems and platforms

Promises of Big Data

(1) Collect large amounts of data

(2) Do something with it

beneficial

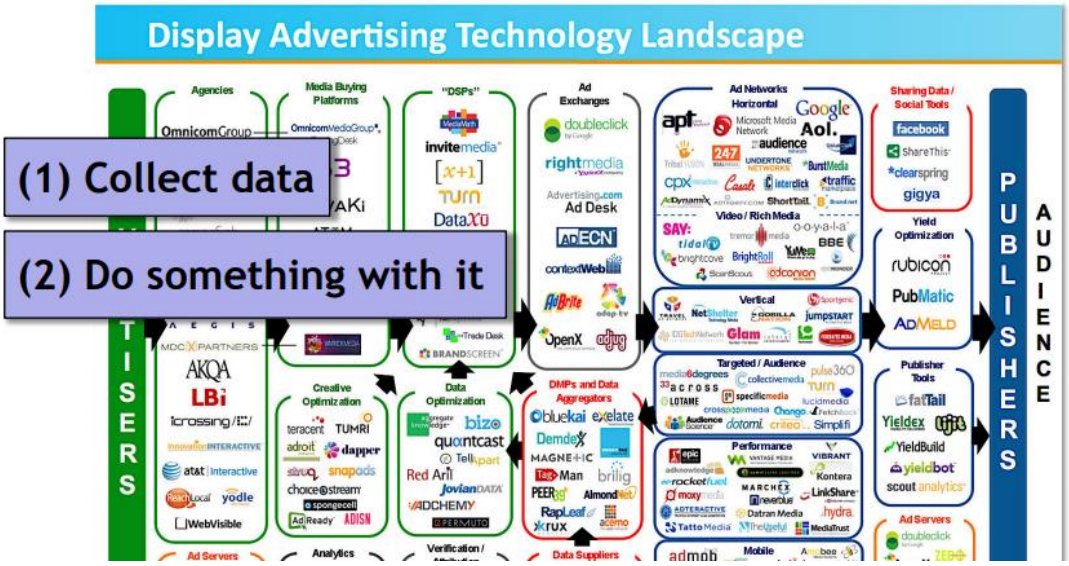
Traffic

(1) Collect data
(2) Do something with it
Menlo Park
DOWNTOWN MENLO PARK
Staples
Safeway
ALLIED ARTS
Allied Arts Guild
Stanford Shopping Center
El Camino Park
west elm
Philz Coffee
Palo Alto Junior Museum and Zoo
PROFESSORVILLE
Arboretum Grove
Town & Country Village
Stanford Hospital
OLD PAL

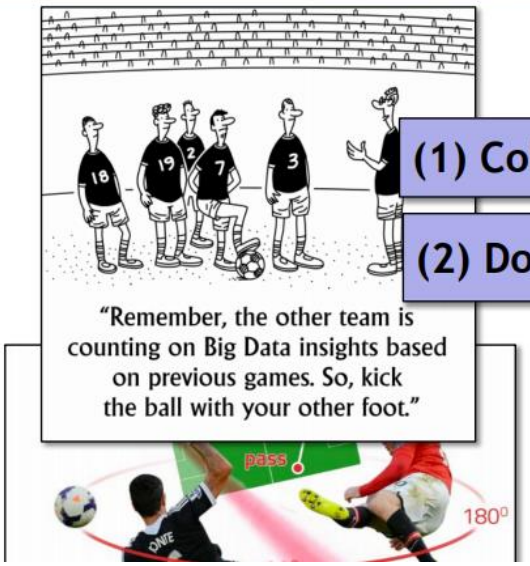
Recommender Systems

+ music, news, friends, romantic partners, and many more!

Online Advertising



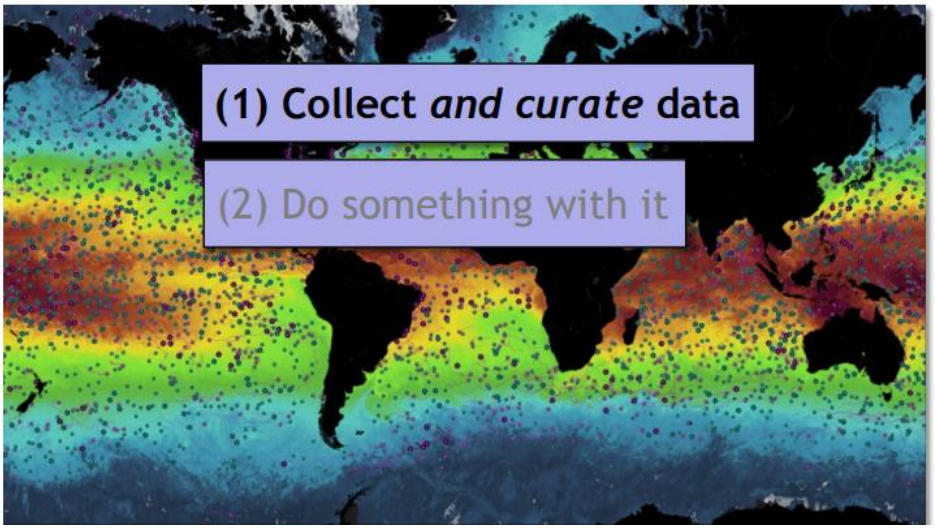
Sports



(1) Collect data

(2) Do something with it

Ocean Health

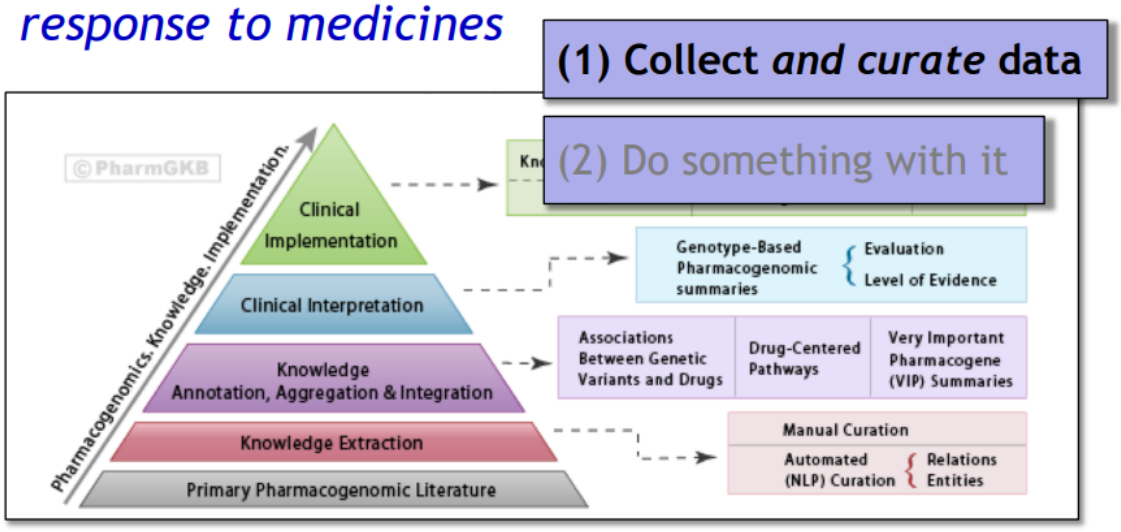


44,000 sensors, over 2 billion measurements

Physical, chemical, biological …

Genetics-Medicine Relationships

PharmGKB collects, curates, and disseminates knowledge about how human genetics affects response to medicines



And Many More

- Weather prediction
- Medical diagnosis
- Financial markets
- Resource management
- Computational social science
- Smart buildings and cities
-

The list goes on and on, and it's still early days

Big Data Tools and Techniques

- Basic Data Manipulation and Analysis
  Performing well-defined computations or asking well-defined questions ("queries")
- Data Mining
  Looking for patterns in data
- Machine Learning
  Using data to build models and make predictions
- Data Visualization
  Graphical depiction of data
- Data Collection and Preparation

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Average January low temperature for each country over last 20 years
- Number of items over $100 bought by females between ages 20 and 30
- Frequency of specific medicine relieving specific symptoms
- The ten stocks whose price varied the most over the past year

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Ave … h cou…
- Nu… fe…
- Fre… specific symptoms
- The ten stocks whose price varied the most over the past year

- Spreadsheets
- Relational (SQL) database systems
- "NoSQL" / scalable systems
- Programming languages with big-data support (e.g., Python, R)

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People who like movie X also like movie Y
- Patients who respond well to medicines X and Y also respond well to medicine Z
- Students going to the same university are frequently online friends
- Wealthier people are moving from cities to suburbs

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People ... ie Y
- Patien ... es X and Y ...
- Studer ... are frequently online friends
- Wealthier people are moving from cities to suburbs

- Frequent item-sets
- Association rules
- Specialized techniques for networks, text, multimedia

Data Visualization

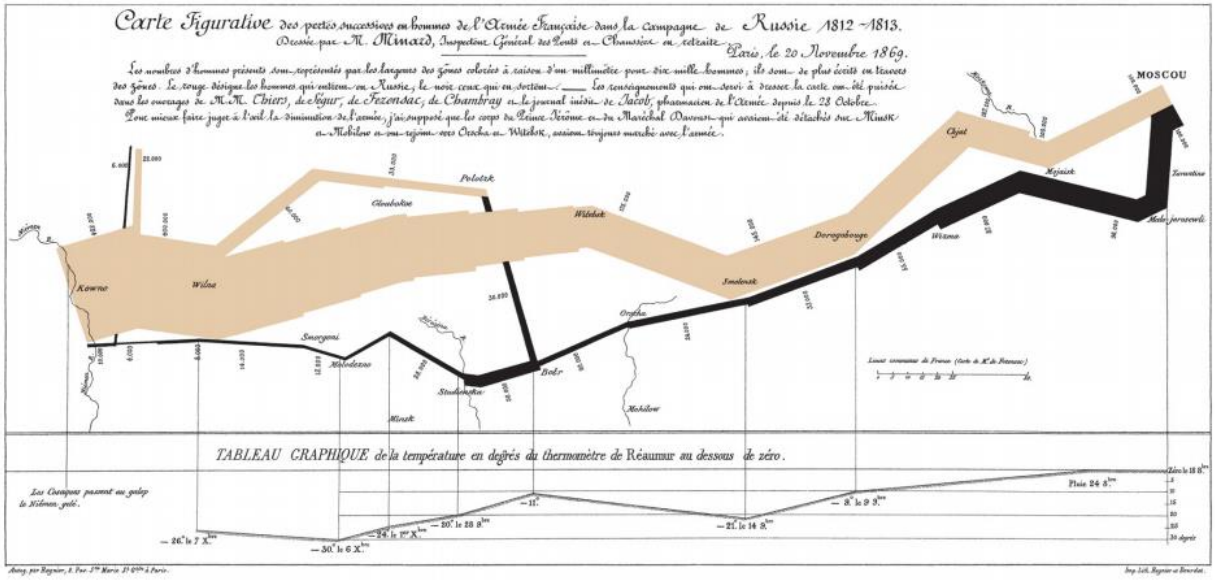
"A picture is worth a thousand words"

Data Visualization

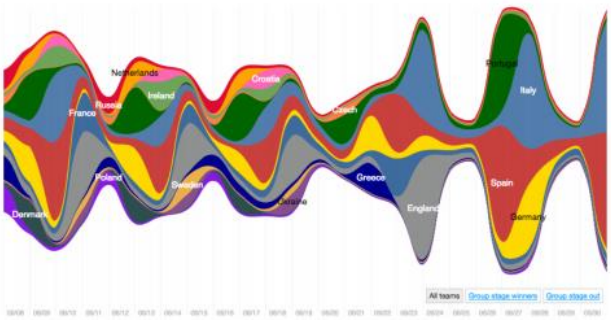
"A picture is worth a ~~thousand words~~"
trillion data points

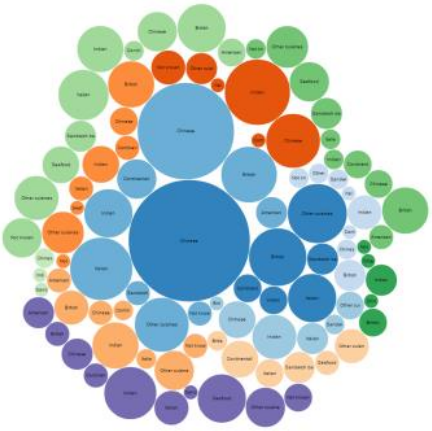
Early Data Visualization

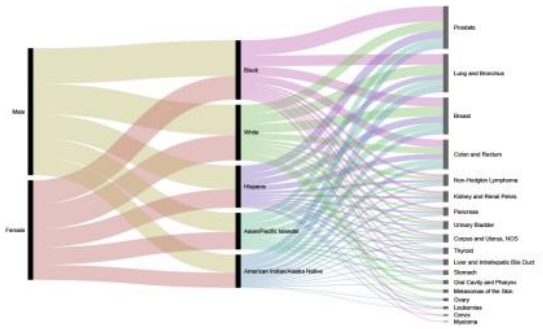
Napoleon's Army



Fancy Data Visualization



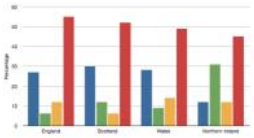


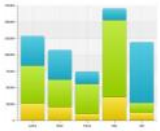


Basic Data Visualization

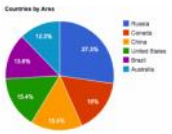
Don't underestimate the power of basic visualizations

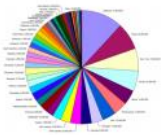
- Bar charts



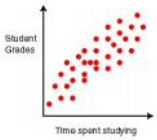


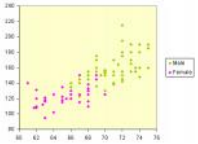
- Pie charts





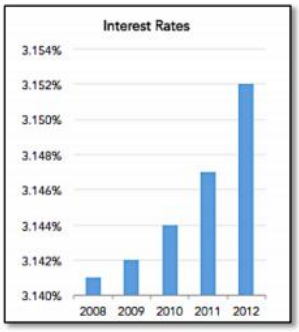
- Scatterplots

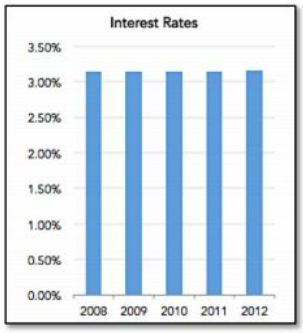


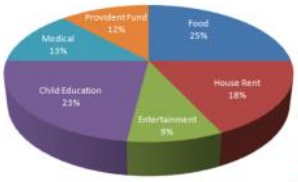


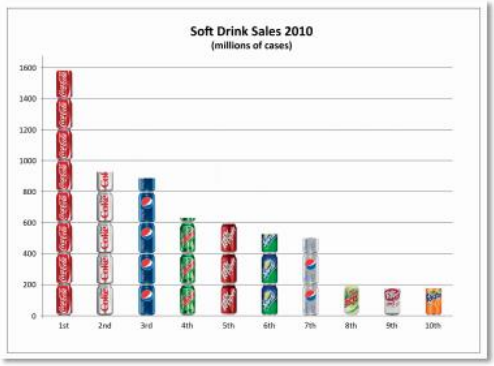
- Maps

Misleading Data Visualization









Data Collection and Preparation

The "dirty" secret of Big Data

- Extracting data from difficult sources
- Filling in missing values
- Removing suspicious data
- Making formats, encoding, and units consistent
- De-duplicating and matching

Data preparation often consumes 80% or more of the effort in a Big Data project

Pitfalls of Big Data

(1) Collect large amounts of data

(2) Do something with it

correct

Correlation and Causation

Data analysis, data mining, and machine learning can reveal relationships between data values

Correlation – Values track each other

- Height and Shoe Size
- Grades and SAT Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

Correlation – Values track each other

- Height and Shoe Size
- Grades and SAT Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B
  - ❖ Homeless population and crime rate
    Confounding variable: unemployment
  - ❖ Forgetfulness and poor eyesight
    Confounding variable: age
  - ❖ Height and shoe size
  - ❖ Grades and SAT scores

Correlation and Causation

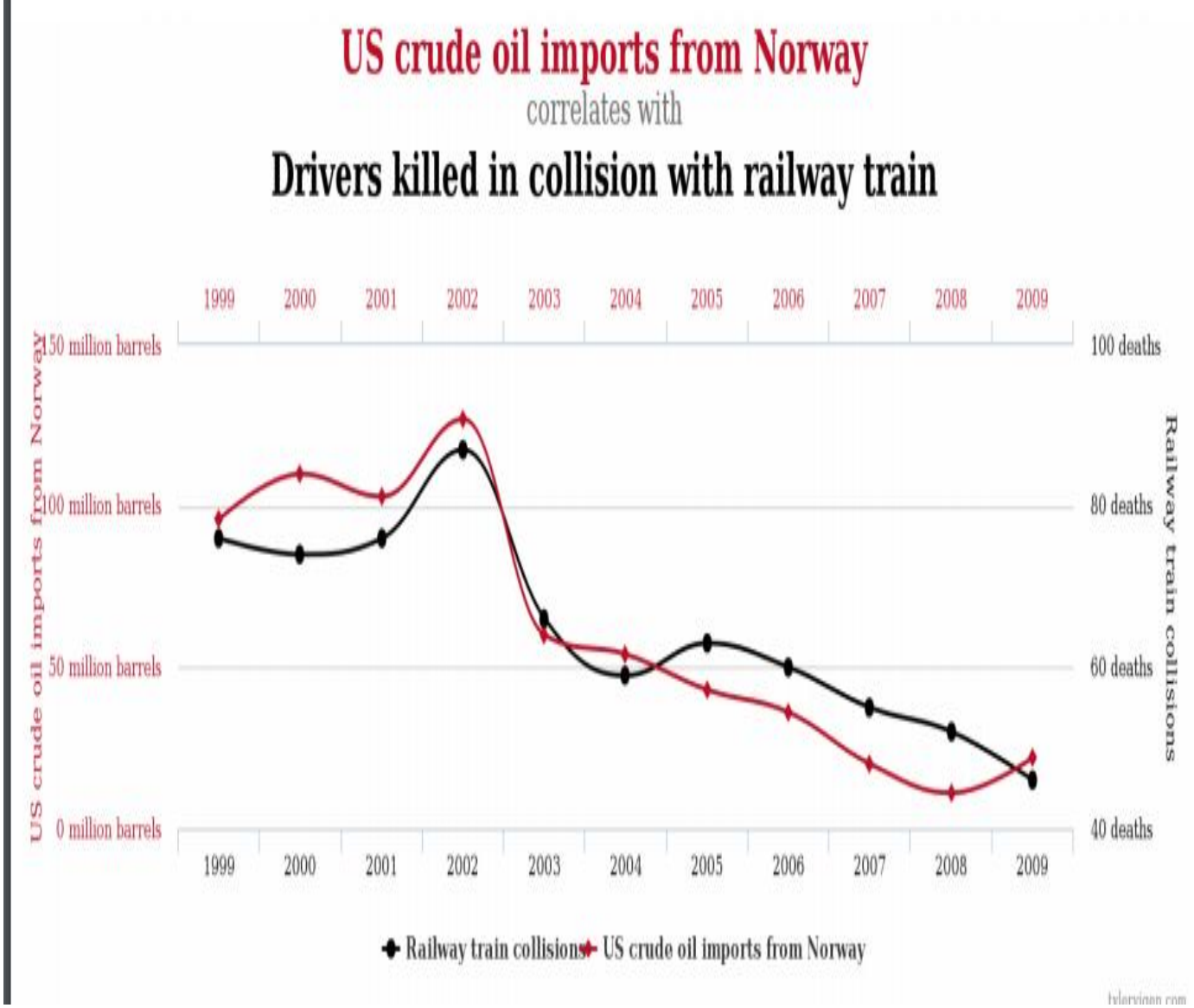
"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B

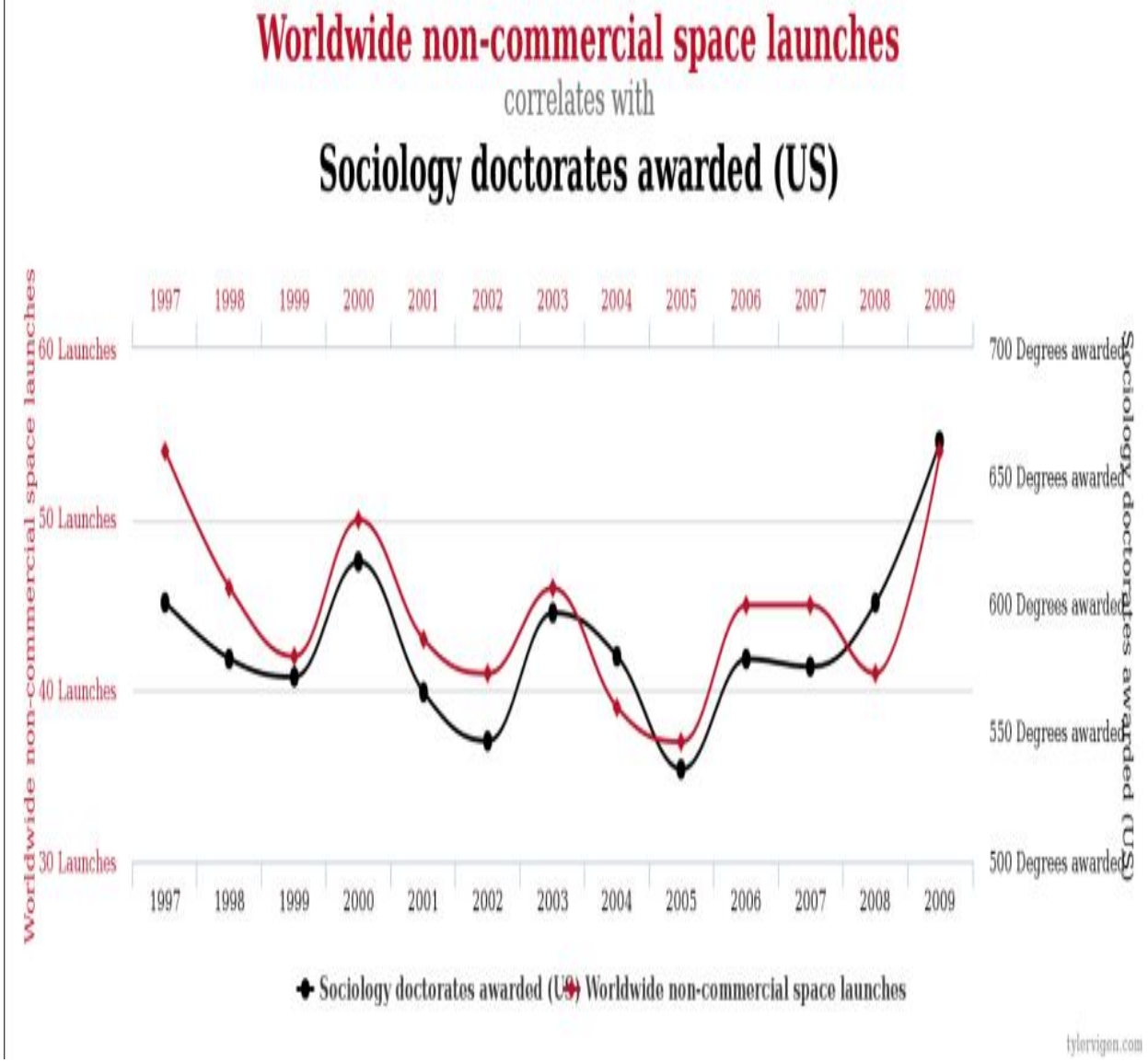
- Correlation is usually "easy" to test
- Causation is impossible to test

Correlation and Causation

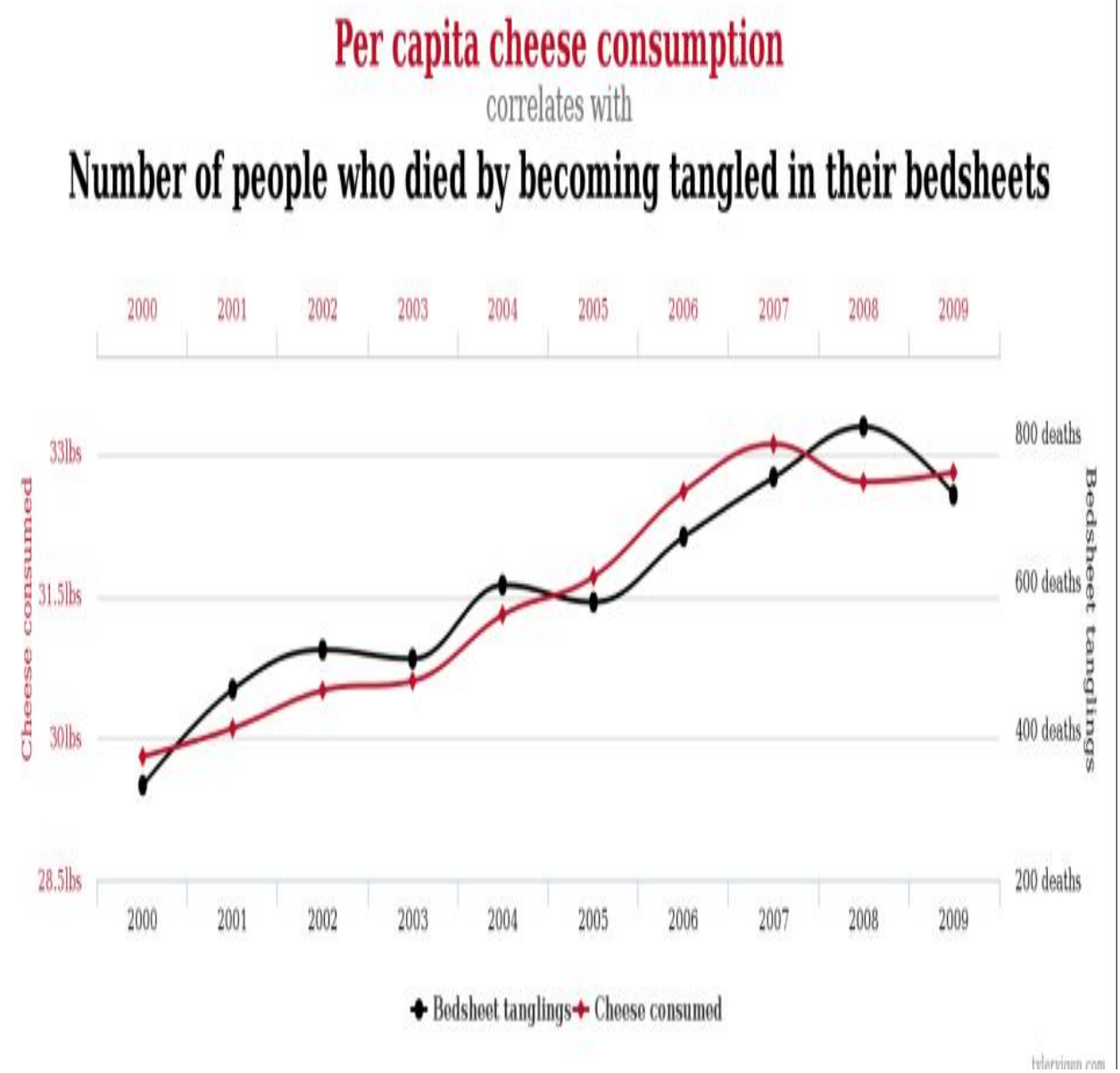
Surprising Correlation #1



Surprising Correlation #2



Surprising Correlation #3



“Spurious Correlations” Website

http://www.tylervigen.com/

Underfitting and Overfitting

Machine learning uses data to create a "model" and uses model to make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on the SAT
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Underfitting

Model used for predictions is **too simplistic**

- 60% of men and 70% of women responded to an advertisement, therefore all future ads should go to women
- If a furniture item has four legs and a flat top it is a dining room table
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

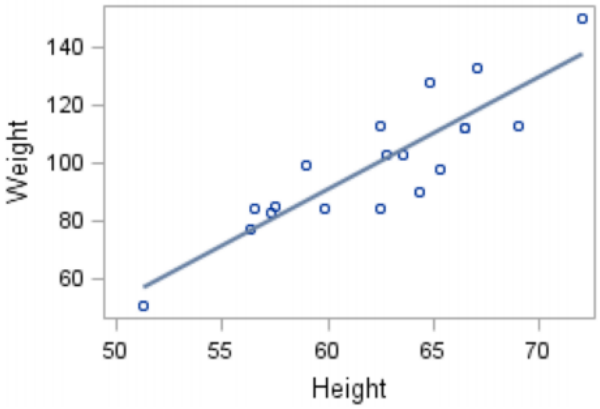
Overfitting

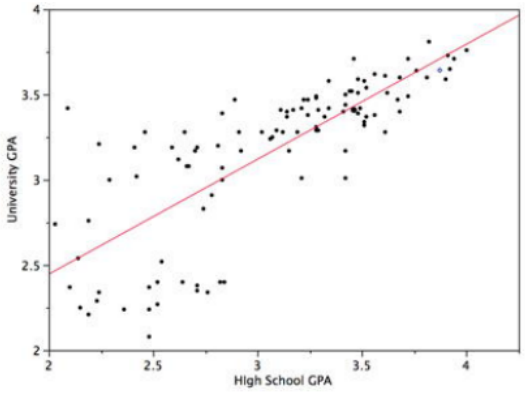
Model used for predictions is **too specific**

- The best targets for an advertisement are married women between 25 and 27 years with short black hair, one child, and one pet dog
- If a furniture item has four 100 cm legs with decoration and a flat polished wooden top with rounded edges then it is a dining room table

Regression

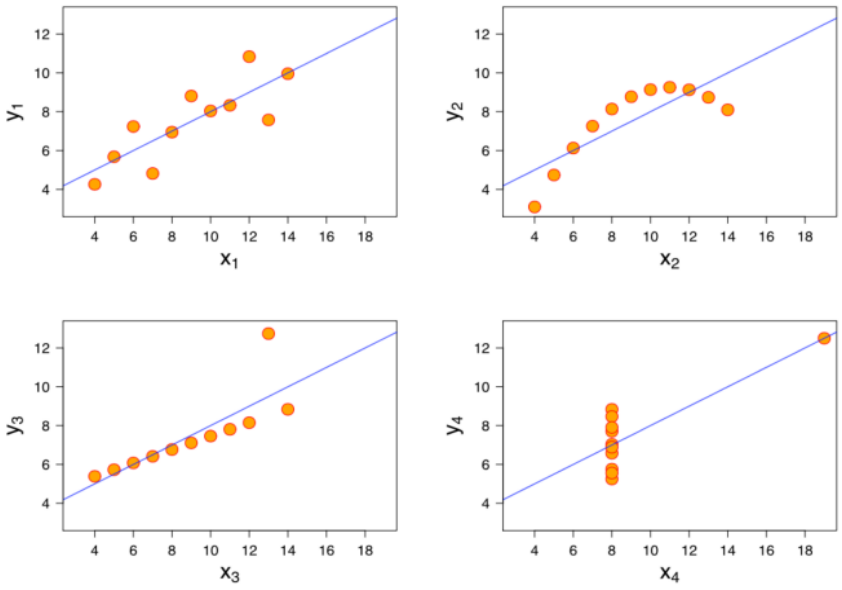
- Fit a line or curve to a set of points (model)
- Use model to predict values for new points





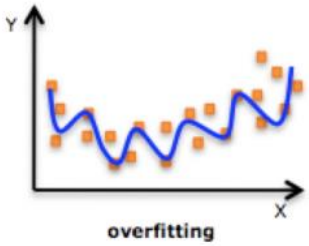
Underfitting

Model is too simplistic

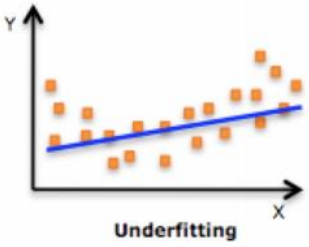


Overfitting

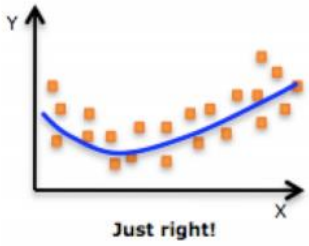
Model is **too specific**



overfitting



Underfitting



Just right!

Big Data Scam: Soccer Match Prediction

- Friday: receive email from "Psychic Sally" predicting which teams will be the winners in the weekend's five soccer matches. She's right about all of them!
- Same thing the following weekend: five games, all winners predicted correctly
- And the following one: five more correct
- Fourth Friday: Sally offers to give you her predictions for the coming weekend's games, for a fee

Should you do it?

Big Data Scam: Soccer Match Prediction

How many contacts must Sally start with on week one to ensure she has 100 potential buyers by week four, i.e., 100 people who received 15 correct predicted winners?

(Assume no draws)

Data Privacy

- Individual data collected covertly
  - Edward Snowden, “metadata” argument
- Individual data collected legally but used questionably
  - Individual “digital footprints” are enormous
  - Target stores pregnancy mailing
  - Engagement ring purchase broadcast on Facebook
- Individual data deduced from “anonymous” public data
  - Governor of Massachusetts health record

Languages, Systems, Platforms

- Spreadsheets

  Surprisingly versatile and powerful for data analysis tasks, but not truly big data

- Programming languages with big-data support
  - R Language – powerful statistical features
  - Python – general-purpose language with R-like add-ons (Pandas, SciPy, scikit-learn)

Languages, Systems, Platforms

- Relational Database Management Systems
  - Also called RDBMS, SQL Systems
  - Long-standing solution for reliability, efficiency, powerful query processing
  - Works for all but truly extreme data sizes, or highly unstructured data
- "NoSQL" Systems
  - Distributed/scalable processing, unstructured data
  - Key-value row stores (e.g., Cassandra, Dynamo)
  - Document databases (e.g., MongoDB, CouchDB)
  - Graph databases (e.g., Neo4J, Giraph)

Languages, Systems, Platforms

- Specialized languages on scalable systems
  - MapReduce / Hadoop
  - Spark generalized data flow
- Systems for data preparation
- Systems for data visualization

Languages, Systems, Platforms

- Data processing in the cloud
  - Amazon Web Services, Google Cloud, Microsoft Azure
  - Data storage
  - Data processing: SQL, Hadoop, Spark
  - Machine learning libraries
  - Integration with visualization systems

Complete works of William Shakespeare
5 megabytes

Average individual
50 gigabytes (10,000 Shakespeares)

USA Library of Congress
10 terabytes (2 million Shakespeares)

Uploaded to Facebook daily
1 petabyte (200 million Shakespeares)

Produced by humanity daily
2.5 exabytes (500 trillion Shakespeares)

Size Isn't Everything

- Tools and techniques apply to data of all sizes
- Big insights can come from small/medium data

> Some applications actually do need twenty Spark servers in the cloud. More often a laptop with SQL, Python, or simple spreadsheets does the job.

What We'll Cover

- Data Analysis & Visualization Using Spreadsheets
- Advanced Data Visualization Using Tableau
- Relational Databases and SQL
- Python for Data Analysis & Visualization
- Machine Learning – Regression, Classification, Clustering
- Using Python for Machine Learning
- The R Language - Data Analysis, Visualization, and Machine Learning
- Data Mining Algorithms
- Data Mining Using SQL and Python
- Network Analysis
- Unstructured Data
- Guest speaker –Big Data Platform and Services

Chapter 5:

Big Data Analytics

What is big data analytics?

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media — much of it generated in real time and at a very large scale.

Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

Making Sense of Big Data: A Day in the Life of an Enterprise Architect

The rise of cloud storage — and the demand for remote, cloud-based access — has transformed the role of the enterprise data architect. Instead of simply managing an in-house database, architects now need to handle a host of new challenges related to data integrity, security, integration and analysis.

This Enterprise Management Associates (EMA) white paper reviews the challenges that confront an enterprise architect, including new data sources, more data projects, and more data platforms. The authors explain how enterprise architects can nurture an in-house approach that's able to handle these challenges, including:

- **Using customizable applications** to make implementations faster
- **Training internal staff** instead of hiring expensive contractors
- **Collaborating closely with business stakeholders** to help align data strategy with business needs
- **Deploying a hybrid data ecosystem** to provide the foundation for self-service projects

EMA argues that enterprise data architects need to go beyond data management and help design the enterprise data strategy, future-proofing it against other challenges that will inevitably arise.

Big Data Analytics

What it is and why it matters

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze your data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions.

History and evolution of big data analytics

The concept of big data has been around for years; most organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get significant value from it. But even in the 1950s, decades before anyone uttered the term "big data," businesses were using basic analytics (essentially numbers in a spreadsheet that were manually examined) to uncover insights and trends.

The new benefits that big data analytics brings to the table, however, are speed and efficiency. Whereas a few years ago a business would have gathered information, run analytics and unearthed information that could be used for future decisions, today that business can identify insights for immediate decisions. The ability to work faster – and stay agile – gives organizations a competitive edge they didn't have before.

Why is big data analytics important?

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report *Big Data in Big Companies*, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways:

1. **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Big data analytics in today's world

Most organizations have big data. And many understand the need to harness that data and extract value from it. But how? These

resources cover the latest thinking on the intersection of big data and analytics.

Statistics and Machine Learning at Scale

The concept of machine learning has been around for decades and now it can now be applied to huge quantities of data.

Big Data Analytics

Big Data analytics is the process of collecting, organizing and analyzing large sets of data (*called* Big Data) to discover patterns and other useful information. Big Data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with Big Data typically want the *knowledge* that comes from analyzing the data.

High-Performance Analytics Required

To analyze such a large volume of data, Big Data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics. Using Big Data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.

Zoho Analytics is business intelligence & analytics software that transforms large amounts of raw data into actionable reports and dashboards. Track key business metrics, see long term trends, determine outliers, and discover hidden insights with a business intelligence platform trusted by some of the world's largest companies. Customers include big name brands such as Tesla, Suzuki, and more.

The Challenges

For most organizations, Big Data analysis is a challenge. Consider the sheer volume of data and the different formats of the data (both structured and unstructured data) that is collected across the entire organization and the many different ways different types of data can be combined, contrasted and analyzed to find patterns and other useful business information.

The first challenge is in breaking down data silos to access all data an organization stores in different places and often in different systems. A

second challenge is in creating platforms that can pull in unstructured data as easily as structured data. This massive volume of data is typically so large that it's difficult to process using traditional database and software methods.

How Big Data Analytics is Used Today

As the technology that helps an organization to break down data silos and analyze data improves, business can be transformed in all sorts of ways. Today's advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook.

Another example comes from one of the biggest mobile carriers in the world. France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety. Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment.

The Benefits of Big Data Analytics

Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.

Webopedia parent company, QuinStreet, surveyed 540 enterprise decision-makers involved in big data purchases to learn which business areas companies plan to use Big Data analytics to improve operations. About half of all respondents said they were applying big data analytics to improve customer retention, help with product development and gain a competitive advantage.

Notably, the business area getting the most attention relates to increasing efficiency and optimizing operations. Specifically, 62 percent of respondents said that they use big data analytics to improve speed and reduce complexity.

Big Data Analytics

Big data analytics enables companies to increase revenues, decrease costs and become more competitive within their industries. Many firms are investing heavily in Big Data analytics.

Big data analytics is quickly gaining adoption. Enterprises have awakened to the reality that their big data stores represent a largely untapped gold mine that could help them lower costs, increase revenue and become more competitive. They don't just want to store their vast quantities of data, they want to convert that data into valuable insights that can help improve their companies.

As a result, investment in big data analytics tools is seeing remarkable gains. According to IDC, worldwide sales of big data and business analytics tools are likely to reach $150.8 billion in 2017, which is 12.4 percent higher than in 2016. And the market research firm doesn't see that trend stopping anytime soon. It

forecasts 11.9 percent annual growth through 2020 when revenues will top $210 billion.

A business intelligence & analytics tool from Zoho Analytics can transform big data into actionable reports and dashboards. Track key business metrics, see long term trends, determine outliers, and discover hidden insights with a business intelligence platform trusted by some of the world's largest companies. Customers include notable brands such as Tesla, Suzuki, and more.

Clearly, the trend toward big data analytics is here to stay. IT professionals need to familiarize themselves with the topic if they want to remain relevant within their companies.

What is Big Data Analytics?

The term "big data" refers to digital stores of information that have a high volume, velocity and variety. Big data analytics is the process of using software to uncover trends, patterns, correlations or other useful insights in those large stores of data.

Data analytics isn't new. It has been around for decades in the form of business intelligence and data mining software. Over the years, that software has improved dramatically so that it can handle much larger data volumes, run queries more quickly and perform more advanced algorithms.

The market research firm Gartner categories big data analytics tools into four different categories:

1. **Descriptive Analytics**: These tools tell companies what happened. They create simple reports and visualizations that show what occurred at a particular point in time or over a period of time. These are the least advanced analytics tools.
2. **Diagnostic Analytics**: Diagnostic tools explain why something happened. More advanced than descriptive reporting tools, they allow analysts to dive deep into the data and determine root causes for a given situation.
3. **Predictive Analytics**: Among the most popular big data analytics tools available today, predictive analytics tools use highly advanced algorithms to forecast what might happen next. Often these tools make use of artificial intelligence and machine learning technology.

4. **Prescriptive Analytics**: A step above predictive analytics, prescriptive analytics tell organizations what they should do in order to achieve a desired result. These tools require very advanced machine learning capabilities, and few solutions on the market today offer true prescriptive capabilities.

Types of Big Data Analytics				
	Descriptive	**Diagnostic**	**Predictive**	**Prescriptive**
Answers the question…	What happened?	Why did it happen?	What will happen next?	What should I do?
Level of advancement	Low	Medium	High	Very high
Incorporates AI and machine learning?	Not usually	Sometimes	Usually	Always
Level of popularity	Used by almost all organizations	Used by many organizations	Used by a smaller but growing group of organizations	Not yet widespread

Source: Gartner and others

Benefits of Big Data Analytics

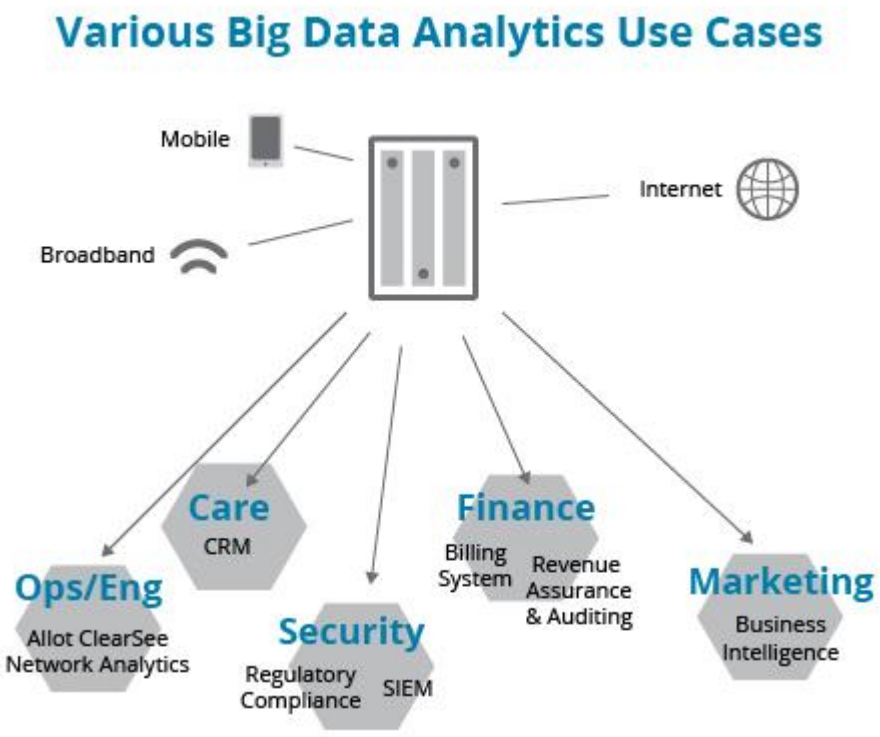
Organizations decide to deploy big data analytics for a wide variety of reasons, including the following:

- **Business Transformation** In general, executives believe that big data analytics offers tremendous potential to

revolution their organizations. In the 2016 Data & Analytics Survey from IDGE, 78 percent of people surveyed agreed that over the next one to three years the collection and analysis of big data could fundamentally change the way their companies do business.

- **Competitive Advantage** In the MIT Sloan Management Review Research Report Analytics as a Source of Business Innovation, sponsored by SAS, 57 percent of enterprises surveyed said their use of analytics was helping them achieve competitive advantage, up from 51 percent who said the same thing in 2015.
- **Innovation**Big data analytics can help companies develop products and services that appeal to their customers, as well as helping them identify new opportunities for revenue generation. Also in the MIT Sloan Management survey, 68 percent of respondents agreed that analytics has helped their company innovate. That's an increase from 52 percent in 2015.
- **Lower Costs** In the NewVantage Partners Big Data Executive Survey 2017, 49.2 percent of companies surveyed said that they had successfully decreased expenses as a result of a big data project.

- **Improved Customer Service** Organizations often use big data analytics to examine social media, customer service, sales and marketing data. This can help them better gauge customer sentiment and respond to customers in real time.
- **Increased Security** Another key area for big data analytics is IT security. Security software creates an enormous amount of log data. By applying big data analytics techniques to this data, organizations can sometimes identify and thwart cyberattacks that would otherwise have gone unnoticed.



Big data analytics can offer key advantages across many verticals.

Big Data Analytics Challenges

Implementing a big data analytics solution isn't always as straightforward as companies hope it will be. In fact, most surveys find that the number of organizations experiencing a measurable financial benefit from their big data analytics lags behind the number of organizations implementing big data analytics. Several different obstacles can make it difficult to achieve the benefits promised by big data analytics vendors:

- **Data Growth** One of the biggest challenges of big data analytics is the explosive rate of data growth. According to IDC, the amount of data in the world's servers is roughly doubling every two years. By 2020, those servers will likely hold 44 zettabytes of digital information. To put that in perspective, that is enough data to fill a stack of iPads stretching from the earth to the moon 6.6 times. Big data analytics solutions must be able to perform well at scale if they are going to be useful to enterprises.
- **Unstructured Data** Must of the data stored in an enterprise's systems doesn't reside in structured databases.

Instead, it is unstructured data, such as email messages, images, reports, audio files, videos and other types of files. This unstructured data can be very difficult to search—unless you have advanced artificial intelligence capabilities. Vendors are constantly updating their big data analytics tools to make them better at examining and extracting insights from unstructured data.

- **Data Siloes** Enterprise data is created by a wide variety of different applications, such as enterprise resource planning (ERP) solutions, customer relationship management (CRM) solutions, supply chain management software, ecommerce solutions, office productivity programs, etc. Integrating the data from all these different sources is one of the most difficult challenges in any big data analytics project.
- **Cultural Challenges** Although big data analytics is becoming commonplace, it hasn't infiltrated the corporate culture everywhere yet. In the NewVantage Partners Survey, 52.5 percent of executives said that organizational hurdles like lack of alignment, internal resistance or lack of

Big Data Analytics Trends

What's coming next for the big data analytics market? Experts offer a number of predictions.

Open Source

As big data analytics increases its momentum, the focus is on open-source tools that help break down and analyze data. Hadoop, Spark and NoSQL databases are the winners here. Even proprietary tools now incorporate leading open source technologies and/or support those technologies. That seems unlikely to change for the foreseeable future.

Market Segmentation

Plenty of general-purpose big data analytics platforms have hit the market, but expect even more to emerge that focus on specific niches, such as security, marketing, CRM, application performance monitoring and hiring. Analytics tools are also being integrated into existing enterprise software at a rapid rate.

Artificial Intelligence and Machine Learning

As interest in AI has skyrocketed, vendors have rushed to incorporate machine learning and cognitive capabilities into their big data analytics tools. According to Gartner, by 2020, almost

every new software product, including big data analytics, will incorporate AI technologies. In addition, the company says, "By 2020, AI will be a top five investment priority for more than 30 percent of CIOs."

Prescriptive Analytics

Fueled by this rush to AI, expect companies to become more interested in prescriptive analytics. Seen by many as the "ultimate" type of big data analytics, these tools will not only be able to predict the future, they will be able to suggest courses of action that might lead to desirable results for organizations. But before these types of solutions can become mainstream, vendors will need to make advancements in both hardware and software.

Refocusing on the Human Decision-Making?

As machine learning improves and becomes a table stakes feature in analytics suites, don't be surprised if the human element initially gets downplayed, before coming back into vogue.

Two of the most famous Big Data prognosticators/pioneers are Billy Beane and Nate Silver. Beane popularized the idea of correlating various statistics with under-valued player traits in order to field an A's baseball team on the cheap that could compete with deep-pocketed teams like the Yankees.

Meanwhile, Nate Silver's effect was so strong that people who didn't want to believe his predictions created all sorts of analysis-free zones, such as Unskewed Polls (which, ironically, were ridiculously skewed). Many think of Silver as a polling expert, but Silver is also a master at Big Data analysis.

In each case, what mattered most was not the machinery that gathered in the data and formed the initial analysis, but the human on top analyzing what this all means. People can look at polling data and pretty much treat them as Rorscharch tests. Silver, on the other hand, pours over reams of data, looks at how various polls have performed historically, factors in things that could influence the margin of error (such as the fact that younger voters are often under-counted since they don't have landline phones) and emerges with incredibly accurate predictions.

Similarly, every baseball GM now values on-base percentage and other advanced stats, but few are able to compete as consistently on as little money as Beane's A's teams can. There's more to finding under-valued players than crunching numbers. You also need to know how to push the right buttons in order to negotiate trades with other GMs, and you need to find players who will fit into your system.

As Big Data analytics becomes mainstream, it will be like many earlier technologies. Big Data analytics will be just another tool. What you do with it, though, will be what matters.

Big Data Analytics Tools

Big data analytics has become so trendy that nearly every major technology company sells a product with the "big data analytics" label on it, and a huge crop of startups also offers similar tools. Cloud-based big data analytics have become particularly popular. In fact, the 2016 Big Data Maturity Survey conducted by AtScale found that 53 percent of those surveyed planned to use cloud-based big data solutions, and 72 percent planned to do so in the future. Open source tools like Hadoop are also very important, often providing the backbone to commercial solution.

The lists below are not exhaustive, but do include a sampling of some of better known big data analytics solutions.

Open Source Big Data Analytics Tools

- Hadoop
- HPCC
- Ignite
- Lumify

- Pandas
- Pivotal GemFire, Greenplum Database and HAWQ
- Spark
- Storm

Big Data Analytics Vendors

- Zoho Analytics *(**free trial available**)*
- Alpine Data
- Alteryx
- Amazon Web Services
- Cloudera
- Datameer
- Domo
- Google Cloud Platform
- HPE
- IBM
- Microsoft
- MicroStrategy
- Oracle
- Pentaho
- SAP
- SAS

- SiSense
- Splunk
- Tableau
- Teradata
- TIBCO Jaspersoft
- VMware

Big Data Analytics

Definition - What does *Big Data Analytics* mean?

Big data analytics refers to the strategy of analyzing large volumes of data, or big data. This big data is gathered from a wide variety of sources, including social networks, videos, digital images, sensors, and sales transaction records. The aim in analyzing all this data is to uncover patterns and connections that might otherwise be invisible, and that might provide valuable insights about the users who created it. Through this insight, businesses may be able to gain an edge over their rivals and make superior business decisions.

Techopedia explains *Big Data Analytics*

Big data analytics allows data scientists and various other users to evaluate large volumes of transaction data and other data sources that traditional business systems would be unable to

tackle. Traditional systems may fall short because they're unable to analyze as many data sources.

Sophisticated software programs are used for big data analytics, but the unstructured data used in big data analytics may not be well suited to conventional data warehouses. Big data's high processing requirements may also make traditional data warehousing a poor fit. As a result, newer, bigger data analytics environments and technologies have emerged, including Hadoop, MapReduce and NoSQL databases. These technologies make up an open-source software framework that's used to process huge data sets over clustered systems.

What is big data analytics? Fast answers from diverse data sets

Analyzing large volumes of data is only part of what makes big data analytics different from traditional data analytics.

There’s data, and then there’s big data. So, what’s the difference?

Big data defined As

A clear big data definition can be difficult to pin down because big data can cover a multitude of use cases. But in general the term refers to sets of data that are so large in volume and so complex that traditional data processing software products are not capable of capturing, managing, and processing the data within a reasonable amount of time.

These big data sets can include structured, unstructured, and semi structured data, each of which can be mined for insights.

How much data actually constitutes?

“Big” is open to debate, but it can typically be in multiples of petabytes—and for the largest projects in the exabytes range.

Often, big data is characterized by the three Vs:

- an extreme *volume* of data
- a broad *variety* of types of data
- the *velocity* at which the data needs to be processed and analyzed

The data that constitutes big data stores can come from sources that include web sites, social media, desktop and mobile apps, scientific experiments, and—increasingly—sensors and other devices in the internet of things (IoT).

The concept of big data comes with a set of related components that enable organizations to put the data to practical use and solve a number of business problems. These include the IT infrastructure needed to support big data technologies, the analytics applied to the data; the big data platforms needed for projects, related skill sets, and the actual use cases that make sense for big data.

What is data analytics?

What really delivers value from all the big data organizations are gathering is the analytics applied to the data? Without analytics, which involves examining the data to discover patterns, correlations, insights, and trends, the data is just a bunch of ones and zeros with limited business use.

By applying analytics to big data, companies can see benefits such as increased sales, improved customer service, greater efficiency, and an overall boost in competitiveness.

Data analytics involves examining data sets to gain insights or draw conclusions about what they contain, such as trends and predictions about future activity.

By analyzing information using big data analysis tools, organizations can make better-informed business decisions such as when and where to run a marketing campaign or introduce a new product or service.

Analytics can refer to basic business intelligence applications or more advanced, predictive analytics such as those used by scientific organizations. Among the most advanced type of data analytics is data mining, where analysts evaluate large data sets to identify relationships, patterns, and trends.

Data analytics can include exploratory data analysis (to identify patterns and relationships in data) and confirmatory data analysis (applying statistical techniques to find out whether an assumption about a particular data set is true.

Another distinction is quantitative data analysis (or analysis of numerical data that has quantifiable variables that can be compared statistically) vs. qualitative data analysis (which focuses on nonnumerical data such as video, images, and text).

IT infrastructure to support big data

For the concept of big data to work, organizations need to have the infrastructure in place to gather and house the data, provide access to it, and secure the information while it's in storage and in transit. This requires the deployment of big data analytics tools.

At a high level, these include storage systems and servers designed for big data, data management and integration software, business intelligence and data analytics software, and big data applications.

Much of this infrastructure will likely be on-premises, as companies look to continue leveraging their datacenter investments. But increasingly organizations rely on cloud computing services to handle much of their big data requirements.

Data collection requires having sources to gather the data. Many of these—such as web applications, social media channels, mobile apps, and email archives—are already in place. But as IoT becomes entrenched, companies might need to deploy sensors on all sorts of devices, vehicles, and products to gather data, as well as new applications that generate user data. (IoT-

oriented big data analytics has its own specialized techniques and tools.)

To store all the incoming data, organizations need to have adequate data storage in place. Among the storage options are traditional data warehouses, data lakes, and cloud-based storage.

Security infrastructure tools might include data encryption, user authentication and other access controls, monitoring systems, firewalls, enterprise mobility management, and other products to protect systems and data,

Big data technologies

In addition to the foregoing IT infrastructure used for data in general. There several technologies specific to big data that your IT infrastructure should support.

Hadoop ecosystem

Hadoop is one of the technologies most closely associated with big data. The Apache Hadoop project develops open source software for scalable, distributed computing.

The Hadoop software library is a framework that enables the distributed processing of large data sets across clusters of computers using simple programming models. It's designed to scale up from a single server to thousands, each offering local computation and storage.

The project includes several modules:

- Hadoop Common, the common utilities that support other Hadoop modules
- Hadoop Distributed File System, which provides high-throughput access to application data
- Hadoop YARN, a framework for job scheduling and cluster resource management
- Hadoop MapReduce, a YARN-based system for parallel processing of large data sets.

Apache Spark

Part of the Hadoop ecosystem, Apache Spark is an open source cluster-computing framework that serves as an engine for processing big data within Hadoop. Spark has become one of the key big data distributed processing frameworks, and can be deployed in a variety of ways. It provides native bindings for the Java, Scala, Python (especially the Anaconda Python distro), and R programming languages (R is especially well suited for big data), and it supports SQL, streaming data, machine learning, and graph processing.

Data lakes

Data lakes are storage repositories that hold extremely large volumes of raw data in its native format until the data is needed

by business users. Helping to fuel the growth of data lakes are digital transformation initiatives and the growth of the IoT. Data lakes are designed to make it easier for users to access vast amounts of data when the need arises.

NoSQL databases

Conventional SQL databases are designed for reliable transactions and ad hoc queries, but they come with restrictions such as rigid schema that make them less suitable for some types of applications. NoSQL databases address those limitations, and store and manage data in ways that allow for high operational speed and great flexibility. Many were developed by companies that sought better ways to store content or process data for massive websites. Unlike SQL databases, many NoSQL databases can be scaled horizontally across hundreds or thousands of servers.

In-memory databases

An in-memory database (IMDB) is a database management system that primarily relies on main memory, rather than disk, for data storage. In-memory databases are faster than disk-optimized databases, an important consideration for big data analytics uses and the creation of data warehouses and data marts.

Big data skills

Big data and big data analytics endeavors require specific skills, whether they come from inside the organization or through outside experts.

Many of these skills are related to the key big data technology components, such as Hadoop, Spark, NoSQL databases, in-memory databases, and analytics software.

Others are specific to disciplines such as data science, data mining, statistical and quantitative analysis, data visualization, general-purpose programming, and data structure and algorithms. There is also a need for people with overall management skills to see big data projects through to completion.

Given how common big data analytics projects have become and the shortage of people with these types of skills, finding experienced professionals might be one of the biggest challenges for organizations.

Big data analytics use cases

Big data and analytics can be applied to many business problems and use cases. Here are a few examples:

- Customer analytics. Companies can examine customer data to enhance customer experience, improve conversion rates, and increase retention.

- Operational analytics. Improving operational performance and making better use of corporate assets are the goals of many companies. Big data analytics tools can help businesses find ways to operate more efficiently and improve performance.
- Fraud prevention. Big data tools and analysis can help organizations identify suspicious activity and patterns that might indicate fraudulent behavior and help mitigate risks.
- Price optimization. Companies can use big data analytics to optimize the prices they charge for products and services, helping to boost revenue.

What is Big Data Analytics?

A Definition of Big Data Analytics

Big Data Analytics is "the process of examining large data sets containing a variety of data types – i.e., Big Data – to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other

useful information.” Companies and enterprises that implement Big Data Analytics often reap several business benefits, including more effective marketing campaigns, the discovery of new revenue opportunities, improved customer service delivery, more efficient operations, and competitive advantages. Companies implement Big Data Analytics because they want to make more informed business decisions. Big Data Analytics gives analytics professionals, such as data scientists and predictive modelers, the ability to analyze Big Data from multiple and varied sources, including transactional data and other structured data.

Types of Big Data Analytics Tools

Big Data Analytics tools are important for companies and enterprises because of the sheer volume of Big Data now generated and managed by modern organizations. Big Data Analytics tools also help businesses save time and money and aid in gaining insights to inform data-driven decisions. There are various types of tools that may fall under the umbrella of Big

Data Analytics or serve to improve the process of analyzing data: data storage and management, data cleaning, data mining, data analysis, data visualization, data integration, and data collection.

Benefits of Big Data Analysis for Modern Enterprises

Big Data Analytics enables enterprises to analyze their data in full context quickly, and some offer real-time analysis. With high-performance data mining, predictive analytics, text mining, forecasting, and optimization, enterprises that utilize Big Data Analytics are able to drive innovation and make the best business decisions. Companies that take advantage of all that Big Data Analytics solutions have to offer are better positioned to optimize machine

learning and address their Big Data needs in groundbreaking ways.

Specifically, Big Data Analytics enables enterprises to narrow their Big Data to the most relevant information and analyze it to inform critical business decisions. This proactive approach to business is transformative because it gives analysts and decision makers the power to move ahead with the best knowledge and insights available, often in real time. This means that companies can improve their customer retention, develop better products, and gain a competitive advantage by taking rapid action to respond to market changes, indications of critical customer shifts, and other metrics that impact business.

Enterprises utilizing Big Data Analytics with fidelity also have the ability to boost sales and marketing results, discover new revenue opportunities, improve customer service, optimize operational efficiency, reduce risk, and drive other business results.

Trends in Big Data Analytics

Big Data Analytics is changing the way the world does business, which also means that it is changing technology and business practices. Robert L. Mitchell, Computerworld contributor and chief editor of TechBeacon.com, explains that Big Data technologies and practices move quickly and states that "top emerging technologies and trends should be on your watch list." Mitchell compiled a list of the hottest trends in Big Data Analytics, based on input from IT leaders, consultants, and industry analysts:

1. Big Data Analytics in the cloud
2. Hadoop: The new enterprise data operating system
3. Big Data lakes
4. More predictive analytics
5. SQL on Hadoop: Faster, better
6. More, better NoSQL
7. Deep learning

8. In-memory analytics

Mitchell explains that these trends are emerging so quickly that IT organizations must “create conditions that will allow analysts and data scientists to experiment.” Collaboration between IT and analysts should occur in order to meet the needs of organizations and deliver the resources necessary for Big Data Analytics.

The increasingly widespread use of Big Data Analysis solutions is a clear indication that Big Data is not just a fad: it’s a business practice that is here to stay because of the insights it delivers to enterprises that want to gain a competitive edge, improve sales and marketing team performance, increase revenue, and make proactive data-driven business decisions.

Big Data Analytics: Defined By Value

Recently, I talked about defining big data, and how there seems to be some confusion in the market around what, exactly, constitutes this term that we're seeing everywhere. But big data is just the tip of the iceberg. Big data *analytics* have taken on all sorts of forms and functions, as well. Whether it's charts and graphs that show images of the business or prescriptive analytics that provide recommendations in real time, there's an abundance of definitions for big data analytics, too.

One of the articles I linked to in my post about defining big data was an interview with Ventana Research's Tony Cosentino (@TonyCosentinoVR). In it, he talks about the "Three V's" of big data. He says that, "…if we think about big data analytics in terms of the Vs (volume, velocity and variety), the research shows that organizations prioritize these by variety, volume and then velocity." He goes on to say that thinking of analytics this way tends to turn big data into an ethos instead of something tangible that delivers value to an organization.

I couldn't agree more.

So when Mr. Cosentino talks about prioritizing the Three V's of velocity, volume and variety, I actually think that's not the full picture. The only way for analytics to deliver real, tangible value is to combine all three of them. You've got to take large data sets (volume) from disparate sources (variety) coming at you at the speed of light (velocity) and provide recommendations in real time. Removing any of them from the equation is like

chopping off the leg of a barstool – it leaves you without support.

The industry is moving in the right direction – but until the default definition of big data analytics centers on the value it brings to the organization, we've still got some work to do.

Machine learning and Big Data analytics: the perfect marriage

While 2012 has been the year of Big Data technologies, 2013 is becoming the year of Big Data analytics. Gathering and maintaining large collections of data is one thing, but extracting useful information from these collections is even more challenging. Big Data not only changes the tools one can use for predictive analytics, it also changes our entire way of

thinking about knowledge extraction and interpretation. Traditionally, data science has always been dominated by trial-and-error analysis, an approach that becomes impossible when datasets are large and heterogeneous. Ironically, availability of more data usually leads to fewer options in constructing predictive models, because very few tools allow for processing large datasets in a reasonable amount of time. In addition, traditional statistical solutions typically focus on static analytics that is limited to the analysis of samples that are frozen in time, which often results in surpassed and unreliable conclusions.

However, more clever alternatives that overcome those problems have been recently proposed in a novel and rapidly expanding research domain: machine learning. At the edge of statistics, computer science and emerging applications in industry, this research community focuses on the development of fast and efficient algorithms for real-time processing of data with as a main goal to deliver accurate predictions of various kinds. To name only a few applications, think of business cases such as product recommendation, segmentation of customers,

fraud detection or churn prevention. Machine learning techniques can solve such applications using a set of generic methods that differ from more traditional statistical techniques. The emphasis is on real-time and highly scalable predictive analytics, using fully automatic and generic methods that simplify some of the typical data scientist tasks. And yes, machine learning is finding its way to industry at this moment!

Traditional methods in machine learning and statistics provide data-driven models for predicting one-dimensional targets, such as binary outputs in classification and real-valued outputs in regression. In recent years, novel application domains have triggered fundamental research on more complicated problems where multi-target predictions are required. Such problems arise in diverse application domains, such as document categorization, recommender systems, tag prediction of images, videos and music, information retrieval, natural language processing, drug discovery, marketing, biology, etc. Specific

multi-target prediction problems have been studied in a variety of subfields of machine learning and statistics, such as multi-label classification, multivariate regression, sequence learning, structured output prediction, preference learning, multi-task learning, recommender systems and collective learning. Despite their commonalities, work on solving problems in the above domains has typically been performed in isolation, without much interaction between the different sub-communities. The main goal of the tutorial is to present a unifying overview of the above-mentioned subfields of machine learning, by focusing on the simultaneous prediction of multiple, mutually dependent output variables.

Three key ideas we like from Data Mining Techniques in CRM to Improve Data Quality Management:

- Anomaly detection is a key data mining technique because anomalies can provide actionable information, since they deviate from the data set’s average

- Clustering is important for identifying similar data sets and understanding the similarities and differences within data
- Regression analysis is an advanced data mining technique that determines customer satisfaction levels and how they affect customer loyalty

As Big Data takes center stage for business operations, data mining becomes something that salespeople, marketers, and C-level executives need to know how to do and do well. Generally, data mining is the process of finding patterns and correlations in large data sets to predict outcomes. There are a variety of techniques to use for data mining, but at its core are statistics, artificial intelligence, and machine learning. Companies and organizations are using data mining to get the insights they need about pricing, promotions, social media, campaigns, customer experience, and a plethora of other business practices.

5 Skills You Need To Know To Become a Big Data Analyst:

Traditional data analysis fails to cope with the advent of Big Data which is essentially huge data, both structured and unstructured. Much more is needed that being able to navigate on relational database management systems and draw insights using statistical algorithms.

The good news is that the analytics part remains the same whether you are dealing with small datasets, large datasets or even unstructured datasets. What is needed the most in big data is the ability to draw relevant information from the humungous amounts of data being processed every minute. This requires technology to join hands with traditional analytics.

Let us now look at some of the key skills needed for being a big data analyst –

1) Programming

While traditional data analyst might be able to get away without being a full-fledged programmer, a big data analyst needs to be very comfortable with coding. One of the main reasons for this requirement is that big data is still in an evolution phase. Not many standard processes are set around the large complex datasets a big data analyst has to deal with. A lot of customization is required on daily basis to deal with the unstructured data.
Which languages are required – R, Python, Java, C++, Ruby, SQL, Hive, SAS, SPSS, MATLAB, Weka, Julia, Scala. As you can not knowing a language should not be a barrier for a big data scientist. At the minimum one needs to know R, Python, and Java. While working you may end up using various tools. Programming Language is only a tool and more tools you have in your kitty, merrier it is.

2) Data Warehousing

Experience with relational and non -relational database systems is a must. Examples of non- relational database include – Mysql, Oracle, DB2. Examples of non-relational database include – NoSql : Hbase, HDFS, MongoDB, CouchDB, Cassandra, Teradeta, etc.

3) Computational frameworks

A good understanding and familiarity with frameworks such as Apache Spark, Apache Storm, Apache Samza, Apache Flink and the classic MapReduce and Hadoop. These technologies help in Big Data processing which can be streamed to a great extent.

4) Quantitative Aptitude and Statistics

While the processing of Big Data requires great use of technology, fundamental to any analysis of data is good knowledge of Statistics and linear algebra. Statistics is a basic building block of data science and understanding of core concepts like summary statistics, probability distribution, random variables, Hypothesis testing framework is important if you are data scientist of any genre.

5) Business Knowledge

To keep the analysis focused, to validate, sort, relate, evaluate the data, the most critical skill of a big data scientist is to have a good knowledge of the domain one is working on. In fact, the reason big data analysts are so much in demand is that its very rare to find resources who have a thorough understanding of technical aspects, statistics and business. There are analysts good in business and statistics but not in programming. There are expert programmers without the knowhow of how to put the programs in the context of the business goal.

To keep the analysis focused, to validate, sort, relate, evaluate the data, the most critical skill of a big data scientist is to have a good knowledge of the domain one is working on. In fact, the reason big data analysts are so much in demand is that it's very rare to find resources that have a thorough understanding of technical aspects, statistics and business.

There are analysts good in business and statistics but not in programming. There are expert programmers without the knowhow of how to put the programs in the context of the business goal.

Lastly, a good hold on machine learning is highly beneficial as it helps in managing complex data structures and learning patterns that are too difficult to handle using traditional data analytics

11 Best Big Data Analytics Tools in 2019

Big Data Analytics software is widely used in providing meaningful analysis of a large set of data. This software helps in finding current market trends, customer preferences, and other information.

Here are the 11 Top Big Data Analytics Tools with key feature and download links.

1) Xplenty

Xplenty is a cloud-based ETL solution providing simple visualized data pipelines for automated data flows across a wide range of sources and destinations. Xplenty's powerful on-platform transformation tools allow you to clean, normalize, and transform data while also adhering to compliance best practices.

Features:

- Powerful, code-free, on-platform data transformation offering
- Rest API connector - pull in data from any source that has a Rest API

- Destination flexibility - send data to databases, data warehouses, and Salesforce
- Security focused - field-level data encryption and masking to meet compliance requirements
- Rest API - achieve anything possible on the Xplenty UI via the Xplenty API
- Customer-centric company that leads with first-class support

---

2) Microsoft HDInsight:

Azure HDInsight is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads.

Features:

- Reliable analytics with an industry-leading SLA
- It offers enterprise-grade security and monitoring
- Protect data assets and extend on-premises security and governance controls to the cloud

- High-productivity platform for developers and scientists
- Integration with leading productivity applications
- Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs

Download link: https://azure.microsoft.com/en-in/free/

---

3) Skytree:

Skytree is a big data analytics tool that empowers data scientists to build more accurate models faster. It offers accurate predictive machine learning models that are easy to use.

Features:

- Highly Scalable Algorithms
- Artificial Intelligence for Data Scientists
- It allows data scientists to visualize and understand the logic behind ML decisions
- Skytree via the easy-to-adopt GUI or programmatically in Java
- Model Interpretability
- It is designed to solve robust predictive problems with data preparation capabilities
- Programmatic and GUI Access

Download link: http://www.skytree.net/

---

4) Talend:

Talend is a big data tool that simplifies and automates big data integration. Its graphical wizard generates native code. It also allows big data integration, master data management and checks data quality.

Features:

- Accelerate time to value for big data projects
- Simplify ETL & ELT for big data
- Talend Big Data Platform simplifies using MapReduce and Spark by generating native code
- Smarter data quality with machine learning and natural language processing
- Agile DevOps to speed up big data projects
- Streamline all the DevOps processes

Download Link: https://www.talend.com/download/

---

5) Splice Machine:

Splice Machine is a big data analytic tool. Their architecture is portable across public clouds such as AWS, Azure, and Google.

Features:

- It can dynamically scale from a few to thousands of nodes to enable applications at every scale
- The Splice Machine optimizer automatically evaluates every query to the distributed HBase regions
- Reduce management, deploy faster, and reduce risk
- Consume fast streaming data, develop, test and deploy machine learning models

Download link: https://www.splicemachine.com

---

6) Spark:

Apache Spark is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

Features:

- It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk
- It offers lighting Fast Processing
- Support for Sophisticated Analytics
- Ability to Integrate with Hadoop and Existing Hadoop Data
- It provides built-in APIs in Java, Scala, or Python

Download link: https://spark.apache.org/downloads.html

---

7) Plotly:

Plotly is an analytics tool that lets users create charts and dashboards to share online.

Features:

- Easily turn any data into eye-catching and informative graphics

- It provides audited industries with fine-grained information on data provenance
- Plotly offers unlimited public file hosting through its free community plan

Download link: https://plot.ly/

---

8) Apache SAMOA:

Apache SAMOA is a big data analytics tool. It enables development of new ML algorithms. It provides a collection of distributed algorithms for common data mining and machine learning tasks.

Download link: https://samoa.incubator.apache.org/

---

9) Lumify:

Lumify is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

Features:

- It provides both 2D and 3D graph visualizations with a variety of automatic layouts
- It provides a variety of options for analyzing the links between entities on the graph
- It comes with specific ingest processing and interface elements for textual content, images, and videos
- It spaces feature allows you to organize work into a set of projects, or workspaces
- It is built on proven, scalable big data technologies

Download link: http://www.altamiracorp.com/index.php/lumify/

---

10) Elasticsearch:

Elasticsearch is a JSON-based Big data search and analytics engine. It is a distributed, RESTful search and analytics engine for solving numbers of use cases. It offers horizontal scalability, maximum reliability, and easy management.

Features:

- It allows combine many types of searches such as structured, unstructured, geo, metric, etc

- Intuitive APIs for monitoring and management give complete visibility and control
- It uses standard RESTful APIs and JSON. It also builds and maintains clients in many languages like Java, Python, NET, and Groovy
- Real-time search and analytics features to work big data by using the Elasticsearch-Hadoop
- It gives an enhanced experience with security, monitoring, reporting, and machine learning features

Download link: https://www.elastic.co/downloads/elasticsearch

---

11) R-Programming:

R is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

- Effective data handling and storage facility,
- It provides a suite of operators for calculations on arrays, in particular, matrices,

- It provides coherent, integrated collection of big data tools for data analysis
- It provides graphical facilities for data analysis which display either on-screen or on hardcopy

Download link: https://www.r-project.org/

---

12) IBM SPSS Modeler:

IBM SPSS Modeler is a predictive big data analytics platform. It offers predictive models and delivers to individuals, groups, systems and the enterprise. It has a range of advanced algorithms and analysis techniques.

Features:

- Discover insights and solve problems faster by analyzing structured and unstructured data
- Use an intuitive interface for everyone to learn
- You can select from on-premises, cloud and hybrid deployment options
- Quickly choose the best performing algorithm based on model performance

Download link: https://www.ibm.com/us-en/marketplace/spss-modeler/purchase#product-header-top

Big Data Analytics – Turning Insights into Action

Just like the entire universe and our galaxy is said to have formed due to the Big Bang explosion, similarly, due to so many technological advancements, data has also been growing exponentially leading to the Big Data explosion. This data comes in from various sources, has different formats, is generated at a variable rate and may also contain inconsistencies. Thus, we can simply term the explosion of such data as ***Big Data***. I will be explaining the following topics in this blog to give you insights on Big Data Analytics :

- Why Big Data Analytics?
- What is Big Data Analytics?
- Big Data Tools
- Big Data Domains
- Big Data Use Cases
- Trends in Big Data Analytics

Why Big Data Analytics?

Before I jump on to tell you about what is ***Big Data*** Analytics, let me tell you guys about why it is needed. Let me also reveal to you guys that we create about 2.5 quintillion bytes of data every day! So now that we have accumulated Big Data, neither can we ignore it nor can we let it stay idle and make it go to waste.

Various organizations and sectors all across the globe started adopting Big Data Analytics in order to gain numerous benefits. Big Data Analytics gives insights which many companies are turning into actions and making huge profits as well as discoveries. I am going to list down four such reasons along with interesting examples.

The first reason is,

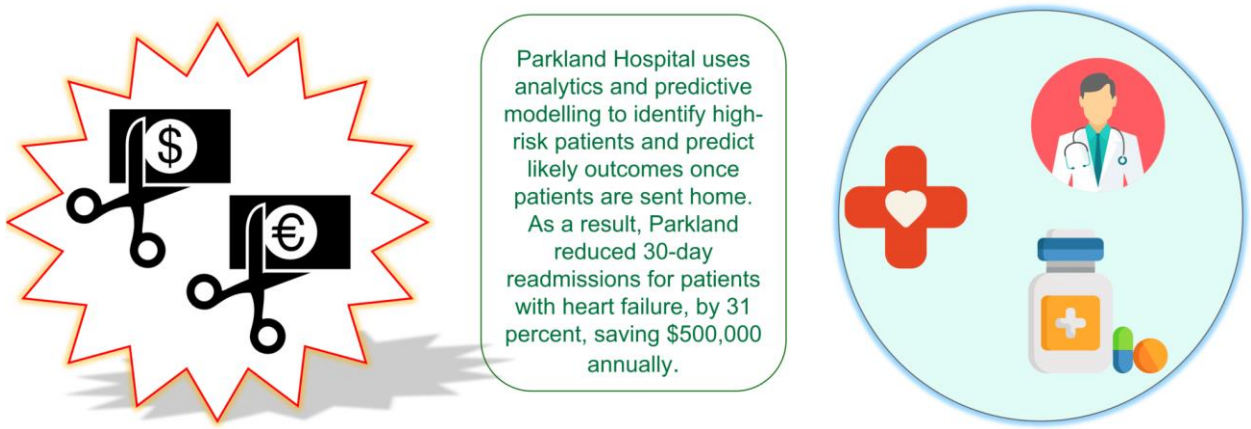
1. **Making Smarter and More Efficient Organization**.

Let me tell you about one such organisation, the New York Police Department (NYPD). The NYPD brilliantly uses Big Data and analytics to detect and identify crimes before they occur. They analyse historical arrest patterns and then maps them with events such as federal holidays, paydays, traffic flows, rainfall etc. This aids them in analyzing the information immediately by utilizing these data patterns. Big Data and analytics strategy helps them identify crime locations, through which they deploy their officers to these locations. Thus by reaching these locations before the crimes were committed, they prevent the occurrence of crime.

2. Optimize Business Operations by Analyzing Customer Behavior.

Most organizations use behavioral analytics of customers in order to provide customer satisfaction and hence, increase their customer base. The best example of this is Amazon. Amazon is one of the best and most widely used e-commerce websites with a customer base of about 300 million. They use customer click-stream data and historical purchase data to provide them with customized results on customized web pages. Analyzing the clicks of every visitor on their website aids them in understanding their site-navigation behavior, paths the user took to buy the product, paths that led them to leave the site and more. All this information helps Amazon to improve their user experience, thereby improving their sales and marketing.

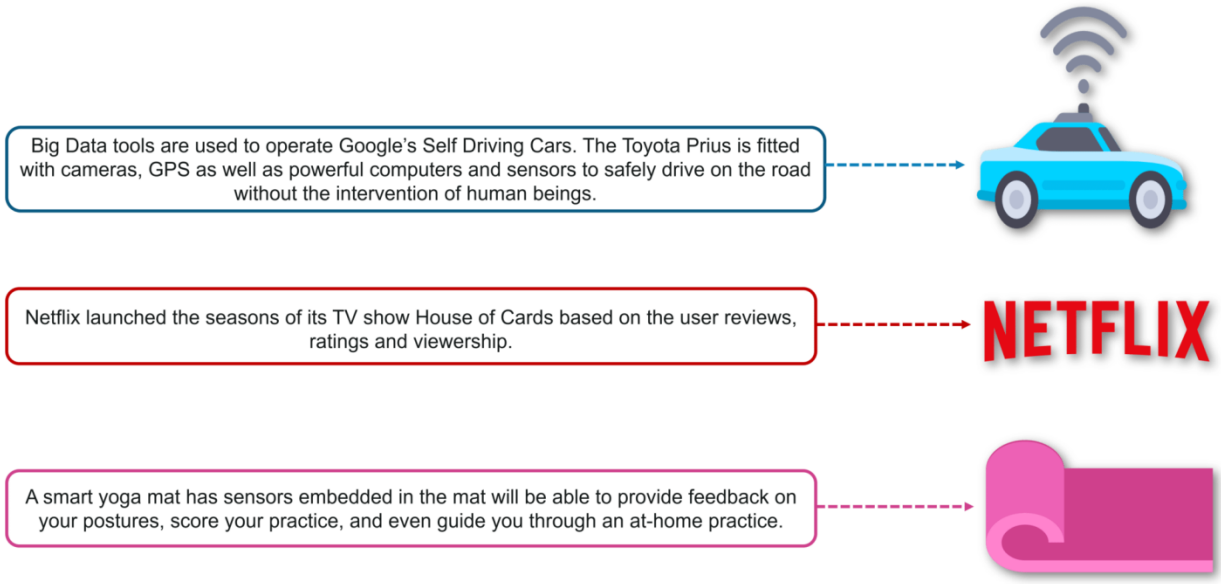
3. Cost Reduction



Big data technologies and technological advancements like cloud computing bring significant cost advantages when it comes to store and process Big Data. Let me tell you how healthcare utilizes Big Data Analytics to reduce their costs. Patients nowadays are using new sensor devices when at home or outside, which send constant streams of data that can be monitored and analyzed in real-time to help patients avoid hospitalization by self-managing their conditions. For hospitalized patients, physicians can use predictive analytics to optimize outcomes and reduce readmissions. Parkland Hospital uses analytics and predictive modeling to identify high-risk patients and predict likely outcomes once patients are sent home. As a result, Parkland reduced 30-day readmissions for patients with heart failure, by 31%, saving $500,000 annually.

New Generation Products

With the ability to gauge customer needs and satisfaction through analytics, comes the power to give customers what they want. I have found there such interesting products to cite here.



First, Google's self-driving car which makes millions of calculations on every trip that help the car decide when and where to turn, whether to slow down or speed up and when to change lanes — the same decisions a human driver makes behind the wheel.

The **second** one is Netflix which committed for two seasons of its extremely popular show House of Cards, by completely trusting Big Data Analytics! Last year, Netflix grew its US subscriber base by 10% and added nearly 20 million subscribers from around the globe.

The **third** example is one of the really cool new things I have come across, is a smart yoga mat. The first time you use your Smart Mat, it will take you through a series of movements to calibrate your body shape, size and personal limitations. This personal profile information is stored in your Smart Mat App and will help Smart Mat detect when you're out of alignment or balance. Over time, it will automatically evolve with updated data as you improve your Yoga practice.

What is Big Data Analytics?

Now let us formally define “What is Big Data Analytics?” Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights. Basically, Big Data Analytics is largely used by companies to facilitate their growth and development. This majorly involves applying various data mining algorithms on the given set of data, which will then aid them in better decision making.

Stages in Big Data Analytics

These are the following stages involved in the Big Data Analytics process:

Types of Big Data Analytics

There are four types:

1. **Descriptive Analytics:** It uses data aggregation and data mining to provide insight into the past and answer: “What has happened?” The descriptive analytics does exactly what the name implies they “describe” or summarize raw data and

make it interpretable by humans.

2. **Predictive Analytics:** It uses statistical models and forecasts techniques to understand the future and answer: “What could happen?” Predictive analytics provides companies with actionable insights based on data. It provides estimates about the likelihood of a future outcome.

What might happen in the future

For example, Southwest Airlines analyses sensor data on their planes in order to identify patterns that indicate a potential malfunction, thus allowing the airlines to the necessary repairs before its schedule.

3. **Prescriptive Analytics:** It uses optimization and simulation algorithms to advice on possible outcomes and answers: “What should we do?” It allows users to “prescribe” a number of different possible actions and guide them towards a solution. In a nutshell, this analytics is all about providing advice.

What action should be taken.

Google's self-driving car is a perfect example of prescriptive analytics. It analyses the environment and decides the direction to take based on data.

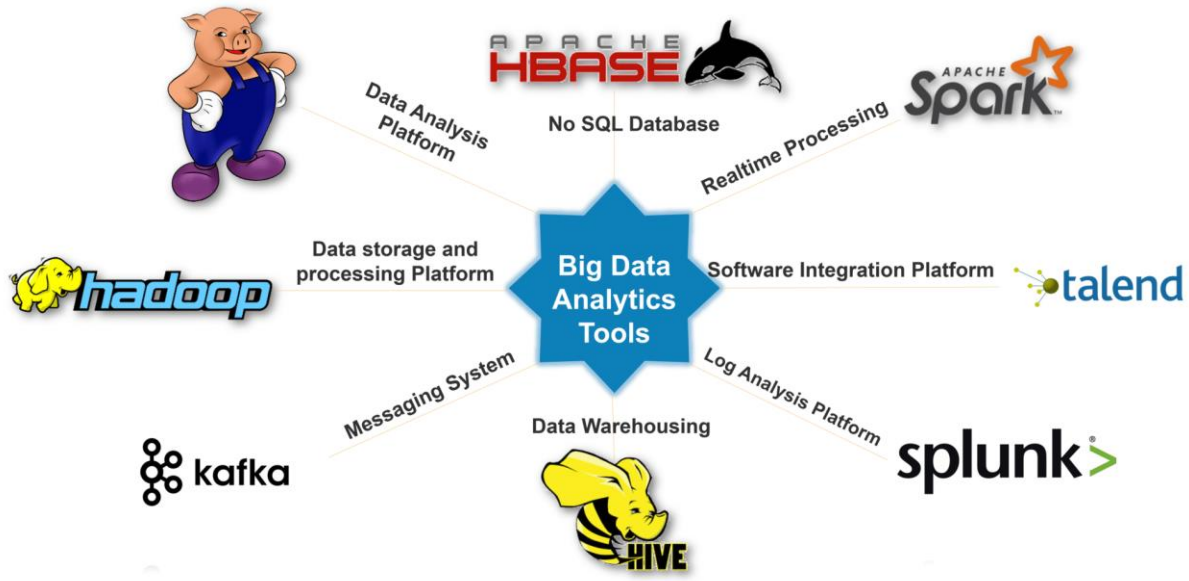
4. **Diagnostic Analytics:** It is used to determine why something happened in the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations. Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

Why did it happen

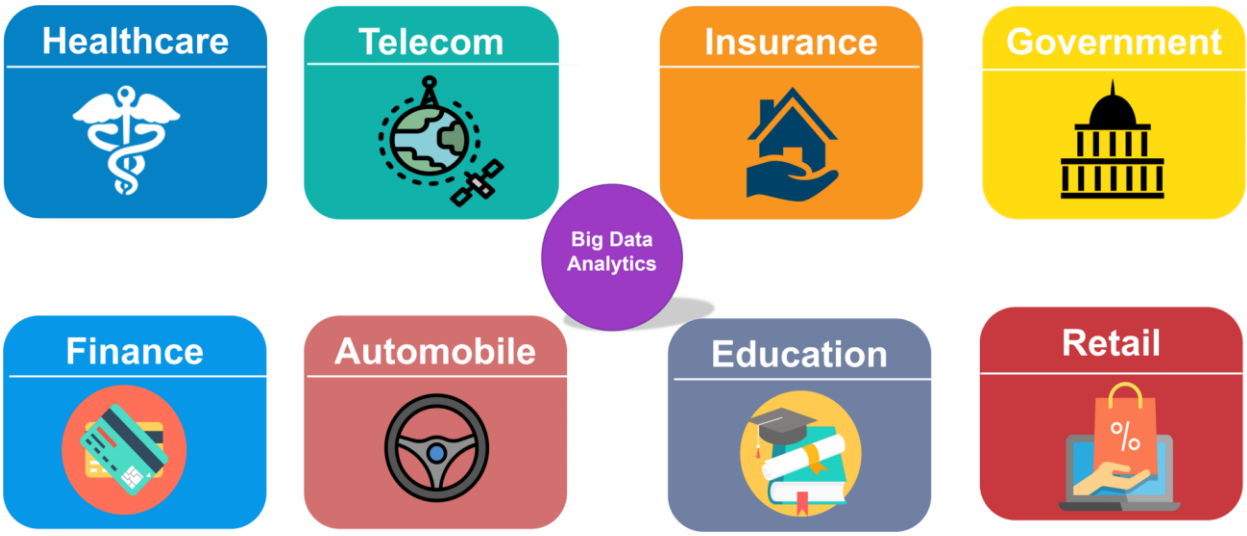
For a Social Media marketing campaign, you can use diagnostic analytics to assess the number of posts, mentions, followers, fans, page views, reviews, pins, etc. and analyse the failure and success rate of the campaign at a fundamental level.

Big Data Tools

These are some of the following tools used for Big Data Analytics: ***Hadoop***, ***Pig***, ***Apache HBase***, ***Apache Spark***, ***Talend***, ***Splunk***, ***Apache Hive***, ***Kafka***.



Big Data Domains

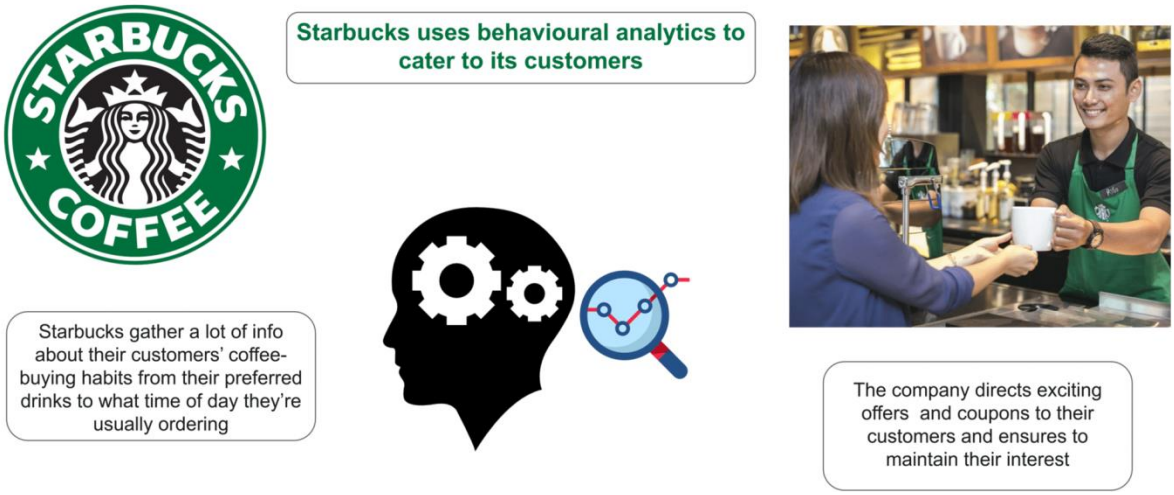


- **Healthcare:** Healthcare is using big data analytics to reduce costs, predict epidemics, avoid preventable diseases and improve the quality of life in general. One of the most widespread applications of Big Data in healthcare is Electronic Health Record (EHRs).

- **Telecom:** They are one of the most significant contributors to Big Data. Telecom industry improves the quality of service and routes traffic more effectively. By analyzing call data records in real-time, these companies can identify fraudulent behavior and act on them immediately. The marketing division can modify its campaigns to better target its customers and use insights gained to develop new products and services.
- **Insurance:** These companies use big data analytics for risk assessment, fraud detection, marketing, customer insights, customer experience and more.
- **Government:** The Indian government used big data analytics to get an estimate of trade in the country. They used Central sales tax invoices to analyze the extent to which states trade with each other.
- **Finance:** Banks and financial services firms use analytics to differentiate fraudulent interactions from legitimate business transactions. The analytics systems suggest immediate actions, such as blocking irregular transactions, which stops fraud before it occurs and improves profitability.
- **Automobile:** Rolls Royce which has embraced Big Data by fitting hundreds of sensors into its engines and propulsion systems, which record every tiny detail about their operation. The changes in data in real-time are reported to engineers who will decide the best course of action such as scheduling maintenance or dispatching engineering teams.
- **Education:** This is one field where Big Data Analytics is being absorbed slowly and gradually. Opting for big data powered technology as a learning tool instead of traditional lecture methods, enhanced the learning of students as well as aided the teachers to track their performance better.
- **Retail:** Retail including e-commerce and in-stores are widely using Big Data Analytics to optimize their business. For example, Amazon, Wal-Mart etc.

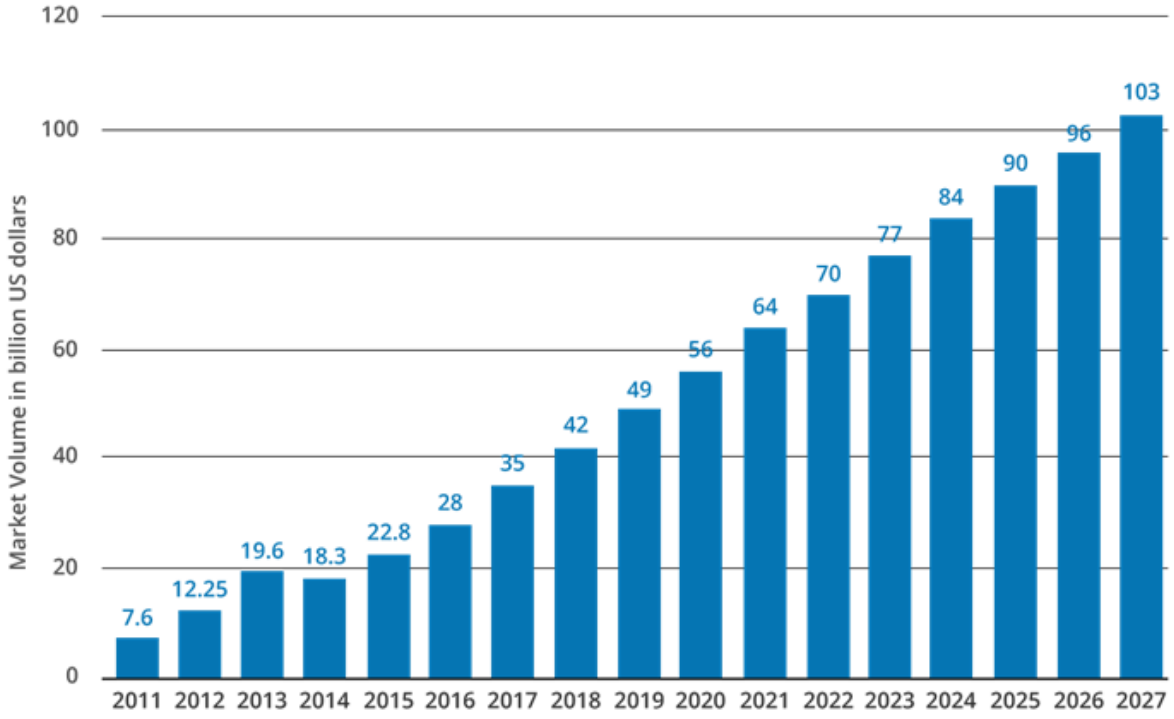
Big Data Use Cases

The first use case that I have taken here is of Starbucks.

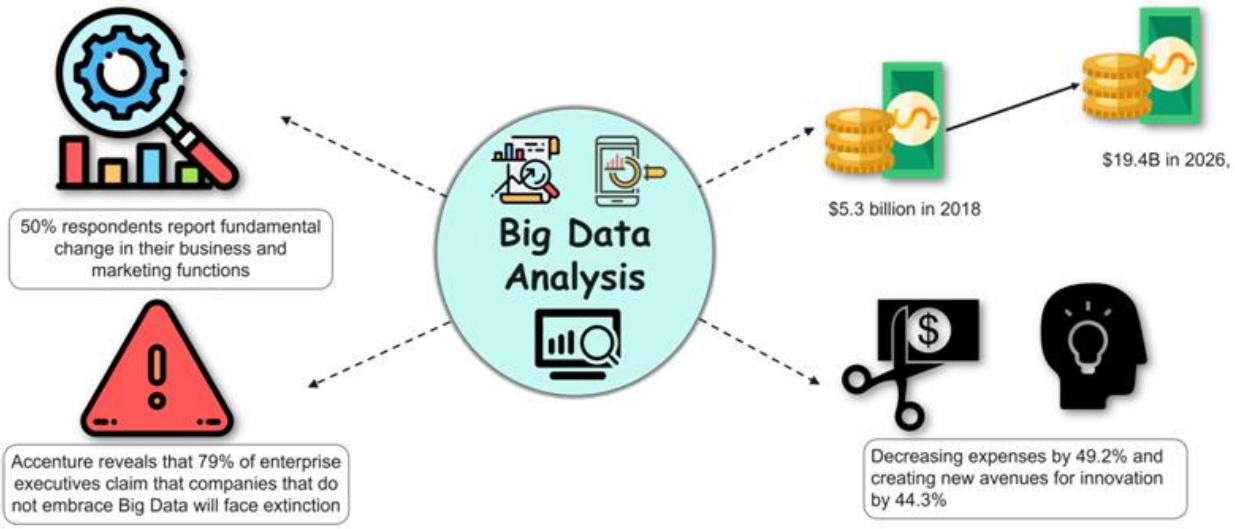


The second use case I want to share with you guys is of Procter&Gamble.

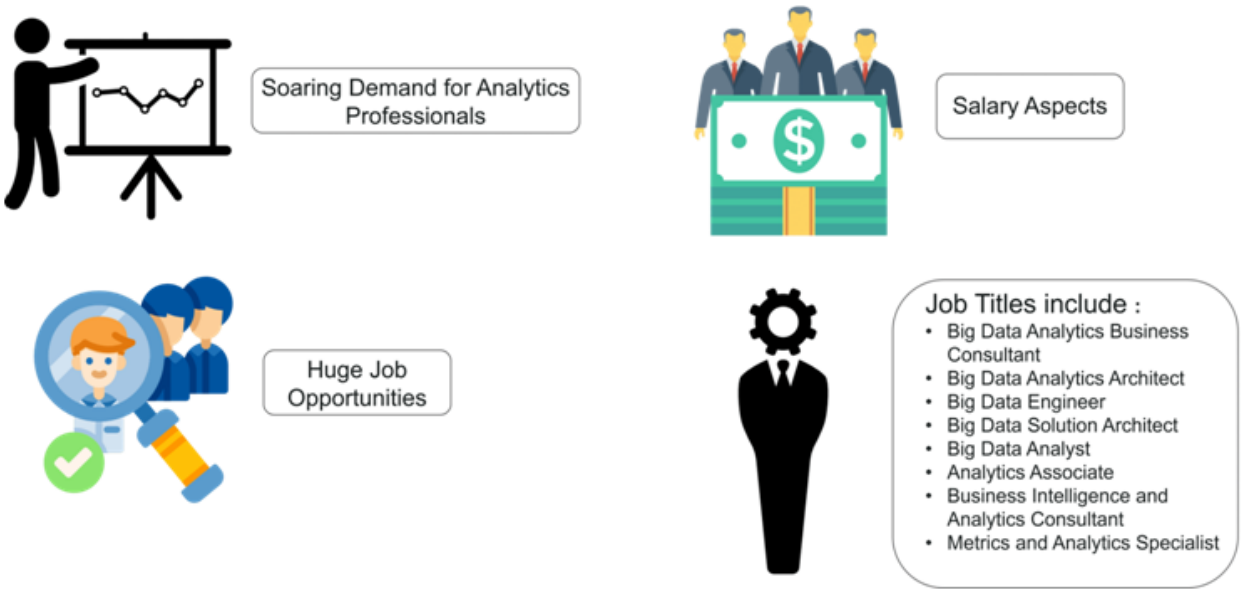
The image below depicts the **market revenue of Big Data** in billion U.S. dollars from the year 2011 to 2027.



Here is some **Facts and Statistics by Forbes**:



Career prospects in Big Data Analytics:



- **Salary Aspects:** The average salary of the analytics jobs is around $94,167. Data Scientist has been named the best job in America for three years running, with a median base salary of $110,000 and 4,524 job openings. In India the percentage of analytics professionals commanding salaries less than INR 10 Lakhs has gone lower; percentage of analytics professionals earning more than INR 15 Lakhs has increased from **17%** in 2016 to **21%** in 2017 to **22.3%** in 2018.
- **Huge Job Opportunities:** Companies like Google, Apple, IBM, Adobe, Qualcomm and many more hire Big Data Analytics Professionals.

Skill set

These are some of the skills which are required depending upon the role in the field of Big Data Analytics:

- **Basic programming:** One should have knowledge about at least some general purpose programming language such as Java and Python.
- **Statistical and Quantitative Analysis:** Having an idea about statistics and quantitative analysis is ideal.
- **Data Warehousing:** Knowledge of SQL and NoSQL databases is required.
- **Data Visualization:** It is very important to know how to visualize the data in order to be able to understand the insights and apply it in action.
- **Specific Business Knowledge:** One must necessarily be aware of the business where they are applying analytics in order to optimize their operations.

- **Computational Frameworks:** Preferably one should know about at least one or two tools which are required for Big Data Analytics.

What Is Big Data Analytics and why do I Need It for My Business?

Big data is primarily defined by the volume of a data set. Big data sets are generally huge – measuring tens of terabytes – and sometimes crossing the threshold of petabytes. The term big data was preceded by very large databases (VLDBs) which were managed using database management systems (DBMS). Today, big data falls under three categories of data sets – structured, unstructured and semi-structured.

Structured data sets comprise of data which can be used in its original form to derive results. Examples include relational data such as employee salary records. Most modern computers and applications are programmed to generate structured data in preset formats to make it easier to process.

Unstructured data sets, on the other hand, are without proper formatting and alignment. Examples include human texts, Google search result outputs, etc. These random collections of data sets require more processing power and time for conversion into structured data sets so that they can help in deriving tangible results.Semi-Structured data sets are a combination of both structured and unstructured data. These data sets might have a proper structure and yet lack defining elements for sorting and processing. Examples include RFID and XML data.

Semi-Structured data sets are a combination of both structured and unstructured data. These data sets might have a proper structure and yet lack defining elements for sorting and processing. Examples include RFID and XML data.

Big data processing requires a particular setup of physical and virtual machines to derive results. The processing is done simultaneously to achieve results as quickly as possible. These days big data processing techniques also include Cloud Computing and Artificial Intelligence. These technologies help in

reducing manual inputs and oversight by automating many processes and tasks.

The evolving nature of big data has made it difficult to give it a commonly accepted definition. Data sets are consigned the big data status based on technologies and tools required for their processing.

BIG DATA ANALYTICS – TECHNOLOGIES AND TOOLS

Big data analytics is the process of extracting useful information by analyzing different types of big data sets. Big data analytics is used to discover hidden patterns, market trends and consumer preferences, for the benefit of organizational decision making. There are several steps and technologies involved in big data analytics.

Data Acquisition

Data acquisition has two components: identification and collection of big data. Identification of big data is done by analyzing the two natural formats of data – born digital and born analogue.

Born Digital Data

It is the information which has been captured through a digital medium, e.g. a computer or Smartphone app, etc. This type of data has an ever expanding range since systems keep on collecting different kinds of information from users. Born digital data is traceable and can provide both personal and demographic business insights. Examples include Cookies, Web Analytics and GPS tracking.

Born Analogue Data

When information is in the form of pictures, videos and other such formats which relate to physical elements of our world, it is termed as analogue data. This data requires conversion into digital format by using sensors, such as cameras, voice recording, digital assistants, etc. The increasing reach of technology has also raised the rate at which traditionally analogue data is being converted or captured through digital mediums.

The second step in the data acquisition process is collection and storage of data sets identified as big data. Since the archaic DBMS

techniques were inadequate for managing big data, a new method is used for collecting and storing big data. The process is called MAD – magnetic, agile and deep. Since, managing big data requires a significant amount of processing and storage capacity, creating such systems is out-of-reach for most entities which rely on big data analytics. Thus, the most common solutions for big data processing today are based on two principles – distributed storage and Massive Parallel processing a.k.a. MPP. Most of the high-end Hadoop platforms and specialty appliances use MPP configurations in their system.

Non-relational Databases

The databases that store these massive data sets have also evolved in how and where the data is stored. JavaScript Object Notation or JSON is the preferred protocol for saving big data nowadays. Using JSON, the tasks can be written in the application layer and allow better cross-platform functionalities. Thus enabling, agile development of scalable and flexible data solutions for the devs. Many companies are using it as a replacement of XML as a

way of transmitting structured data between the server and web application.

In-memory Database Systems

These database storage systems are designed to overcome one of the major hurdles in the way of big data processing – the time taken by traditional databases to access and process information. IMDB systems store the data in the RAM of big data servers, therefore, drastically reducing the storage I/O gap. Apache Spark is an example of IMDB systems. VoltDB, NuoDB and IBM solidDB are some more examples of the same.

Hybrid Data Storage and Processing Systems – Apache Hadoop

Apache Hadoop is hybrid data storage and processing system which provides scalability and speed at reasonable costs for mid and small-scale businesses. It uses a Hadoop Distributed File System (HDFS) for storing large files across multiple systems known as cluster nodes. Hadoop has a replication mechanism to ensure smooth operation even during instances of individual node

failures. Hadoop uses Google's MapReduce parallel programming as its core. The name originates from 'Mapping' and 'Reduction' of functional programming languages in its algorithm for big data processing. MapReduce works on the premise of increasing the number of functional nodes over increasing processing power of individual nodes. Moreover, Hadoop can be run using readily available hardware which has sped up its development and popularity, significantly.

Data Mining

It is a recent concept which is based on contextual analyzing of big data sets to discover the relationship between separate data items. The objective is to use a single data set for different purposes by different users. Data mining can be used for reducing costs and increasing revenues.

TOP 10 SECTORS USING BIG DATA ANALYTICS

Big data is finding usage in almost all industries today. Here is a list of the top segments using big data to give you an idea of its application and scope.

1. **Banking and Securities**: For monitoring financial markets through network activity monitors and natural language

processors to reduce fraudulent transactions. Exchange Commissions or Trading Commissions are using big data analytics to ensure that no illegal trading happens by monitoring the stock market.

2. **Communications and Media**: For real-time reportage of events around the globe on several platforms (mobile, web and TV), simultaneously. Music industry, a segment of media, is using big data to keep an eye on the latest trends which are ultimately used by auto tuning softwares to generate catchy tunes.

3. **Sports:** To understand the patterns of viewership of different events in specific regions and also monitor the performance of individual players and teams by analysis. Sporting events like Cricket world cup, FIFA world cup and Wimbledon make special use of big data analytics.

4. **Healthcare**: To collect public health data for faster responses to individual health problems and identify the global spread of new virus strains such as Ebola. Health Ministries of different countries

incorporate big data analytic tools to make proper use of data collected after Census and surveys.

5. **Education**: To update and upgrade prescribed literature for a variety of fields which are witnessing rapid development. Universities across the world are using it to monitor and track the performance of their students and faculties and map the interest of students in different subjects via attendance.

6. **Manufacturing**: To increase productivity by using big data to enhance supply chain management. Manufacturing companies use these analytical tools to ensure that are allocating the resources of production in an optimum manner which yields the maximum benefit.

7. **Insurance**: For everything from developing new products to handling claims through predictive analytics. Insurance companies use business big data to keep a track of the scheme of policy which is the most in demand and is generating the most revenue.

8. **Consumer Trade**: To predict and manage staffing and inventory requirements. Consumer trading companies are using it

to grow their trade by providing loyalty cards and keeping a track of them.

9. **Transportation**: For better route planning, traffic monitoring and management, and logistics. This is mainly incorporated by governments to avoid congestion of traffic in a single place.
10. **Energy**: By introducing smart meters to reduce electrical leakages and help users to manage their energy usage. Load dispatch centers are using big data analysis to monitor the load patterns and discern the differences between the trends of energy consumption based on different parameters and as a way to incorporate daylight savings.

8 WAYS YOU CAN GROW YOUR BUSINESS WITH DATA SCIENCE

Today, the advent of Internet of Things and the development of AI technology have simplified implementation of big data solutions to the degree that even medium to small scale businesses are benefiting from it. And since the top 10 list comprises of sectors

which are directly or indirectly associated with various businesses, the imperative of this technology increases even further. Using big data analytics, businesses can take informed decisions and better their operational efficiency in a number of ways.

E.g. Using big data analytics, businesses can take informed decisions and better their operational efficiency in a number of ways. E.g.

- Utilizing company data to identify the need for improvement in existing policies and processes.
- Utilizing customer data available with the company such as social media streams, credit information, and internal or external consumer research, to improve or develop new products and services.

Deploying data science for your business –

1. **Empowers management to make better decisions**

   big data analytics acts as a trusted advisor for an organization's strategic planning. It helps your management and staff in enhancing their analytical abilities and thereby improves their overall decision-making skills. Measuring, recording and tracking performance metrics then allow the upper management to set new goals.

2. **Helps identify trends to stay competitive**

as mentioned earlier in this post; one of data analytics' primary objectives is to determine patterns within large data sets. This is particularly useful for identifying new and emerging market trends. Once identified these trends could become the key to gaining a competitive advantage by introducing new products and services.

3. **Increases the efficiency and commitment of staff in handling core tasks and issue**

by making employees aware of benefits of using the organization's analytics product, data science can make them more efficient at their jobs. Working with a greater insight into company goals, these employees will be able to drive more action towards core tasks and issues at every stage. Hence, improving the overall operational efficiency of your business.

4. **Identifies and acts upon opportunities**

Data science is all about constantly looking for areas of improvement in the organizational workings. By discovering inconsistencies in the organizational processes and existing analytical systems, data scientists

can introduce new ways of doing things. This, in turn, can drive innovation and allow new product development, opening profitable avenues for your company.

5. **Promotes low risk data-driven action plans**

Big data analytics has made it possible for small and big businesses to take actions based on quantifiable, data-driven evidence. Such a strategy can save a business from unnecessary tasks and sometimes foreshadow risks.

6. **Validates decisions**

Apart from allowing your business to base decisions on data, analytics also helps you test these decisions by introducing variable factors, to check for flexibility and scalability. Using data science and big data solutions you can introduce favorable changes in your organizational structure and functioning.

7. **Helps in selecting target audience**

one of the key value props of big data analytics is how you can shape customer data to provide more insight into consumer preference and expectations. A deeper analysis of customer data can help companies in

identifying and targeting audience with utmost precision using tailor-made products and services.

8. **Facilitates sensible recruitment of talent**

Human resource departments are constantly at work in companies to find talent that fits the prescribed criteria. Big data has made their task simpler by providing comprehensive data profiles on individuals by merging social media, corporate profiles and job search databases. Now your HR Department can process CVs much faster and recruit the right talent quickly and without compromises.

CONCLUSION

The world is moving towards a more connected future, and big data solutions are going to play a big part in automation and development of AI technologies. Companies like Google are already using Machine Learning processes for greater precision in delivering their services. As technologies around the globe become more synchronous and interoperable, big data will become the core that connects them together. Therefore, companies using big data solutions need to keep up with its evolving nature while those still reluctant to invest should rethink their organizational policies. There are a few pointers which can be helpful in getting the most out of your investment in big data.

- Demand a value proposition from big data by investing in adequate technologies to capture and store data. If you do not have the data, then you do not have the benefits. Data discovery tools can help you in digging up big data which is relevant to your business.

- Make use of big data to improve and innovate your applications and services.
- Arrange organization-wide training to accustom your staff to big data solutions and their usage.
- Interact and collaborate with big data users from associated fields of businesses to derive more benefits and bring down usage costs.
- Avoid soloed big data management and stay open to integration with shared enterprise infrastructures.
- If shifting to a new data platform, choose those which have a special support system for big data, such as in-memory processing, MapReduce, etc.
- Develop a tech strategy for your organization’s data and lay out a plan for capturing and processing them in the long-run.
- Plan the financials for storage and processing of your big data as well.

Moreover, big data is also resonating with government and a public-sector agency, which is a good sign for businesses all around the

world as this will help deepen the public-private collaboration in a range of fields.

Big data Analytics products

Efficiently capture, process, and analyze data with Google Cloud data analytics products.

Fully managed server less approach

Seamlessly unlock business insights with Google Cloud Platform's fully managed, proven, end-to-end data analytics products. Remove operational complexities of data analytics with a server less approach and make important business decisions quickly and efficiently.

BigQuery

Server less data warehouse

BigQuery is Google's fully managed, low cost, serverless data warehouse that scales with your storage and computing power needs. With BigQuery, you get a columnar and ANSI SQL database that can analyze terabytes to petabytes of data in blazing-fast speeds. Analyze geospatial data using familiar SQL with BigQuery GIS. Quickly build and operationalize ML models on large-scale structured or semi-structured data using simple SQL with BigQuery ML, and support real-time interactive dashboarding with sub-second query latency using BigQuery BI Engine. Plus, BigQuery offers data transferring services, flexible data ingestion, and pay-for-what-you-use pricing.

Cloud Dataflow

Batch and stream data processing

Cloud Dataflow is a fully managed service for transforming and enriching data in stream (real time) and batch (historical) modes with equal reliability and expressiveness. Cloud Dataflow's server less approach frees you from operational tasks like capacity planning, resource management, and performance optimization, while you only pay for what you use. Plus, Cloud Dataflow not only works with Google's ingestion, data warehousing, and machine learning products, but also third-party tools like Apache Spark and Apache Beam.

Cloud Dataproc

Managed Apache Spark and Apache Hadoop

Cloud Dataproc is a fast, easy-to-use, fully managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way. Operations that used to take hours or days take seconds or minutes instead — and you pay only for the resources you use with per-second billing. Cloud Dataproc integrates with storage, compute, and monitoring services across Google Cloud products, giving you a powerful and complete data processing platform.

Cloud Pub/Sub

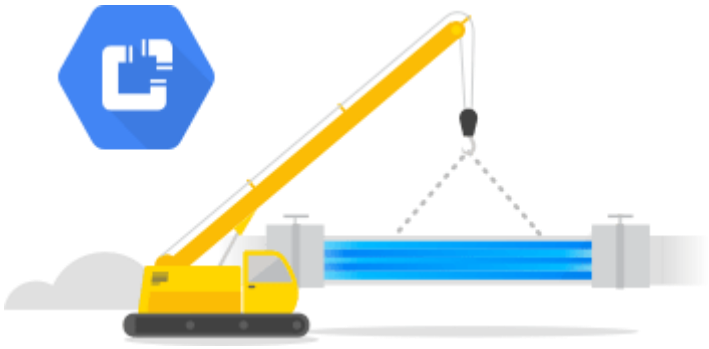
Scalable event ingestion and messaging service

Cloud Pub/Sub is a simple, reliable, scalable foundation for stream analytics and event-driven computing systems. You can send and receive messages between independent applications and syndicate data across projects and applications running on cloud, on-premises, or hybrid environments. You can leverage Cloud Pub/Sub's flexibility to decouple systems and components hosted on GCP or elsewhere on the internet. And Cloud Pub/Sub is designed to provide "at least once" delivery at low latency with on-demand scaling to tens of millions of messages per second.

Cloud Data Fusion

Fully managed, code-free data integration service

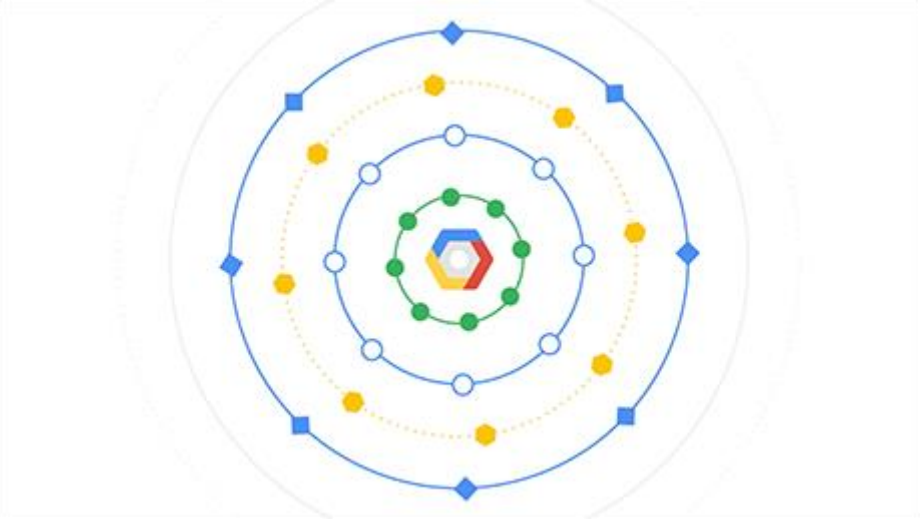
Cloud Data Fusion is a fully managed, cloud-native data integration service that helps users efficiently build and manage ETL/ELT data pipelines. With a graphical interface and a broad open-source library of preconfigured connectors and transformations, Data Fusion shifts an organization's focus away from code and integration to insights and action.



Cloud Composer

Workflow orchestration

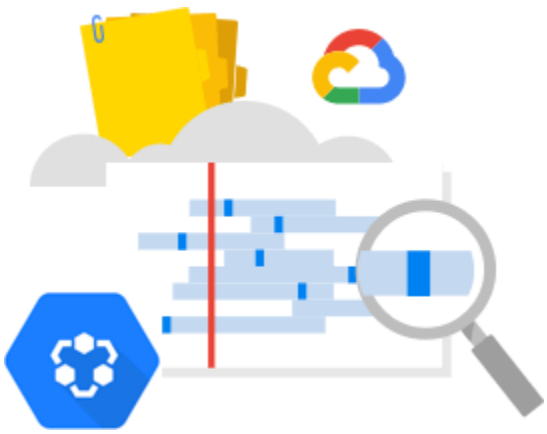
Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor workflows that span across clouds and on-premises data centers. Built on the popular Apache Airflow open-source project and operated using the Python programming language, Cloud Composer is free from lock-in and easy to use. Plus, with end-to-end integration for GCP workloads, you can orchestrate a full workflow with all of Google Cloud's big data products.



Data Catalog

Fully managed data discovery service

Data Catalog is a fully managed and scalable metadata management service that empowers organizations to quickly discover, manage, and understand all their data in Google Cloud. It offers a simple and easy-to-use search interface for data discovery, a flexible and powerful cataloging system for capturing both technical and business metadata, and a strong security and compliance foundation with Cloud Data Loss Prevention (DLP) and Cloud Identity and Access Management (IAM) integrations.



Google Data Studio

Serverless visual analytics

Google Data Studio is fully managed visual analytics services that can help anyone in your organization unlock insights from data through easy-to-create and interactive dashboards that inspire smarter business decision-making. When Data Studio is combined with BigQuery BI Engine, an in-memory analysis service, data exploration and visual interactivity reach sub-second speeds, over massive datasets.

Google Sheets

Big data in familiar spreadsheets

The connected sheet is a new feature of Sheets that activates only when using the Sheets data connector for BigQuery, allowing you to access, analyze, visualize, and collaborate on up to 10 billion rows of BigQuery data — no SQL scripts needed. Sheets end users can surface insights without needing the help of BigQuery experts or analysts fluent in SQL, saving time for everyone. Sheets users can then make sense of that data with spreadsheet formulas or perform deeper analysis with features like Explore, pivot tables, and charts. It's that easy.

Cloud Dataprep

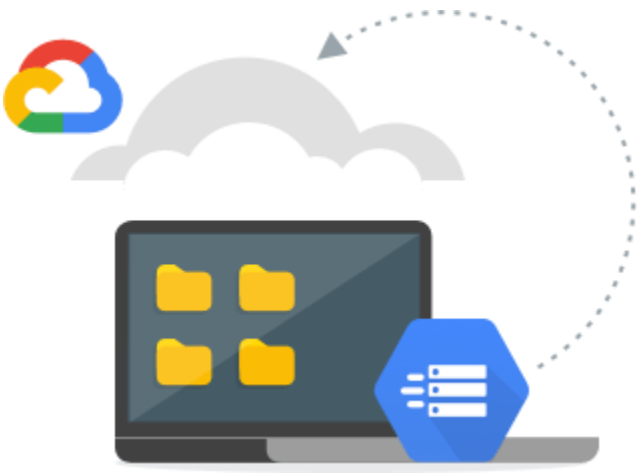
Intelligent data preparation

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Because Cloud Dataprep is server less and works at any scale, there is no infrastructure to deploy or manage. Your next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code. And with automatic schema, data type, possible joins, and anomaly detection, you can skip time-consuming data profiling and focus on data analysis.

Cloud Data Transfer

Data transfer services

Cloud Data Transfer offers solutions that meet your unique data transfer needs and move data to Cloud Storage, BigQuery, or Cloud Dataproc quickly and securely. Whether you have 50 gigabytes or 50 petabytes of data, one-time or recurring transfers, or a T1 line or a 10 Gbps network connection, our data transfer services are here to meet your unique needs.



Cloud Bigtable

Fully managed NoSQL database service

Cloud Bigtable provides a massively scalable NoSQL database suitable for low-latency and high-throughput workloads. It integrates easily with popular big-data tools like Hadoop and Spark, and it supports the open-source, industry-standard HBase API. Cloud Bigtable is a great choice for both operational and analytical applications, including IoT, user analytics, and financial data analysis.

Cloud Storage

Object storage for big data

Cloud Storage is a unified object storage solution that allows worldwide storage and retrieval of any amount of data at any time. You can use Cloud Storage for a range of scenarios, including serving website content, storing data for archival and disaster recovery, or distributing large data objects to users via direct download.

AI Platform Notebooks

An enterprise notebook service to launch projects in minutes

AI Platform Notebooks is a managed service whose integrated JupyterLab environment makes it easy to create instances that come pre-installed with the latest data science and ML frameworks and integrate with BigQuery, Cloud Dataproc, and Cloud Dataflow for easy development and deployment.

Introduction to Big Data Analytics Tools

Big data analytics is the process, it is used to examine the varied and large amount of data sets that to uncover unknown correlations, hidden patterns, market trends, customer preferences and most of the useful information which makes and help organizations to take business

decisions based on more information from Big data analysis. It is possible to process data very quickly and efficiently. That includes analyzing the data and make use of the result. This makes a lesser effort and more efficient where it is not possible with more traditional business intelligence solutions.

Big Data Analytics tools are very important for enterprises and large scale industries because of the huge volume of Data will be generated and managed by modern organizational tools using Bigdata tools. Big Data Analytics

tools help businesses in saving time and money and also in gaining insights to make data-driven decisions.

Big Data analytics is the complete process of collecting, gathering, organizing and analyzing the huge sets of data (known as Big Data) to observe/identify the patterns and also other useful information needed for business decisions. The process – Big Data analytics helps organizations for better understand the information which is present within the sets of data. The guy who works as Analysts working with Big Data typically will have the knowledge that comes from analyzing the data. There are different types of tools are available under the Data Analytics which help to improve the data analyzing the process that are data analysis, data

cleansing, data mining, data visualization, data integration, data storage, and management.

For the process of big data analytics, there is a need for very High-Performance Analytics. Hence to analyze such a huge volume of data, specialized software tools required for Big Data analytics process and applications for predictive analytics, data mining, text mining, forecasting, and data optimization.

There are many advantages to using Big Data Analytics. This process of big data analytics with high-performance in predictive analytics, data mining, text mining, forecasting that data, and optimization helps the enterprises in getting benefit in many areas which includes new revenue opportunities in business, for more effective marketing of the products, in providing much better customer service, in improving operational efficiency and also can make decisions on competitive advantages over market competitors.

To be Specific on the Big Data Analytics process, it enables enterprises to breakdown/narrow their huge volume of data to the most relevant information and analyzes it to inform critical business decisions. This proactive approach to business is transformative because it gives analysts and decision-makers the power to move ahead with the best knowledge and insights available, often in real-time.

As mentioned earlier, a big data analytics process is not a single activity that encompasses a huge volume of data. Instead it advanced analytics that can be applied to large data, but in reality, several types of different technologies work together to achieve the most value from information. Below are the biggest and important technologies which involve in big data analytics process:

- Data management

- Data mining
- Hadoop
- In-memory analytics
- Predictive analytics.
- Text mining

There ‘N’ number of Big Data Analytics tools, below is the list of some of the top tools used to store and analyze Big Data. These Big Data Analytics tools can be further be classified into two Storage and Querying/Analysis.

1. Apache Hadoop:

Apache Hadoop, a big data analytics tool which is a java based free software framework. It helps in effective storage of huge amount of data in a storage place known as a

cluster. The special feature of this framework is it runs in parallel on a cluster and also has an ability to process huge data across all nodes in it. There is a storage system in Hadoop popularly known as the Hadoop Distributed File System (HDFS), which helps to splits the large volume of data and distribute across many nodes present in a cluster. It also performs the replication process of data in a cluster hence providing high availability and recovery from the failure – which increases the fault tolerance.

2. KNIME:

KNIME Analytics Platform is one of the leading open solutions for data-driven innovation. This tool helps in discovering the potential & hidden in a huge volume of data, it also performs mine for fresh insights, or predicts the

new futures. The KNIME Analytics Platform tool is very much helpful toolbox for data scientists.

3. OpenRefine:

OpenRefine is introduced as Google Refine. This tool is one of the efficient tools to work on the messy and large volume of data, that all include: cleansing data, transforming that data from one format another, and also to perform extending it with web services and external data. The open refine tool helps explore large data sets easily.

4. Orange:

Orange is famous open-source data visualization and helps in data analysis for beginner and as well to the expert. This tool provides interactive workflows with a large toolbox

option to create the same which helps in analysis and visualizing of data. An orange tool has many and different visualizations, that include bar charts, trees, scatter plots, to dendrograms, networks and heat maps.

5. RapidMiner:

RapidMiner tool operates using visual programming and also it is much capable of manipulating, analyzing and modeling the data. RapidMiner tools make data science teams easier and productive by using an open-source platform for all their jobs like machine learning, data prep, and model deployment. Because of its uniformity in the data science platform makes accelerates in a building of complete analytical workflows in a single environment which

helps in dramatically improving efficiency and short duration of time to value for data science projects.

Big data analytics tools are very much in need of business/enterprises which depend on quick and agile decisions to stay as competitive, and most likely big data analytics tools are important while business decisions based on their previous business data. Below are some of the different types of organizations which can make use of this technology:

Travel and hospitality:

In the travel and hospitality business, it is a very important and key factor to keep customers happy, but to make customer satisfy is harder to gauge. For example, in Resorts

and casinos, they will be having a very short span of an opportunity to turn around customer experience. Hence applications of Big data analytics gives the ability to collect customer data, apply statistical analytics to better understand and improve these businesses.

Retail:

Nowadays, customer service has emerged as a huge tree compared to past decades, knowledgeable shoppers always keep searching and expect retailers to understand exactly what they want and when those products need it. Here Big data analytics technology comes into the picture and helps the retailers meet customer demands.

Government:

There are few and particular government agencies always face some big challenge like how to prepare the budget for the public without any compromise on quality or productivity. Hence many agencies use big data analytics; which helps them in streamlining operations while giving the agency a more correct view of criminal activity to avoid preparing a feasible and good budget.

Healthcare:

Big data analytics can also be used in the healthcare industry. Maintaining the Patient records, their insurance information and health plans and also all other types of information which are difficult to manage. Hence it is so

important applications of big data analytics technology in the healthcare industry.

In the current situation, the volume of data is growing along with world population growth and technology growth. This is a clear sign/indication of increasingly widespread use and necessity of Big Data Analysis solutions. Big Data is not just a trend of technology but it is a business practice which helps the industries/enterprise to stay in this competitive world to make their proactive data-driven business decisions on to improve sales and marketing team performance, increase revenue.

This has been a guide to Big data analytics tools in the real world. Here we have discussed the Different Big data Analytics tools like Apache Hadoop, RapidMine, OpenRefine, etc.

What Is Big Data Analytics and why do I Need It for My Business?

Big data is primarily defined by the volume of a data set. Big data sets are generally huge – measuring tens of terabytes – and sometimes crossing the threshold of petabytes. The term big data was preceded by very large databases (VLDBs) which were managed using database management systems (DBMS). Today, big data falls under three categories of data sets – structured, unstructured and semi-structured.

Structured data sets comprise of data which can be used in its original form to derive results. Examples include relational data such as employee salary records. Most modern computers and applications are programmed to generate structured data in preset formats to make it easier to process.

Unstructured data sets, on the other hand, are without proper formatting and alignment. Examples include human texts, Google

search result outputs, etc. These random collections of data sets require more processing power and time for conversion into structured data sets so that they can help in deriving tangible results.Semi-Structured data sets are a combination of both structured and unstructured data. These data sets might have a proper structure and yet lack defining elements for sorting and processing. Examples include RFID and XML data.

Semi-Structured data sets are a combination of both structured and unstructured data. These data sets might have a proper structure and yet lack defining elements for sorting and processing. Examples include RFID and XML data.

Big data processing requires a particular setup of physical and virtual machines to derive results. The processing is done simultaneously to achieve results as quickly as possible. These days big data processing techniques also include Cloud Computing and Artificial Intelligence. These technologies help in reducing manual inputs and oversight by automating many processes and tasks.

The evolving nature of big data has made it difficult to give it a commonly accepted definition. Data sets are consigned the big data status based on technologies and tools required for their processing.

BIG DATA ANALYTICS – TECHNOLOGIES AND TOOLS

Big data analytics is the process of extracting useful information by analysing different types of big data sets. Big data analytics is used to discover hidden patterns, market trends and consumer preferences, for the benefit of organizational decision making. There are several steps and technologies involved in big data analytics.

Data Acquisition

Data acquisition has two components: identification and collection of big data. Identification of big data is done by analyzing the two natural formats of data – born digital and born analogue.

Born Digital Data

It is the information which has been captured through a digital medium, e.g. a computer or smartphone app, etc. This type of data has an ever expanding range since systems keep on collecting different kinds of information from users. Born digital data is traceable and can provide both personal and demographic business insights. Examples include Cookies, Web Analytics and GPS tracking.

Born Analogue Data

When information is in the form of pictures, videos and other such formats which relate to physical elements of our world, it is termed as analogue data. This data requires conversion into digital format by using sensors, such as cameras, voice recording, digital assistants, etc. The increasing reach of technology has also raised the rate at which traditionally analogue data is being converted or captured through digital mediums.

The second step in the data acquisition process is collection and storage of data sets identified as big data. Since the archaic DBMS

techniques were inadequate for managing big data, a new method is used for collecting and storing big data. The process is called MAD – magnetic, agile and deep. Since, managing big data requires a significant amount of processing and storage capacity, creating such systems is out-of-reach for most entities which rely on big data analytics. Thus, the most common solutions for big data processing today are based on two principles – distributed storage and Massive Parallel Processing a.k.a. MPP. Most of the high-end Hadoop platforms and specialty appliances use MPP configurations in their system.

Non-relational Databases

The databases that store these massive data sets have also evolved in how and where the data is stored. JavaScript Object Notation or JSON is the preferred protocol for saving big data nowadays. Using JSON, the tasks can be written in the application layer and allow better cross-platform functionalities. Thus enabling, agile development of scalable and flexible data solutions for the devs. Many companies are using it as a replacement of XML as a

way of transmitting structured data between the server and web application.

In-memory Database Systems

These database storage systems are designed to overcome one of the major hurdles in the way of big data processing – the time taken by traditional databases to access and process information. IMDB systems store the data in the RAM of big data servers, therefore, drastically reducing the storage I/O gap. Apache Spark is an example of IMDB systems. VoltDB, NuoDB and IBM solidDB are some more examples of the same.

Hybrid Data Storage and Processing Systems – Apache Hadoop

Apache Hadoop is a hybrid data storage and processing system which provides scalability and speed at reasonable costs for mid and small-scale businesses. It uses a Hadoop Distributed File System (HDFS) for storing large files across multiple systems known as cluster nodes. Hadoop has a replication mechanism to ensure smooth operation even during instances of individual node

failures. Hadoop uses Google's MapReduce parallel programming as its core. The name originates from 'Mapping' and 'Reduction' of functional programming languages in its algorithm for big data processing. MapReduce works on the premise of increasing the number of functional nodes over increasing processing power of individual nodes. Moreover, Hadoop can be run using readily available hardware which has sped up its development and popularity, significantly.

Data Mining

It is a recent concept which is based on contextual analysing of big data sets to discover the relationship between separate data items. The objective is to use a single data set for different purposes by different users. Data mining can be used for reducing costs and increasing revenues.

TOP 10 SECTORS USING BIG DATA ANALYTICS

Big data is finding usage in almost all industries today. Here is a list of the top segments using big data to give you an idea of its application and scope.

1. **Banking and Securities**: For monitoring financial markets through network activity monitors and natural language processors to reduce fraudulent transactions. Exchange Commissions or Trading Commissions are using big data analytics to ensure that no illegal trading happens by monitoring the stock market.
2. **Communications and Media**: For real-time reportage of events around the globe on several platforms (mobile, web and TV), simultaneously. Music industry, a segment of media, is using big data to keep an eye on the latest trends which are ultimately used by autotuning softwares to generate catchy tunes.
3. **Sports:** To understand the patterns of viewership of different events in specific regions and also monitor the performance of individual players and teams by analysis. Sporting events like Cricket world cup, FIFA world cup and Wimbledon make special use of big data analytics.
4. **Healthcare**: To collect public health data for faster responses to individual health problems and identify the global spread of new

virus strains such as Ebola. Health Ministries of different countries incorporate big data analytic tools to make proper use of data collected after Census and surveys.

5. **Education**: To update and upgrade prescribed literature for a variety of fields which are witnessing rapid development. Universities across the world are using it to monitor and track the performance of their students and faculties and map the interest of students in different subjects via attendance.

6. **Manufacturing**: To increase productivity by using big data to enhance supply chain management. Manufacturing companies use these analytical tools to ensure that are allocating the resources of production in an optimum manner which yields the maximum benefit.

7. **Insurance**: For everything from developing new products to handling claims through predictive analytics. Insurance companies use business big data to keep a track of the scheme of policy which is the most in demand and is generating the most revenue.

8. **Consumer Trade**: To predict and manage staffing and inventory requirements. Consumer trading companies are using it to grow their trade by providing loyalty cards and keeping a track of them.
9. **Transportation**: For better route planning, traffic monitoring and management, and logistics. This is mainly incorporated by governments to avoid congestion of traffic in a single place.
10. **Energy**: By introducing smart meters to reduce electrical leakages and help users to manage their energy usage. Load dispatch centers are using big data analysis to monitor the load patterns and discern the differences between the trends of energy consumption based on different parameters and as a way to incorporate daylight savings.

8 WAYS YOU CAN GROW YOUR BUSINESS WITH DATA SCIENCE

Today, the advent of Internet of Things and the development of AI technology has simplified implementation of big data solutions to the degree that even medium to small scale businesses are benefiting

from it. And since the top 10 list comprises of sectors which are directly or indirectly associated with various businesses, the imperative of this technology increases even further. Using big data analytics, businesses can take informed decisions and better their operational efficiency in a number of ways.

E.g. Using big data analytics, businesses can take informed decisions and better their operational efficiency in a number of ways. E.g.

- Utilizing company data to identify the need for improvement in existing policies and processes.
- Utilizing customer data available with the company such as social media streams, credit information, and internal or external consumer research, to improve or develop new products and services.

Deploying data science for your business –

1. **Empowers management to make better decisions**

   Big data analytics acts as a trusted advisor for an organization's strategic planning. It helps your management and staff in enhancing their analytical abilities and thereby improve their overall decision-making skills. Measuring, recording and tracking performance metrics then allow the upper management to set new goals.

2. **Helps identify trends to stay competitive**

As mentioned earlier in this post, one of data analytics' primary objectives is to determine patterns within large data sets. This is particularly useful for identifying new and emerging market trends. Once identified these trends could become the key to gaining a competitive advantage by introducing new products and services.

3. **Increases the efficiency and commitment of staff in handling core tasks and issue**

By making employees aware of benefits of using the organization's analytics product, data science can make them more efficient at their jobs. Working with a greater insight into company goals, these employees will be able to drive more action towards core tasks and issues at every stage. Hence, improving the overall operational efficiency of your business.

4. **Identifies and acts upon opportunities**

Data science is all about constantly looking for areas of improvement in the organizational workings. By discovering inconsistencies in the organizational processes and existing analytical systems, data scientists

can introduce new ways of doing things. This, in turn, can drive innovation and allow new product development, opening profitable avenues for your company.

5. **Promotes low risk data-driven action plans**

Big data analytics has made it possible for small and big businesses to take actions based on quantifiable, data-driven evidence. Such a strategy can save a business from unnecessary tasks and sometimes foreshadow risks.

6. **Validates decisions**

Apart from allowing your business to base decisions on data, analytics also helps you test these decisions by introducing variable factors, to check for flexibility and scalability. Using data science and big data solutions you can introduce favorable changes in your organizational structure and functioning.

7. **Helps in selecting target audience**

One of the key value props of big data analytics is how you can shape customer data to provide more insight into consumer preference and expectations. A deeper analysis of customer data can help companies in

identifying and targeting audience with utmost precision using tailor-made products and services.

8. **Facilitates sensible recruitment of talent**

Human resource departments are constantly at work in companies to find talent that fits the prescribed criteria. Big data has made their task simpler by providing comprehensive data profiles on individuals by merging social media, corporate profiles and job search databases. Now your HR Department can process CVs much faster and recruit the right talent quickly and without compromises.

CONCLUSION

The world is moving towards a more connected future, and big data solutions are going to play a big part in automation and development of AI technologies. Companies like Google are already using Machine Learning processes for greater precision in delivering their services. As technologies around the globe become more synchronous and interoperable, big data will become the core that connects them together. Therefore, companies using big data solutions need to keep up with its evolving nature while those still

reluctant to invest should rethink their organizational policies. There are a few pointers which can be helpful in getting the most out of your investment in big data.

- Demand a value proposition from big data by investing in adequate technologies to capture and store data. If you do not have the data, then you do not have the benefits. Data discovery tools can help you in digging up big data which is relevant to your business.
- Make use of big data to improve and innovate your applications and services.
- Arrange organization-wide training to accustom your staff to big data solutions and their usage.
- Interact and collaborate with big data users from associated fields of businesses to derive more benefits and bring down usage costs.
- Avoid siloed big data management and stay open to integration with shared enterprise infrastructures.

- If shifting to a new data platform, choose those which have a special support system for big data, such as in-memory processing, MapReduce, etc.
- Develop a tech strategy for your organization’s data and lay out a plan for capturing and processing them in the long-run.
- Plan the financials for storage and processing of your big data as well.

Moreover, big data is also resonating with government and a public-sector agency, which is a good sign for businesses all around the world as this will help deepen the public-private collaboration in a range of fields.

Chapter 6:

BIG DATA SICINETEST

What is a Data Scientist

Data scientists are big data wranglers, gathering and analyzing large sets of structured and unstructured data. A data scientist's role combines computer science, statistics, and mathematics. They analyze, process, and model data then interpret the results to create actionable plans for companies and other organizations.

Data scientists are analytical experts who utilize their skills in both technology and social science to find trends and manage data. They use industry knowledge, contextual understanding, skepticism of existing assumptions – to uncover solutions to business challenges.

A data scientist's work typically involves making sense of messy, unstructured data, from sources such as smart devices, social media feeds, and emails that don't neatly fit into a database.

Technical skills are not the only thing that matters, however. Data scientists often exist in business settings and are charged with communicating complex ideas and making data-driven organizational decisions. As a result, it is highly important for them to be effective communicators, leaders and team members as well as high-level analytical thinkers.

Experienced data scientists and data managers typically have over ten years of experience and are tasked with developing a company's best practices, from cleaning to processing and storing data. They work cross functionally with other teams throughout their organization, such as marketing, customer success, and operations. They are highly sought after in today's data and tech heavy economy, and their salaries and job growth clearly reflect that.

Steps to Become a Data Scientist

Here are six steps to consider if you're interested in pursuing a career in data science:

1. Pursue an undergraduate degree in data science or a closely related field
2. Learn required skills to become a data scientist
3. Consider a specialization
4. Get your first entry-level data scientist job
5. Review additional data scientist certifications and post-graduate learning
6. Earn a master's degree in data science

How to Become a Data Scientist in 2019

1. Pursue an undergraduate degree in data science or a closely related field

You will need at least a bachelor's degree in data science, mathematics, statistics, and computer science to get your foot in the door as an entry level data scientist. Degrees also add structure, internships, networking and recognized academic qualifications for your résumé. However, if you've received a bachelor's degree in a different field, you may need to focus on developing skills needed for the job through online short courses or bootcamps.

2. Learn the required skills to become a data scientist

- Programming
- Machine Learning techniques
- Data Visualization and Reporting
- Risk Analysis
- Statistical analysis and Math
- Effective Communication
- Software Engineering Skills
- Data Mining, Cleaning and Munging
- Research
- Big Data Platforms
- Cloud Tools
- Data warehousing and structures

This list is always subject to change. As Anmol Rajpurohit suggests in his guide to becoming a data scientist, "generic programming skills are a lot

more important than being the expert of any particular programming language.

3. Consider a specialization

In demand data scientists typically specialize in a particular industry or develop strong skills in areas such as artificial intelligence, machine learning, research, or database management. Specialization is a good way to increase your earning potential and do work that is meaningful to you. According to the Burtchworks Study, entry data scientists working in the tech industry earn an average salary of $85,143 and senior data scientists working for consulting firms earn an average salary of $158,462.

Get your first entry level job as a data scientist

Once you've acquired the right skills and/or specialization, you should be ready for your first data science role! It may be useful to create an online portfolio to display a few projects and showcase your accomplishments to potential employers. You also may want to consider a company where there's room for growth since your first data science job may not have the title data scientist, but could be more of an analytical role. You'll quickly learn how to work on a team and best practices that will prepare you for more senior positions.

Review additional data scientist certifications and post-graduate learning

Here are a few certifications that focus on useful skills:

Certified Analytics Professional (CAP)
CAP was created by the Institute for Operations Research and the Management Sciences (INFORMS) and is targeted towards data scientists. During the certification exam, candidates must demonstrate their expertise of the end-to-end analytics process. This includes the framing of business and analytics problems, data and methodology, model building, deployment and life cycle management.

SAS Certified Predictive Modeler using SAS Enterprise Miner 14
This certification is designed for SAS Enterprise Miner users who perform

predictive analytics. Candidates must have a deep, practical understanding of the functionalities for predictive modeling available in SAS Enterprise Miner 14.

Earn a master's degree in data science

Academic qualifications may be more important than you imagine. When it comes to most data science jobs, is a master's required? It depends on the job and some working data scientists have a bachelor's or have graduated from a data science bootcamp. However, as Burtchworks notes, data scientists typically have a graduate or advanced degree in a quantitative discipline. The Burtch Works study also shares that most data scientists have an advanced degree, either a master's or Ph.D.

Data Scientist Responsibilities

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician". – Josh Wills on the difference between data scientists and statisticians

On any given day, a data scientist's responsibilities may include:

- Solving business problems through undirected research and framing open-ended industry questions
- Extract huge volumes of structured and unstructured data. They query structured data from relational databases using programming languages such as SQL. They gather unstructured data through web scraping, APIs, and surveys.
- Employ sophisticated analytical methods, machine learning and statistical methods to prepare data for use in predictive and prescriptive modeling
- Thoroughly clean data to discard irrelevant information and prepare the data for preprocessing and modeling
- Perform exploratory data analysis (EDA) to determine how to handle missing data and to look for trends and/or opportunities
- Discovering new algorithms to solve problems and build programs to automate repetitive work

- Communicate predictions and findings to management and IT departments through effective data visualizations and reports
- Recommend cost-effective changes to existing procedures and strategies

Every company will have a different take on data science job tasks. Some treat their data scientists as data analysts or combine their duties with data engineers; others need top-level analytics experts skilled in intense machine learning and data visualizations.

As data scientists achieve new levels of experience or change jobs, their responsibilities invariably change. For example, a person working alone in a mid-size company may spend a good portion of the day in data cleaning and munging. A high-level employee in a business that offers data-based services may be asked to structure big data projects or create new products.

Characteristics of a Successful Data Scientist Professional

Data scientists don't need to just understand programming languages, management of databases and how to transpose data into visualizations – they should be naturally curious about their surrounding world, but through an analytical lens. Possessing personality traits that resemble quality assurance departments, data scientists may be meticulous as they review large amounts of data and seek out patterns and answers. They are also creative in making new algorithms to crawl data or devising organized database warehouses.

Generally, professionals in the data science field must know how to communicate in several different modes, i.e to their team, stakeholders and clients. There may be a lot of dead ends, wrong turns, or bumpy roads, but data scientists should possess drive and grit to stay afloat with patience in their research.

"Successful data scientists have a strong technical background, but the best data scientists also have great intuition about data. Are the features meaningful, and do they reflect what you think they should mean? Given the way your data is distributed, which model should you be using? What

does it mean if a value is missing, and what should you do with it? The best data scientists are also great at communicating, both to other data scientists and non-technical people. In order to be effective at Airbnb, our analyses have to be both technically rigorous and presented in a clear and actionable way to other members of the company."

–Lisa Qian, Data Scientist at Airbnb

Data Science Job Outlook

According to the Bureau of Labor and Statistics (BLS), employment growth of computer information and research scientists by 2028 is 16%, over three times faster than the national average. Demand for experienced data scientists is high, but you have to start somewhere. Some data scientists get their foot in the door working as entry-level data analysts, extracting structured data from MySQL databases or CRM systems, developing basic visualizations in Tableau or analyzing A/B test results. If you'd like to push beyond your analytical role – think about what you could do with a career in data science:

- Data/Big Data Engineer
- Data/Big Data Architect

Companies of every size and industry – from Google, LinkedIn and Amazon to the humble retail store – are looking for experts to help them wrestle big data into submission. There are many different types of data

scientist jobs, but even as demand for data engineers surges, job postings for big data experts are expected to remain high. In certain companies, “new look” data scientists may find themselves responsible for financial planning, ROI assessment, budgets and a host of other duties related to the management of an organization.

Data Scientist Salary

A data scientist’s salary depends on years of experience, skill set, education, and location. According to The Burtch Works Study, employers place greater value on data scientists with specialized skills, such as Natural Language Processing or Artificial Intelligence. West coast Data scientists earn the highest average salary and entry level data scientists can expect to earn at least $90,000. The BLS claims skilled computer research and information scientists, who include data scientists, enjoy excellent job prospects because of high demand.

Data Scientist
Average Data Scientist Salary: $118,370 per year
Lowest 10%: $69,230
Highest 10%: $183,820

Senior Data Scientist
Median Sr. Data Scientist Salary: $171,755
Total Pay Range: $147,000 – $200,000

9 Must-have skills you need to become a Data Scientist, updated

1. Education

Data scientists are highly educated – 88% have at least a Master's degree and 46% have PhDs – and while there are notable exceptions, a very strong educational background is usually required to develop the depth of knowledge necessary to be a data scientist. To become a data scientist, you could earn a Bachelor's degree in Computer science, Social sciences, Physical sciences, and Statistics. The most common fields of study are Mathematics and Statistics (32%), followed by Computer Science (19%) and Engineering (16%). A degree in any of these courses will give you the skills you need to process and analyze big data.

After your degree programmer, you are not done yet. The truth is, most data scientists have a Master's degree or Ph.D and they also undertake online training to learn a special skill like how to use Hadoop or Big Data querying. Therefore, you can enroll for a master's degree program in the field of Data science, Mathematics, Astrophysics or any other related field.

The skills you have learned during your degree programme will enable you to easily transition to data science.

Apart from classroom learning, you can practice what you learned in the classroom by building an app, starting a blog or exploring data analysis to enable you to learn more.

2. R Programming
In-depth knowledge of at least one of these analytical tools, for data science R is generally preferred. R is specifically designed for data science needs. You can use R to solve any problem you encounter in data science. In fact, 43 percent of data scientists are using R to solve statistical problems. However, R has a steep learning curve.

It is difficult to learn especially if you already mastered a programming language. Nonetheless, there are great resources on the internet to get you started in R such as Simplilearn's Data Science Training with R Programming Language. It is a great resource for aspiring data scientists.

Technical Skills: Computer Science

3. Python Coding
Python is the most common coding language I typically see required in data science roles, along with Java, Perl, or C/C++. Python is a great programming language for data scientists. This is why 40 percent of respondents surveyed by O'Reilly use Python as their major programming language.
Because of its versatility, you can use Python for almost all the steps involved in data science processes. It can take various formats of data and you can easily import SQL tables into your code. It allows you to create datasets and you can literally find any type of dataset you need on Google.

4. Hadoop Platform
Although this isn’t always a requirement, it is heavily preferred in many cases. Having experience with Hive or Pig is also a strong selling point. Familiarity with cloud tools such as Amazon S3 can also be beneficial. A

study carried out by CrowdFlower on 3490 LinkedIn data science jobs ranked Apache Hadoop as the second most important skill for a data scientist with 49% rating.
As a data scientist, you may encounter a situation where the volume of data you have exceeds the memory of your system or you need to send data to different servers, this is where Hadoop comes in. You can use Hadoop to quickly convey data to various points on a system. That's not all. You can use Hadoop for data exploration, data filtration, data sampling and summarization.

5. SQL Database/Coding
Even though NoSQL and Hadoop have become a large component of data science, it is still expected that a candidate will be able to write and execute complex queries in SQL. SQL (structured query language) is a programming language that can help you to carry out operations like add, delete and extract data from a database. It can also help you to carry out analytical functions and transform database structures.

You need to be proficient in SQL as a data scientist. This is because SQL is specifically designed to help you access, communicate and work on data. It gives you insights when you use it to query a database. It has concise commands that can help you to save time and lessen the amount of programming you need to perform difficult queries. Learning SQL will help you to better understand relational databases and boost your profile as a data scientist.

6. Apache Spark
Apache Spark is becoming the most popular big data technology worldwide. It is a big data computation framework just like Hadoop. The only difference is that Spark is faster than Hadoop. This is because Hadoop reads and writes to disk, which makes it slower, but Spark caches its computations in memory.

Apache Spark is specifically designed for data science to help run its complicated algorithm faster. It helps in disseminating data processing when you are dealing with a big sea of data thereby, saving time. It also

helps data scientist to handle complex unstructured data sets. You can use it on one machine or cluster of machines.

Apache spark makes it possible for data scientists to prevent loss of data in data science. The strength of Apache Spark lies in its speed and platform which makes it easy to carry out data science projects. With Apache spark, you can carry out analytics from data intake to distributing computing.

7. Machine Learning and AI

A large number of data scientists are not proficient in machine learning areas and techniques. This includes neural networks, reinforcement learning, adversarial learning, etc. If you want to stand out from other data scientists, you need to know Machine learning techniques such as supervised machine learning, decision trees, logistic regression etc. These skills will help you to solve different data science problems that are based on predictions of major organizational outcomes.

Data science needs the application of skills in different areas of machine learning. Kaggle, in one of its surveys, revealed that a small percentage of data professionals are competent in advanced machine learning skills such as Supervised machine learning, Unsupervised machine learning, Time series, Natural language processing, Outlier detection, Computer vision, Recommendation engines, Survival analysis, Reinforcement learning, and Adversarial learning.

Data science involves working with large amounts of data sets. You may want to be familiar with Machine learning.

8. Data Visualization

The business world produces a vast amount of data frequently. This data needs to be translated into a format that will be easy to comprehend. People naturally understand pictures in forms of charts and graphs more than raw data. An idiom says “A picture is worth a thousand words”.

As a data scientist, you must be able to visualize data with the aid of data visualization tools such as ggplot, d3.js and Matplottlib, and Tableau. These tools will help you to convert complex results from your projects to a format that will be easy to comprehend. The thing is, a lot of people do

not understand serial correlation or p values. You need to show them visually what those terms represent in your results.

Data visualization gives organizations the opportunity to work with data directly. They can quickly grasp insights that will help them to act on new business opportunities and stay ahead of competitions.

9. Unstructured data

It is critical that a data scientist be able to work with unstructured data. Unstructured data are undefined content that does not fit into database tables. Examples include videos, blog posts, customer reviews, social media posts, video feeds, audio etc. They are heavy texts lumped together. Sorting these type of data is difficult because they are not streamlined.

Most people referred to unstructured data as 'dark analytics" because of its complexity. Working with unstructured data helps you to unravel insights that can be useful for decision making. As a data scientist, you must have the ability to understand and manipulate unstructured data from different platforms.

Non-Technical Skills

10. Intellectual curiosity

"I have no special talent. I am only passionately curious."
-Albert Einstein.
No doubt you've seen this phrase everywhere lately, especially as it relates to data scientists. Frank Lo describes what it means, and talks about other necessary "soft skills" in his guest blog posted a few months ago.
Curiosity can be defined as the desire to acquire more knowledge. As a data scientist, you need to be able to ask questions about data because data scientists spend about 80 percent of their time discovering and preparing data. This is because data science field is a field that is evolving very fast and you have to learn more to keep up with the pace.
You need to regularly update your knowledge by reading contents online and reading relevant books on trends in data science. Don't be overwhelmed by the sheer amount of data that is flying around the internet,

you have to be able to know how to make sense of it all. Curiosity is one of the skills you need to succeed as a data scientist. For example, initially, you may not see much insight in the data you have collected. Curiosity will enable you to sift through the data to find answers and more insights.

11. Business acumen

To be a data scientist you'll need a solid understanding of the industry you're working in, and know what business problems your company is trying to solve. In terms of data science, being able to discern which problems are important to solve for the business is critical, in addition to identifying new ways the business should be leveraging its data.

To be able to do this, you must understand how the problem you solve can impact the business. This is why you need to know about how businesses operate so you can direct your efforts in the right direction.

12. Communication skills

Companies searching for a strong data scientist are looking for someone who can clearly and fluently translate their technical findings to a non-technical team, such as the Marketing or Sales departments. A data scientist must enable the business to make decisions by arming them with quantified insights, in addition to understanding the needs of their non-technical colleagues in order to wrangle the data appropriately. Check out our recent flash survey for more information on communication skills for quantitative professionals.

As well as speaking the same language the company understands, you also need to communicate by using data storytelling. As a data scientist, you have to know how to create a storyline around the data to make it easy for anyone to understand. For instance, presenting a table of data is not as effective as sharing the insights from those data in a storytelling format. Using storytelling will help you to properly communicate your findings to your employers.

When communicating, pay attention to results and values that are embedded in the data you analyzed. Most business owners don't want to know what you analyzed, they are interested in how it can impact their

business positively. Learn to focus on delivering value and building lasting relationships through communication.

13. Teamwork

A data scientist cannot work alone. You will have to work with company executives to develop strategies, work product managers and designers to create better products, work with marketers to launch better-converting campaigns, work with client and server software developers to create data pipelines and improve workflow. You will literally have to work with everyone in the organization, including your customers.

Essentially, you will be collaborating with your team members to develop use cases in order to know the business goals and data that will be required to solve problems. You will need to know the right approach to address the use cases, the data that is needed to solve the problem and how to translate and present the result into what can easily be understood by everyone involved.

Resources

1. **Advanced Degree** – More Data Science programs are popping up to serve the current demand, but there are also many Mathematics, Statistics, and Computer Science programs.
2. **MOOCs** –Coursera, Udacity, and codeacademy are good places to start.
3. **Certifications** – KDnuggets has compiled an extensive list.
4. **Bootcamps** – For more information about how this approach compares to degree programs or MOOCs, check out this guest blog from the data scientists at Datascope Analytics.
5. **Kaggle** – Kaggle hosts data science competitions where you can practice, hone your skills with messy, real world data, and tackle actual business problems. Employers take Kaggle rankings seriously, as they can be seen as relevant, hands-on project work.
6. **LinkedIn Groups** – Join relevant groups to interact with other members of the data science community.
7. **Data Science Central and KDnuggets** – Data Science Central and KDnuggets are good resources for staying at the forefront of industry trends in data science.

8. **The Burtch Works Study: Salaries of Data Scientists** – If you're looking for more information about the salaries and demographics of current data scientists be sure to download our data scientist salary study.

Chapter 7:

DATA SICENCE

What Is Data Science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract value from data. Data scientists combine a range of skills—including statistics, computer science, and business knowledge—to analyze data collected from the web, smart phones, customers, sensors, and other sources.

Data science reveals trends and produces insights that businesses can use to make better decisions and create more innovative products and services. Data is the bedrock of innovation, but its value comes from the information data scientists can glean from it and then act upon.

Tools for Data Scientists

Data scientists use many types of tools, but one of the most common is open source notebooks, which are web applications for writing and running code, visualizing data, and seeing the results—all in the same environment. Some of the most popular notebooks include Jupyter, RStudio, and Zepplin. Notebooks are very useful for conducting analysis but have their limitations when data scientists need to work as a team. Data science platforms emerged to solve this problem.

Data Science and the Growth of Data

What is big data? Learn more.

As modern technology has enabled the creation and storage of increasing amounts of information, the volume of data has soared. It's estimated that 90 percent of the data in the world was created in the last two years. For example, Facebook users upload 10 million photos every hour. The number of connected devices in the world—the Internet of Things (IoT)—is projected to grow to more than 75 billion by 2025.

The wealth of data being collected and stored by these technologies can bring transformative benefits to organizations and societies around the world, but only if we can interpret it. That's where data science comes in.

Explore this infographic for more information about data and data science.

The Emergence of the Data Scientist

As a specialty, data science is young. It grew out of the fields of statistical analysis and data mining. The *Data Science Journal* debuted in 2002, published by the International Council for Science: Committee on Data for Science and Technology. By 2008, the title of data scientist had emerged and the field quickly took off. There has been a shortage of data scientists ever since, even though more and more colleges and universities have started offering data science degrees.

A data scientist's duties can include developing strategies for analyzing data; preparing data for analysis; exploring, analyzing, and visualizing data; building models with data using programming languages such as Python and R; and deploying models into applications.

The data scientist doesn't work solo. In fact, the most effective data science is done in teams. In addition to a data scientist, this team might include a business analyst who defines the problem, a data engineer who prepares the data and how it is accessed, an IT architect who oversees the underlying processes and infrastructure,

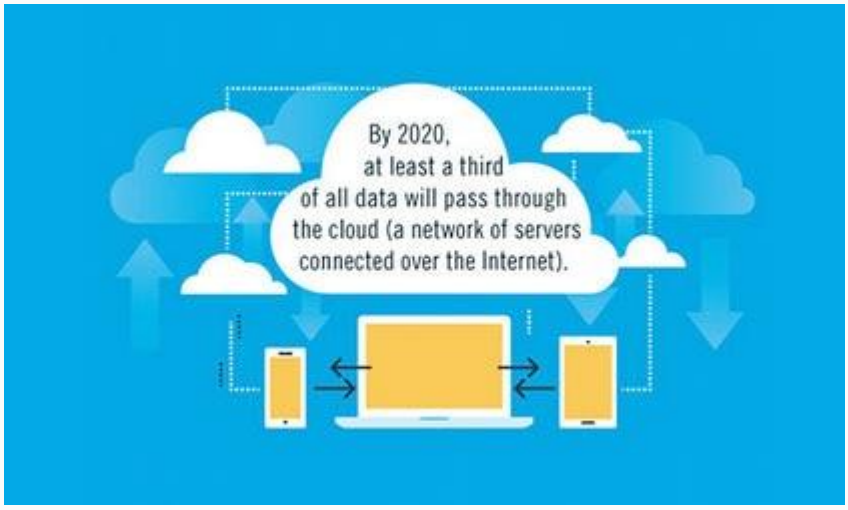
and an application developer who deploys the models or outputs of the analysis into applications and products.

How Data Science Is Transforming Business

Organizations are using data science teams to turn data into a competitive advantage by refining products and services. For example, companies analyze data collected from call centers to identify customers who are likely to churn, so marketing can take action to retain them. Logistics companies analyze traffic patterns, weather conditions, and other factors to improve delivery speeds and reduce costs. Healthcare companies analyze medical test data and reported symptoms to help doctors diagnose diseases earlier and treat them more effectively.

Most companies have made data science a priority and are investing in it heavily. In Gartner's recent survey of more than 3,000 CIOs, respondents ranked analytics and business intelligence as the top differentiating technology for their organizations. The CIOs surveyed see these technologies as the most strategic for their companies; therefore, they are attracting the most new investment.

How Data Science Is Conducted



The process of analyzing and acting upon data is iterative rather than linear, but this is how the work typically flows for a data modeling project:

- **Plan:** Define a project and its potential outputs
- **Prepare:** Build out the work environment, ensuring data scientists have the right tools as well as access to the right data and other resources such as compute power
- **Ingest:** Load the data into the work environment
- **Explore:** Analyze, explore, and visualize the data
- **Model:** Build, train, and validate models so they perform as required
- **Deploy:** Deploy models into production

Who Oversees the Data Science Process?

The data science process is typically overseen by three types of manager:

- **Business managers:** These managers work with the data science team to define the problem and develop a strategy for analysis. They may be the head of a line of business such as marketing, finance, or sales and have a data science team reporting to them.

They work closely with the data science and IT manager to ensure that projects are delivered.

- **IT managers:** Senior IT managers are responsible for the infrastructure planning and architecture that will support data science operations. They are continually monitoring operations and resource utilization to ensure that data science teams are operating efficiently and securely. They may also be responsible for building and updating environments for data science teams.
- **Data science managers:** These managers oversee the data science team and their day-to-day work. They are team builders who can balance team development with project planning and monitoring.

Challenges of Implementing Data Science

Despite the promise of data science and huge investments in data science teams, many companies are not realizing the full value of their data. In their race to hire talent and create data science programs, some companies have experienced inefficient team workflows, with different people using different tools and processes that don't work well together. Without more disciplined, central management, executives might not see a full return on their investments. This chaotic environment presents many challenges.

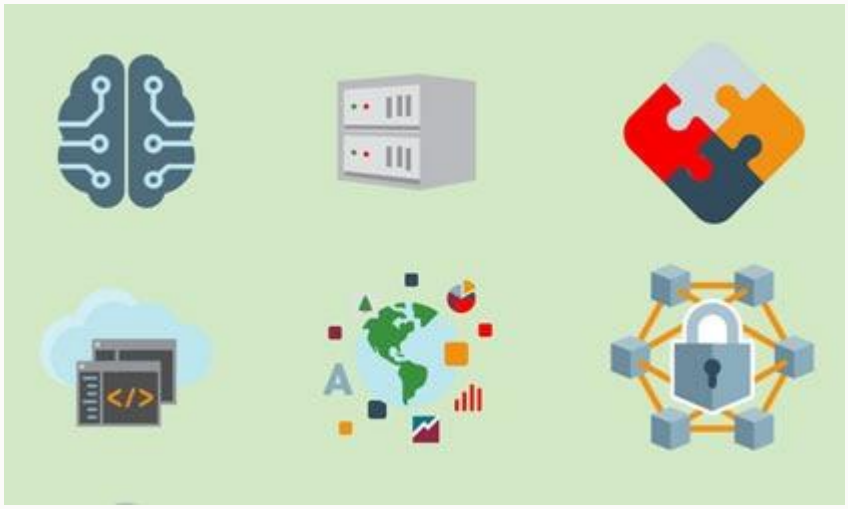
Data scientists can't work efficiently. Because access to data must be granted by an IT administrator, data scientists often have long waits for data and the resources they need to analyze it. Once they have access, the data science team might analyze the data using different and possibly incompatible tools. For example, a scientist might develop a model using the R language, but the application it will be used in is written in a different language. Which is why it can take weeks—or even months—to deploy the models into useful applications.

Application developers can’t access usable machine learning. Sometimes the machine learning models that developers receive must be recoded or are not ready to be deployed in applications. And because access points can be inflexible, models can’t be deployed in all scenarios and scalability is left to the application developer.

IT administrators spend too much time on support. Because of the proliferation of open source tools, IT has an ever-growing list of tools to support. A data scientist in marketing, for example, might be using different tools than a data scientist in finance. Teams might also have different workflows, which means IT must continually rebuild and update environments.

Business managers are too removed from data science. Data science workflows are not always integrated into business decision-making processes and systems, making it difficult for business managers to collaborate knowledgably with data scientists. Without better integration, business managers find it difficult to understand why it takes so long to go from prototype to production—and they are less likely to back the investment in projects they perceive as too slow.

The Data Science Platform Emerges



Find out more about data science platforms.

Companies realized that without an integrated platform, data science work was inefficient, unsecure, and difficult to scale. This realization led to the emergence of data science platforms. These platforms are software hubs around which all data science work takes place. A good platform alleviates many of the challenges of implementing data science and helps businesses turn their data into insights faster and more efficiently.

With a centralized platform, data scientists can work in a collaborative environment using their favorite open source tools, with all their work synced by a version control system.

Make an impact with a data science platform that supports self-sufficiency.

The Benefits of a Data Science Platform

A data science platform reduces redundancy and drives innovation by allowing teams to share code, results, and reports. It removes bottlenecks in the flow of work by simplifying management and using open source tools, frameworks, and infrastructure.

For example, a data science platform might allow data scientists to deploy models as APIs, making it easy to integrate them into different applications. Data scientists can access tools, data, and infrastructure without having to wait for IT.

The demand for data science platforms has exploded in the market. In fact, the platform market is expected to grow at a compounded annual rate of more than 39 percent over the next few years and is projected to reach US$385 billion by 2025.

What a Data Scientist Needs in a Platform

Learn why Forrester positioned Oracle as a leader of notebook-based machine learning solution providers.

If you're ready to explore the capabilities of data science platforms, there are some key capabilities to consider:

- **Choose a project-based UI that encourages collaboration.** . The platform should empower people to work together on a model, from conception to final development. It should give each team member self-service access to data and resources.
- **Prioritize integration and flexibility.** Make sure the platform includes support for the latest open source tools; common version control providers such as GitHub, GitLab, and Bitbucket; and tight integration with other resources.
- **Include enterprise-grade capabilities.** Ensure the platform can scale with your business as your team grows. The platform should be highly available, have robust access controls, and support a large number of concurrent users.
- **Make data science more self-service.** Look for a platform that takes the burden off of IT and engineering and makes it easy for data scientists to spin up environments instantly, track all of their work, and easily deploy models into production.

Companies Struggle to Find Data Science Talent

Finding and recruiting talent is the biggest barrier that companies face when they want to use data science for competitive advantage. In a recent McKinsey & Company survey, half of executives across geographies and industries reported greater difficulty in recruiting analytical talent than any other kind of skill. Retention is also a problem according to 40 percent of those surveyed.
In addition to data scientists, McKinsey reports that there are shortages in other analytics categories. In particular, there are shortages of skilled workers who can translate between business problems and the proper application of data science, and workers who are skilled at data visualization.

Indeed.com, Glassdoor, and Bloomberg provide further proof that there is significant demand for data science talent:

- Job postings for data scientists on Indeed.com rose 75 percent between January 2015 and January 2018. Job searches for data scientist roles rose 65 percent, according to Bloomberg.
- Glassdoor estimates that demand for data scientists in 2018 exceeded supply by 50 percent.
- Glassdoor rated data scientist as the top job in America—for the third year in a row.

Learn More About Artificial Intelligence

Artificial intelligence (AI) enables technology and machines to process data to learn, evolve, and execute human tasks.

Share this

What is Artificial Intelligence?

Learn More about Machine Learning

Machine learning, a subset of artificial intelligence (AI), focuses on building systems that learn through data with a goal to automate and speed time to decision and accelerate time to value.

What is Machine Learning?

News and Opinions

Machine learning, artificial intelligence, and data science are changing the way businesses approach complex problems to alter the trajectory of their respective industries. Read the latest articles to understand how the industry and your peers are approaching these technologies.

10 influential data scientists and why you should follow them on Twitter

No matter what industry you work in, chances are the applications your business uses are generating vast quantities of data that are increasing exponentially. The confluence of technologies that generate enormous amounts of data, store it, and analyze it in a timely manner affords businesses an unprecedented opportunity to put that data to work. Big data, when processed with predictive analytics algorithms, lets you find new patterns in data, and make increasingly accurate predictions about future business trends and opportunities. It's also playing a role in the software development life cycle, according to the 2015 World Quality Report survey.

The role of predictive analytics has become such a disruptive force in business that the *Harvard Business Review* recently suggested that data scientists might have the sexiest jobs of the 21st century.

Who are the people leading the charge? To help you keep up with the current trends and opportunities in data science and predictive analytics, we pulled together this list of influential data scientists worth following. Here, in

alphabetical order, are 10 data scientists you should be following.

1. Dean Abbott

Dean Abbott is co-founder and chief data scientist at SmarterHQ, and founder and president of Abbott Analytics. He is a co-author of the *IBM SPSS Modeler Cookbook*, and the author of *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Follow his blog at http://abbottanalytics.blogspot.com.

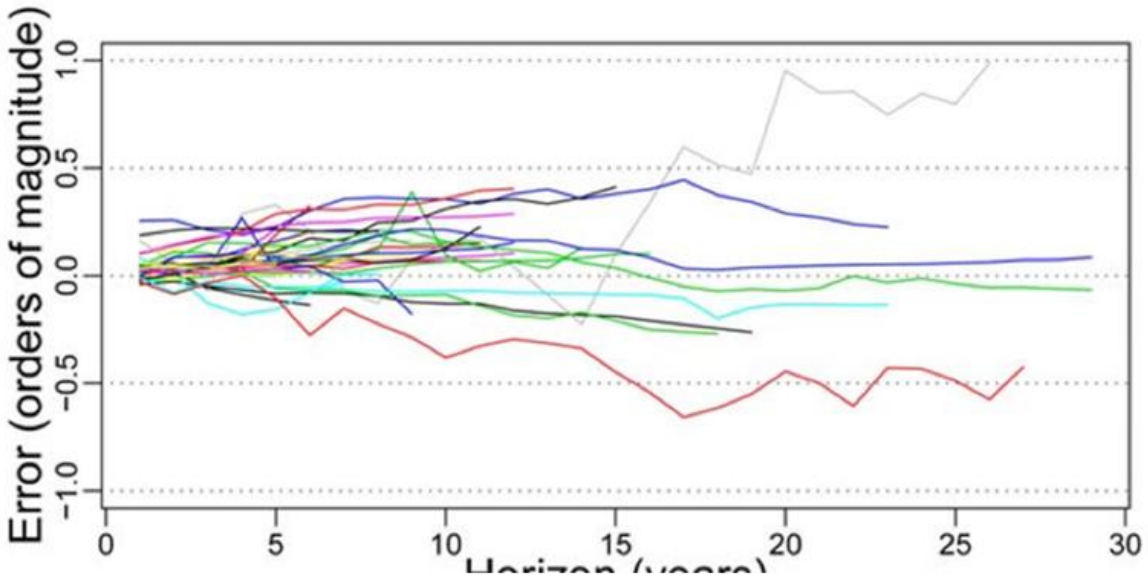
Dean Abbott@deanabb

2. Kenneth Cukier

Kenneth Cukier is the data editor for *The Economist*. He's a co-author of the book *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, and is a popular speaker. Watch him gave a fascinating TED talk on "Big data is better data."

Kenneth Cukier

✔@kncukier



3. John Elder

John Elder is the founder of data mining consultancy Elder Research, Inc. He is a frequent keynote speaker, and served for five years on a presidential panel tasked with guiding technology for national security. He is the author of several books, including the *Handbook of Statistical Analysis and Data Mining Applications*, *Ensemble Methods in Data Mining*, and *Practical Text Mining*, and is an adjunct professor at the University of Virginia, where he teaches the optimization of data mining. You can watch many of his presentations on YouTube.

John Elder@johnelder4

4. Bernard Marr

Bernard Marr, founder and CEO of the Advanced Performance Institute, regularly advises businesses and government organizations on how to gain better insights from their data. He is a contributor to the World Economic Forum, and is recognized by LinkedIn as one of the world's top 50 business influencers. He is a sought-after keynote speaker, and is the author of many articles and books, including *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*.

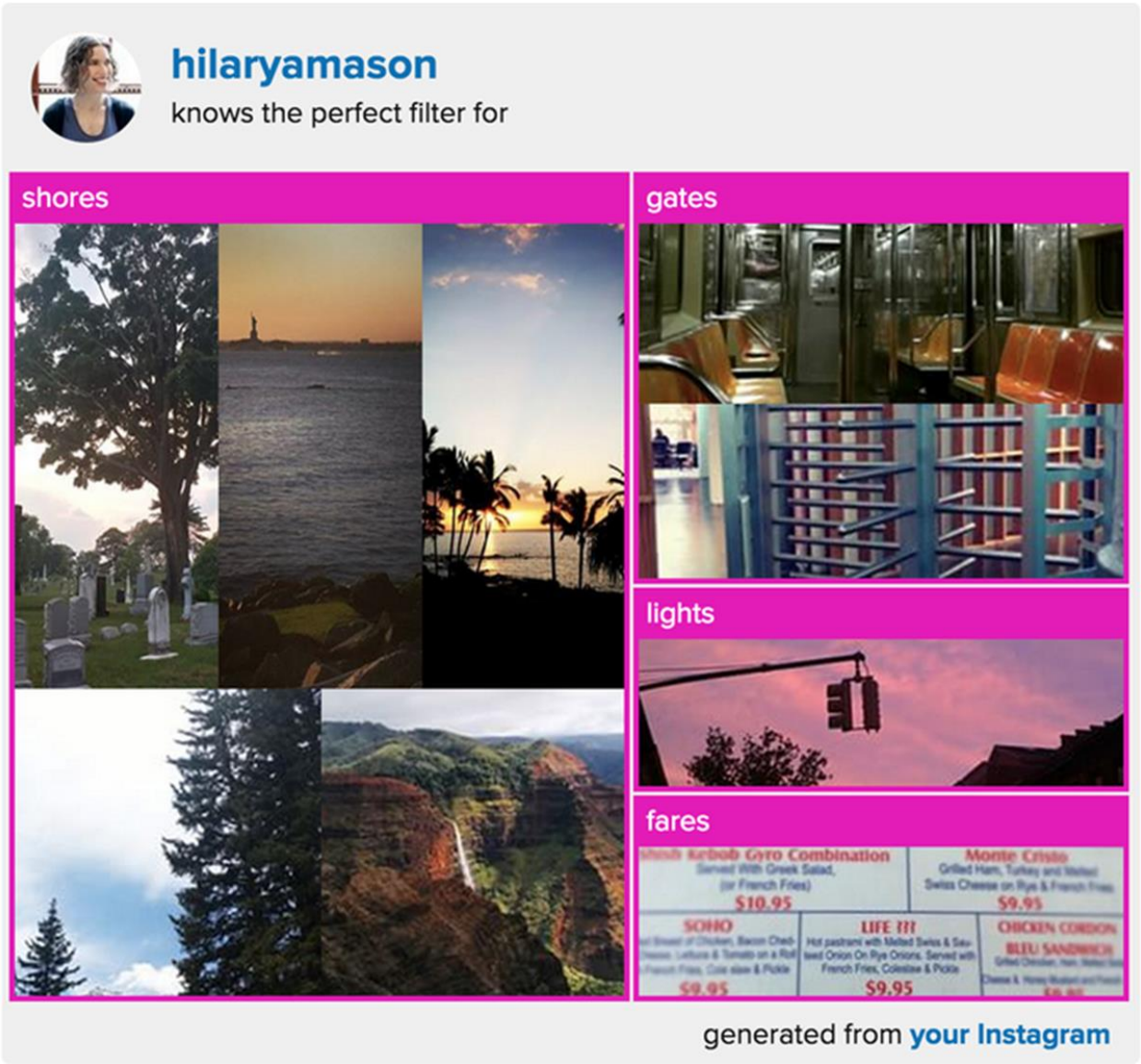
Bernard Marr@BernardMarr

5. Hilary Mason

Hilary Mason, founder of Fast Forward Labs, has also served as chief scientist at Bitly, Inc., co-founded HackNY, and is a member of NYCResistor. Mason was cited by Fortune in its 40 Under 40 list, and is a popular influencer on LinkedIn, where she has a large following. She enjoys speaking, and you can find many of her presentations on YouTube.

Hilary Mason

✔@hmason


hilaryamason
knows the perfect filter for
shores
gates
lights
fares
Monte Cristo
$9.95
SOHO
LIFE !!!
$9.95
$9.95
generated from your Instagram

6. Andrew Ng

Andrew Ng is chief scientist at Baidu Research, associate professor at Stanford University, and founder and chairman of the board at Coursera. He is the founder of the Google Brain project, which developed large-scale artificial neural networks, one of which taught itself to recognize cats in videos. He specializes in deep learning and has published over 100 papers in machine learning and other fields. You can find links to many of them from the Andrew Ng web page.

Andrew Ng

✔@AndrewYNg

7. Lillian Pierson

Lillian Pierson is the founder and chief data scientist at Data-Mania. She speaks, trains, and consults on data science, analytics, and big data, and is the author of the recently published *Data Science for Dummies*. She has a blog, and a YouTube channel where she explains concepts in data science.

Lillian Pierson, PE@Strategy_Gal

Using R & Google Analytics
TO BOOST YOUR
ONLINE VISIBILITY

8. Kira Radinsky

Kira Radinsky is the founder and CTO at SalesPredict, where she is pioneering artificial intelligence??based, predictive analytics applications that transform the way companies do business. She has won many awards, and has been recognized by Forbes 30 Under 30 as a rising star of enterprise technology. She maintains a blog, where she writes about the application of predictive analytics, and has contributed several articles to the *Harvard Business Review*. You can find many of her talks on YouTube.

Kira Radinsky@KiraRadinsky

9. Sebastian Thrun

Sebastian Thrun is CEO at Udacity and a research professor at Stanford University. He founded Google[x], where he led the Google driverless car project, and continues as an advisor. He is the recipient of numerous awards and prizes, a popular influencer on LinkedIn, and has a large following on Twitter. Thrun has authored and contributed to 11 books and hundreds of papers, and was named one of *Popular Science*'s Brilliant Ten scientists of 2005.

Sebastian Thrun@SebastianThrun

10. John Myles White

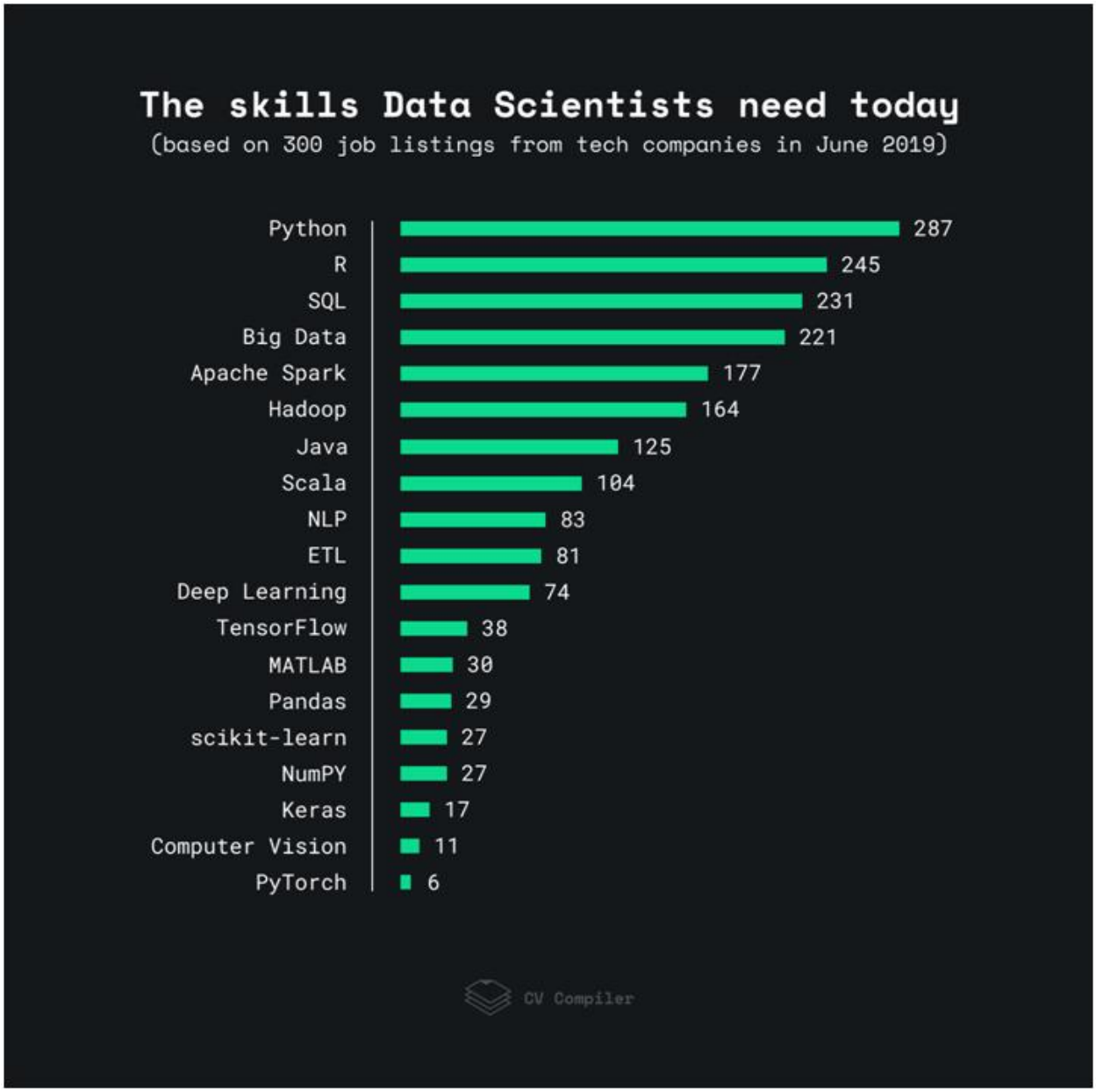
John Myles White is a data scientist at Facebook and a Julia developer. His areas of expertise include machine learning, statistics, data science, and the R programming language for statistics. He is a speaker and writer, and is the author of several books on machine learning, including *Bandit Algorithms for Website Optimization*, *Machine Learning for Hackers*, and *Machine Learning for Email*. You can find his presentations on YouTube.

John Myles White@johnmyleswhite

How to Become More Marketable as a Data Scientist

This headline may seem a bit odd to you. After all, if you're a data scientist in 2019, you're **already marketable**. Since data science has a huge impact on today's businesses, the demand for DS experts is growing. At the moment I'm writing this, there are 144,527 data science jobs on LinkedIn alone.
But still, it's important to keep your finger on the pulse of the industry to be aware of the fastest and most efficient data science solutions.

The following chart represents the skills employers are seeking from data science engineers in 2019:



For this analysis, we looked at 300 Data Science vacancies from StackOverflow, AngelList, and similar websites. Some terms might have been repeated more than once within one job listing.

Note: Bear in mind, this research represents the preferences of the employers, rather than the data science engineers them.

Key takeaways and Data Science trends

Obviously, Data Science is more about fundamental knowledge than frameworks and libraries, yet there are still some trends and technologies worth noting.

Big Data

According to the 2018 Big Data Analytics Market Study, Big Data adoption in enterprises soared from 17% in 2015 to 59% in 2018. Thus the popularity of Big Data tools also grew. If we don't take Apache Spark and Hadoop into account, (we will talk about the latter in detail in the next section), the most popular ones are **MapReduce** (36), and **Redshift** (29).

Hadoop

Despite the popularity of Spark and cloud storage, the 'era' of Hadoop hasn't yet ended. Ergo, some employers still expect candidates to be familiar with **Apache Pig** (30), **HBase** (32), and similar technologies. **HDFS** (20) is still being mentioned in vacancies as well.

Real-time data processing

With the increasing use of various sensors, mobile devices, and **IoT** (18), companies are aiming to get more insights from real-time data processing. Thus the stream analytics platforms such as **Apache Flink** (21) are popular among some employers.

Feature Engineering and Hyperparameter Tuning

Preparing data and selecting the model parameters is a key part of any data scientist's job. The term **Data Mining** (128) is quite popular among employers. Some employers also pay great attention to **Hyperparameter Tuning** (21). However, as a data scientist, you need to first pay attention to **Feature Engineering**. Choosing the best features for your model matters as they determine the success of your model on the earliest stage of its creation.

Data visualization

The ability to process data and extract valuable insights from it is vital. However, **Data Visualization** (55) is no less important of a skill for any data scientist. It's crucial that you could represent the outcomes of your work in a format, understandable to any team member or a customer. As for the data visualization tools, employers prefer **Tableau** (54).

General trends

In the vacancies, we encountered such terms as **AWS** (86), **Docker** (36), and **Kubernetes** (24). Hence the general trends in the software development industry are applicable to Data Science field, too.

What experts say

The technologies in this rating are on par. However, in Data Science, there are some things that are just as important as coding. It's the ability to glean insights from "data output" such as final data sets and trends, visualization, and telling the story with that data. Also, it's the ability to present the findings in a manner that is understandable. Know your audience — if they are Ph.D.'s, talk to them in an appropriate manner, but if they're from the C Suite, they won't care about programming — only results and ROI.
Carla Gentry,
Data Scientist/Owner at Analytical Solution
LinkedIn | Twitter

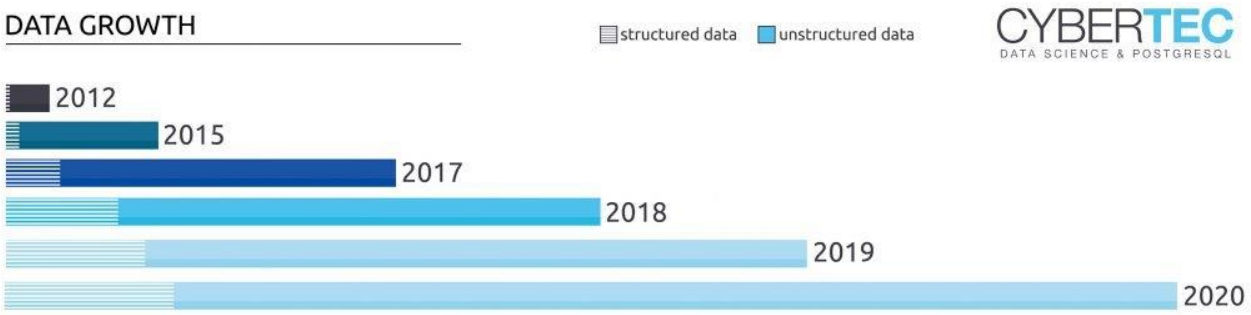
The snapshot data is useful to see the current state of the market but it doesn't represent the trends, so it's hard to plan for the future based on the snapshot alone. I would say that the usage of R will continue to steadily decline (the same can be said about MATLAB), while the popularity of Python among data scientists will keep rising. Hadoop and Big Data are on the list because the industry has some inertia: Hadoop will disappear (no one seriously invests in it anymore) and big data is no longer a hot trend. Whether one has to invest their time in learning Scala is unclear: Google officially supports Kotlin (also a JVM language), it's simpler to learn while Scala has a steep learning curve. I'm also skeptical about the future of TensorFlow: academia already switched to PyTorch and academia's influence is the strongest in data science compared to other industries. (The opinions are mine and might not represent Gartner's views.)
Andriy Burkov,
Director of Machine Learning at Gartner,
author of the Hundred-Page Machine Learning Book.
LinkedIn

PyTorch is the driving force of reinforcement learning with mathematical operations on CUDA tensors with GPUs. It is also a stronger framework for parallelizing the code natively on multiple GPUs at the same time unlike TensorFlow that requires to wrap each operation to a device.

PyTorch also builds dynamic graphs which are efficient for recurrent neural networks. Theano-based TensorFlow produces static charts and is more complicated to learn compared to Torch-based PyTorch. The TensorFlow reflects the larger community of developers and researchers. PyTorch will show more momentum, when it builds machine learning dashboard visualization tools such as TensorBoard. PyTorch is more Pythonic in terms of debugging and data visualization libraries with matplotlib and seaborn. Most of the debugging tools of Python can be leveraged to debug PyTorch as well. TensorFlow comes with its own debugging tool tfdbg.
Dr. Ganapathi Pulipaka,
Chief Data Scientist, Accenture,
winner of Top 50 Tech Leader Awards.
LinkedIn | Twitter

I think of data science "jobs" differently than data science "careers." Job listings offer insights into specific skills the market needs now but for a career, one of the most important skills I've seen is the ability to learn. Data science is a fast moving field and you need to be able to easily pick up new techniques, tools, and domain knowledge if you're going to succeed over the long term. Do that by challenging you and avoid getting too comfortable.
Lon Riesberg,
Founder/curator of Data Elixir,
Ex-NASA.
LinkedIn | Twitter
Data Science is a fast-evolving and complicated industry, where general knowledge matters as well as the experience with particular technologies. I hope this article helps you get valuable insights on what skills of both kinds you need to stay marketable in 2019. Good luck!

Data Science is the art of gaining knowledge from data, or of getting meaning into large databases. Due to the growing digitalization, industry 4.0 and IoT (Internet of Things), the amount of data collected is exploding. Statistics reveal that today we store as much data in just two days, as we did from the beginning of civilized mankind until the year 2000. Data science is an indispensible factor for the world of business, industry and research and it is gaining enormous importance.



Big Data: Massive Data Growth over the Years

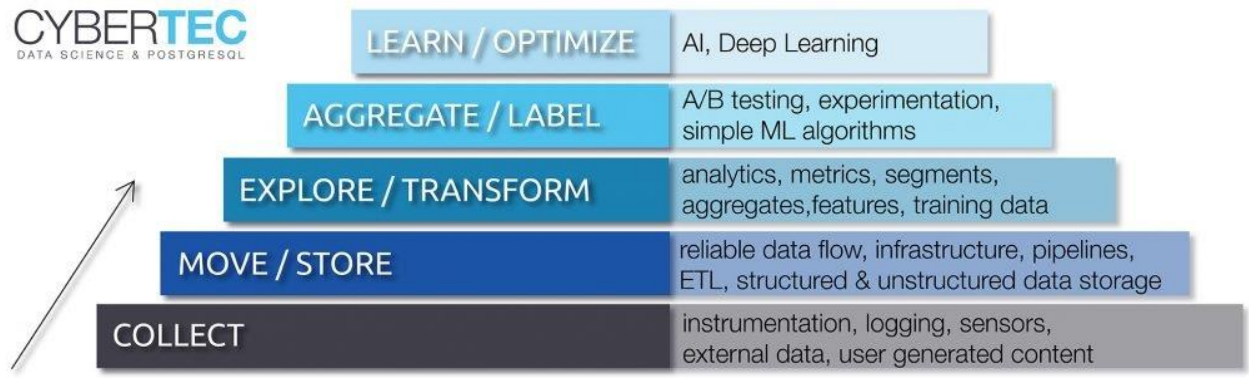
WHAT IS “DATA SCIENCE“

Data science is the general term and basis for numerous technological disciplines such as

- Artificial Intelligence
- Machine Learning
- Deep Learning
- Data Mining
- Big Data
- Business Intelligence
- Data Analysis
- Predictive Analytics
- And many more.

In order to use all these technologies wisely, effectively and profitably, proven specialists are required to support you in creating and implementing the right data strategy, architecture, infrastructure and the development of algorithms.

PROCEDURE OF DATA SCIENCE:



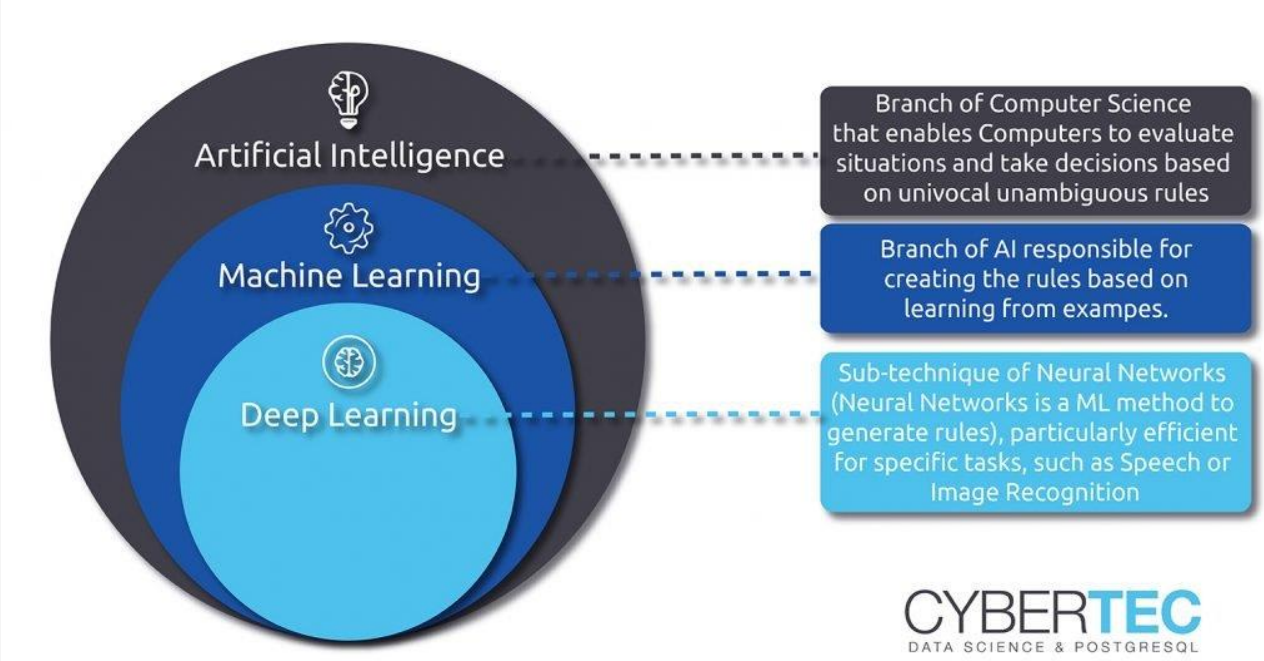
Data Science: How it works

„GOOD DATA SCIENTISTS ARE AS RARE AS UNICORNS“

CYBERTEC has specialized in these services for many years and has an international team of data scientists, architects and engineers. We support our customers worldwide in generating sustainable value from their data.

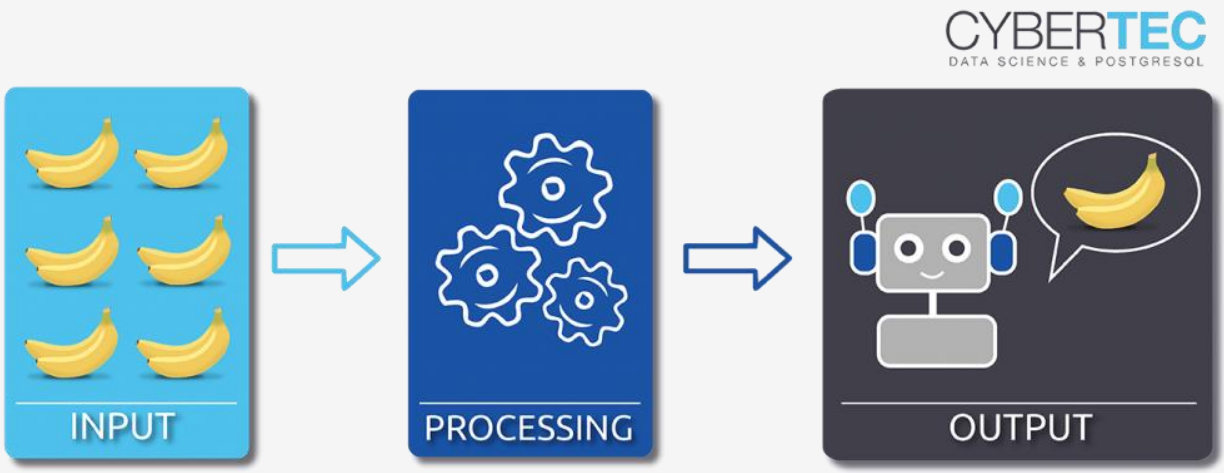
- sensor technology
- financial data
- insurance data
- control data
- geodata
- Healthcare…

Artificial Intelligence VS Machine Learning VS Deep Learning



Data Science: What are Artificial Intelligence, Machine Learning and Deep Learning?

Machine Learning



Machine Learning: How it works

CYBERTEC supports you in evaluating your database according to your wishes as well as your specifications and making it available for further processing. If required, we will take care of developing individual software that corresponds exactly to your application.

> „ [...] almost everything that has something to do with data:
> Collecting, analyzing, modeling.....
> yet the most important part is its applications --- all sorts of applications."
>
> Journal of Data Science, http://www.jds-online.com/about

Examples of Using Data Science in Industry and Research

In 2018, 70% of all companies were expecting the implementation of Artificial Intelligence and Machine Learning within the next twelve months. Of these, 20% assume that future business-critical decisions can be made automatically and 25% demand simple dashboards and interfaces for analyses and evaluations.

Financial sector:

- Minimization of errors and wrong decisions
- Fraud Detection: Automatic detection of any abnormal financial behavior

- Predictive analytics and market research (e.g. stock market prices)
- Chat bots and automatic assistance systems

Healthcare:

- Proactive identification of symptoms and diseases (e.g. cancer cells)
- Surgical interventions with robotic assistance systems that learn from historical data
- …

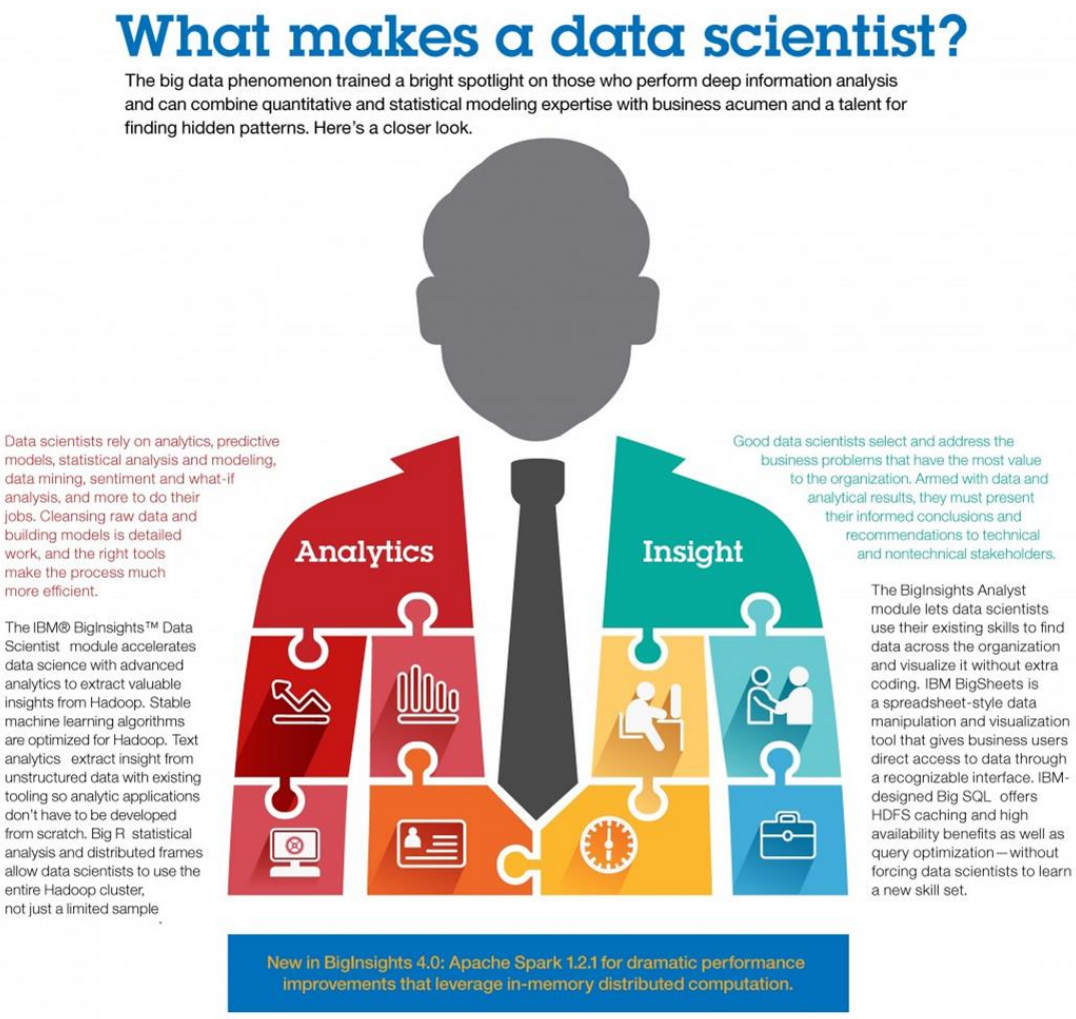
Trade and commercial providers:

- Improved recommendation systems
- Search by image recognition system instead of text input
- Automatic analysis of customers' mood and behavior
- Predictive merchandising and automatic price adjustment

Industry:

- Detailed sensorial evaluations – Predictive Maintenance
- Better decisions by learning historical data
- Improved automatic image and object recognition (e.g. autonomous driving systems)

What makes a data scientist?



What Is Data Science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract value from data. Data scientists combine a range of skills—including statistics, computer science, and business knowledge—to analyze data collected from the web, smart phones, customers, sensors, and other sources.

Data science reveals trends and produces insights that businesses can use to make better decisions and create more innovative products and services. Data is the bedrock of innovation, but its value comes from the information data scientists can glean from it and then act upon.

Tools for Data Scientists

Data scientists use many types of tools, but one of the most common is open source notebooks, which are web applications for writing and running code, visualizing data, and seeing the results—all in the same environment. Some of the most popular notebooks include Jupyter, RStudio, and Zepplin. Notebooks are very useful for conducting analysis but have their limitations when data scientists need to work as a team. Data science platforms emerged to solve this problem.

Data Science and the Growth of Data

What is big data? Learn more.

As modern technology has enabled the creation and storage of increasing amounts of information, the volume of data has soared. It's estimated that 90 percent of the data in the world was created in the last two years. For example, Facebook users upload 10 million photos every hour. The number of connected devices in the world—the Internet of Things (IoT)—is projected to grow to more than 75 billion by 2025.
The wealth of data being collected and stored by these technologies can bring transformative benefits to organizations and societies around the world, but only if we can interpret it. That's where data science comes in.

Explore this infographic for more information about data and data science.

The Emergence of the Data Scientist

As a specialty, data science is young. It grew out of the fields of statistical analysis and data mining. The *Data Science Journal* debuted in 2002, published by the International Council for Science: Committee on Data for Science and Technology. By 2008, the title of data scientist had emerged and the field quickly took off. There has been a shortage of data scientists ever since, even though more and more colleges and universities have started offering data science degrees.
A data scientist's duties can include developing strategies for analyzing data; preparing data for analysis; exploring, analyzing, and visualizing data; building models with data using programming languages such as Python and R; and deploying models into applications.

The data scientist doesn't work solo. In fact, the most effective data science is done in teams. In addition to a data scientist, this team might include a business analyst who defines the problem, a data engineer who prepares the data and how it is accessed, an IT architect who oversees the underlying processes and

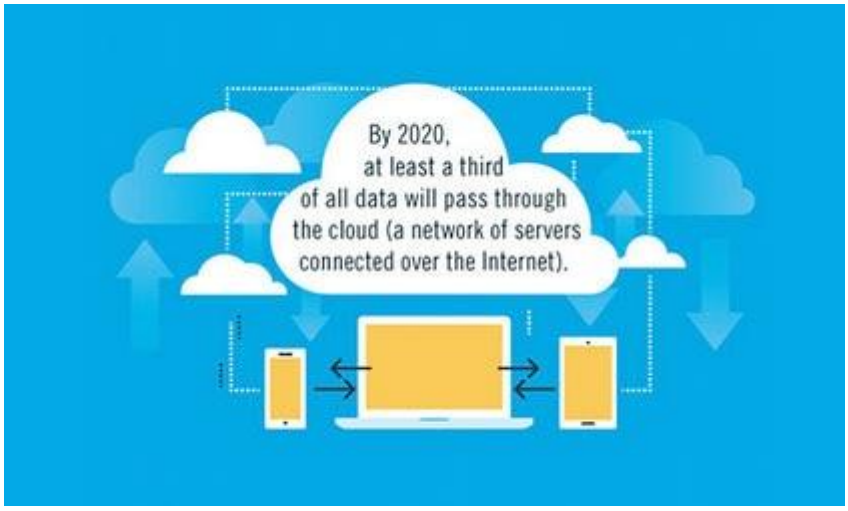
infrastructure, and an application developer who deploys the models or outputs of the analysis into applications and products.

How Data Science Is Transforming Business

Organizations are using data science teams to turn data into a competitive advantage by refining products and services. For example, companies analyze data collected from call centers to identify customers who are likely to churn, so marketing can take action to retain them. Logistics companies analyze traffic patterns, weather conditions, and other factors to improve delivery speeds and reduce costs. Healthcare companies analyze medical test data and reported symptoms to help doctors diagnose diseases earlier and treat them more effectively.

Most companies have made data science a priority and are investing in it heavily. In Gartner's recent survey of more than 3,000 CIOs, respondents ranked analytics and business intelligence as the top differentiating technology for their organizations. The CIOs surveyed see these technologies as the most strategic for their companies; therefore, they are attracting the most new investment.

How Data Science Is Conducted



The process of analyzing and acting upon data is iterative rather than linear, but this is how the work typically flows for a data modeling project:

- **Plan:** Define a project and its potential outputs
- **Prepare:** Build out the work environment, ensuring data scientists have the right tools as well as access to the right data and other resources such as compute power
- **Ingest:** Load the data into the work environment
- **Explore:** Analyze, explore, and visualize the data
- **Model:** Build, train, and validate models so they perform as required
- **Deploy:** Deploy models into production

Who Oversees the Data Science Process?

The data science process is typically overseen by three types of manager:

- **Business managers:** These managers work with the data science team to define the problem and develop a strategy for analysis. They may be the head of a line of business such as marketing, finance, or sales and have a data science team

reporting to them. They work closely with the data science and IT manager to ensure that projects are delivered.

- **IT managers:** Senior IT managers are responsible for the infrastructure planning and architecture that will support data science operations. They are continually monitoring operations and resource utilization to ensure that data science teams are operating efficiently and securely. They may also be responsible for building and updating environments for data science teams.
- **Data science managers:** These managers oversee the data science team and their day-to-day work. They are team builders who can balance team development with project planning and monitoring.

Challenges of Implementing Data Science

Despite the promise of data science and huge investments in data science teams, many companies are not realizing the full value of their data. In their race to hire talent and create data science programs, some companies have experienced inefficient team workflows, with different people using different tools and processes that don't work well together. Without more disciplined, central management, executives might not see a full return on their investments. This chaotic environment presents many challenges.

Data scientists can't work efficiently. Because access to data must be granted by an IT administrator, data scientists often have long waits for data and the resources they need to analyze it. Once they have access, the data science team might analyze the data using different and possibly incompatible tools. For example, a scientist might develop a model using the R language, but the application it will be used in is written in a

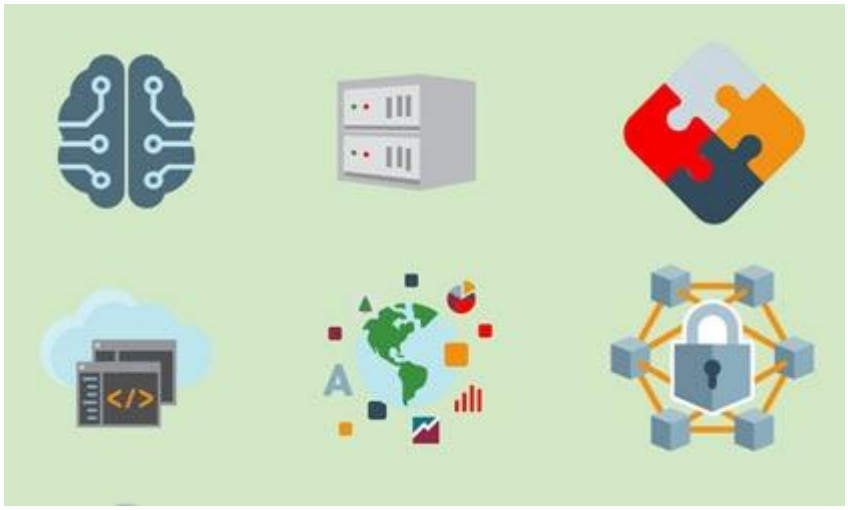
different language. Which is why it can take weeks—or even months—to deploy the models into useful applications.

Application developers can't access usable machine learning. Sometimes the machine learning models that developers receive must be recoded or are not ready to be deployed in applications. And because access points can be inflexible, models can't be deployed in all scenarios and scalability is left to the application developer.

IT administrators spend too much time on support. Because of the proliferation of open source tools, IT has an ever-growing list of tools to support. A data scientist in marketing, for example, might be using different tools than a data scientist in finance. Teams might also have different workflows, which means IT must continually rebuild and update environments.

Business managers are too removed from data science. Data science workflows are not always integrated into business decision-making processes and systems, making it difficult for business managers to collaborate knowledgably with data scientists. Without better integration, business managers find it difficult to understand why it takes so long to go from prototype to production—and they are less likely to back the investment in projects they perceive as too slow.

The Data Science Platform Emerges



Find out more about data science platforms.

Companies realized that without an integrated platform, data science work was inefficient, unsecure, and difficult to scale. This realization led to the emergence of data science platforms. These platforms are software hubs around which all data science work takes place. A good platform alleviates many of the challenges of implementing data science and helps businesses turn their data into insights faster and more efficiently.

With a centralized platform, data scientists can work in a collaborative environment using their favorite open source tools, with all their work synced by a version control system.

Make an impact with a data science platform that supports self-sufficiency.

The Benefits of a Data Science Platform

A data science platform reduces redundancy and drives innovation by allowing teams to share code, results, and reports. It removes bottlenecks in the flow of work by simplifying management and using open source tools, frameworks, and infrastructure.

For example, a data science platform might allow data scientists to deploy models as APIs, making it easy to integrate them into different applications. Data scientists can access tools, data, and infrastructure without having to wait for IT.

The demand for data science platforms has exploded in the market. In fact, the platform market is expected to grow at a compounded annual rate of more than 39 percent over the next few years and is projected to reach US$385 billion by 2025.
What a Data Scientist Needs in a Platform

Learn why Forrester positioned Oracle as a leader of notebook-based machine learning solution providers.

If you're ready to explore the capabilities of data science platforms, there are some key capabilities to consider:

- **Choose a project-based UI that encourages collaboration.** . The platform should empower people to work together on a model, from conception to final development. It should give each team member self-service access to data and resources.
- **Prioritize integration and flexibility.** Make sure the platform includes support for the latest open source tools; common version control providers such as GitHub, GitLab, and Bitbucket; and tight integration with other resources.
- **Include enterprise-grade capabilities.** Ensure the platform can scale with your business as your team grows. The platform should be highly available, have robust access controls, and support a large number of concurrent users.
- **Make data science more self-service.** Look for a platform that takes the burden off of IT and engineering and makes it easy for data scientists to spin up environments instantly, track all of their work, and easily deploy models into production.

Companies Struggle to Find Data Science Talent

Finding and recruiting talent is the biggest barrier that companies face when they want to use data science for competitive advantage. In a recent McKinsey & Company survey, half of executives across geographies and industries reported greater difficulty in recruiting analytical talent than any other kind of skill. Retention is also a problem according to 40 percent of those surveyed.

In addition to data scientists, McKinsey reports that there are shortages in other analytics categories. In particular, there are shortages of skilled workers who can translate between business

problems and the proper application of data science, and workers who are skilled at data visualization.

Indeed.com, Glassdoor, and Bloomberg provide further proof that there is significant demand for data science talent:

- Job postings for data scientists on Indeed.com rose 75 percent between January 2015 and January 2018. Job searches for data scientist roles rose 65 percent, according to Bloomberg.
- Glassdoor estimates that demand for data scientists in 2018 exceeded supply by 50 percent.
- Glassdoor rated data scientist as the top job in America—for the third year in a row.

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.[1][2] Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems".[3]

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[4] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[5][6] In 2015, the American Statistical Association identified database management, statistics and machine learning, and distributed and parallel systems as the three emerging foundational professional communities.[7]

In 2012, when Harvard Business Review called it "The Sexiest Job of the 21st Century",[8] the term "data science" became a buzzword. It is now often used interchangeably with earlier concepts like business analytics,[9] business

intelligence, predictive modeling, and statistics. Even the suggestion that data science is sexy was paraphrasing Hans Rosling, featured in a 2011 BBC documentary with the quote, "Statistics is now the sexiest subject around."[10] Nate Silver referred to data science as a sexed up term for statistics.[11] In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness."[12] While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents.[9] To the discredit of the discipline, however, many data-science and big-data projects fail to deliver useful results, often as a result of poor management and utilization of resources.[13][14][15][16]

History

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage, it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy".[17] In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

The modern definition of "data science" was first sketched during the second Japanese-French statistics symposium organized at the University of Montpellier II (France) in 1992[18]. The attendees acknowledged the emergence of a new

discipline with a specific focus on data from various origins, dimensions, types and structures. They shaped the contour of this new science based on established concepts and principles of statistics and data analysis with the extensive use of the increasing power of computer tools.

In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods"),[19] after the term was introduced in a roundtable discussion by Chikio Hayashi.[4]

In November 1997, C.F. Jeff Wu gave the inaugural lecture entitled "Statistics = Data Science?"[20] for his appointment to the H. C. Carver Professorship at the University of Michigan.[21] In this lecture, he characterized statistical work as a trilogy of data collection, data modeling and analysis, and decision making. In his conclusion, he initiated the modern, non-computer science, usage of the term "data science" and advocated that statistics be renamed data science and statisticians data scientists.[20] Later, he presented his lecture entitled "Statistics = Data Science?" as the first of his 1998 P.C. Mahalanobis Memorial Lectures.[22] These lectures honor Prasanta Chandra Mahalanobis, an Indian scientist and statistician and founder of the Indian Statistical Institute.

In 2001, William S. Cleveland introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" in his article "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," which was published in Volume 69, No. 1, of the April 2001 edition of the International

Statistical Review / Revue Internationale de Statistique.[23] In his report, Cleveland establishes six technical areas which he believed to encompass the field of data science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory.

In April 2002, the International Council for Science (ICSU): Committee on Data for Science and Technology (CODATA)[24] started the *Data Science Journal*,[25] a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.[26] Shortly thereafter, in January 2003, Columbia University began publishing *The Journal of Data Science*,[27] which provided a platform for all data workers to present their views and exchange ideas. The journal was largely devoted to the application of statistical methods and quantitative research. In 2005, The National Science Board published "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" defining data scientists as "the information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection" whose primary activity is to "conduct creative inquiry and analysis."[28]

Around 2007,[*citation needed*] Turing award winner Jim Gray envisioned "data-driven science" as a "fourth paradigm" of science that uses the computational analysis of large data as primary scientific method[5][6] and "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other."[29]

In the 2012 Harvard Business Review article "Data Scientist: The Sexiest Job of the 21st Century",[8] DJ Patil claims to have coined this term in 2008 with Jeff Hammerbacher to define their jobs at LinkedIn and Facebook, respectively. He asserts that a data scientist is "a new breed", and that a "shortage of data scientists is becoming a serious constraint in some sectors", but describes a much more business-oriented role.

In 2013, the IEEE Task Force on Data Science and Advanced Analytics[30] was launched. In 2013, the first "European Conference on Data Analysis (ECDA)" was organised in Luxembourg, establishing the European Association for Data Science (EuADS). The first international conference: IEEE International Conference on Data Science and Advanced Analytics was launched in 2014.[31] In 2014, General Assembly launched student-paid bootcamp and The Data Incubator launched a competitive free data science fellowship.[32] In 2014, the American Statistical Association section on Statistical Learning and Data Mining renamed its journal to "Statistical Analysis and Data Mining: The ASA Data Science Journal" and in 2016 changed its section name to "Statistical Learning and Data Science".[33] In 2015, the International Journal on Data Science and Analytics [34] was launched by Springer to publish original work on data science and big data analytics. In September 2015 the Gesellschaft für Klassifikation (GfKl) added to the name of the Society "Data Science Society" at the third ECDA conference at the University of Essex, Colchester, UK.

Relationship to statistics

"Data science" has recently become a popular term among business executives.[35] However, many critical academics and journalists see no distinction between data science and statistics, whereas others consider it largely a popular term for "data mining" and "big data".[3] Writing in Forbes, Gil Press argues that data science is a buzzword without a clear definition and has simply replaced "business analytics" in contexts such as graduate degree programs.[9] In the question-and-answer section of his keynote address at the Joint Statistical Meetings of American Statistical Association, noted applied statistician Nate Silver said, "I think data-scientist is a sexed up term for a statistician....Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician."[11]

On the other hand, responses to criticism are as numerous. In a 2014 Wall Street Journal article, Irving Wladawsky-Berger compares the data science enthusiasm with the dawn of computer science. He argues data science, like any other interdisciplinary field, employs methodologies and practices from across the academia and industry, but then it will morph them into a new discipline. He brings to attention the sharp criticisms of computer science, now a well respected academic discipline, had to once face.[36] Likewise, NYU Stern's Vasant Dhar, as do many other academic proponents of data science,[36] argues more specifically in December 2013 that data science is different from the existing practice of data analysis across all disciplines, which focuses only on explaining data sets. Data science seeks actionable and consistent pattern for predictive uses.[1] This practical

engineering goal takes data science beyond traditional analytics. Now the data in those disciplines and applied fields that lacked solid theories, like health science and social science, could be sought and utilized to generate powerful predictive models.[1]

In an effort similar to Dhar's, Stanford professor David Donoho, in September 2015, takes the proposition further by rejecting three simplistic and misleading definitions of data science in lieu of criticisms.[37] First, for Donoho, data science does not equate to big data, in that the size of the data set is not a criterion to distinguish data science and statistics.[37] Second, data science is not defined by the computing skills of sorting big data sets, in that these skills are already generally used for analyses across all disciplines.[37] Third, data science is a heavily applied field where academic programs right now do not sufficiently prepare data scientists for the jobs, in that many graduate programs misleadingly advertise their analytics and statistics training as the essence of a data science program.[37][38] As a statistician, Donoho, following many in his field, champions the broadening of learning scope in the form of data science,[37] like John Chambers who urges statisticians to adopt an inclusive concept of learning from data,[39] or like William Cleveland who urges to prioritize extracting from data applicable predictive tools over explanatory theories.[23] Together, these statisticians envision an increasingly inclusive applied field that grows out of traditional statistics and beyond.

For the future of data science, Donoho projects an ever-growing environment for open science where data sets used for academic publications are accessible to all researchers.[37] US National Institute of Health has already announced plans to enhance

reproducibility and transparency of research data.[40] Other big journals are likewise following suit.[41][42] This way, the future of data science not only exceeds the boundary of statistical theories in scale and methodology, but data science will revolutionize current academia and research paradigms.[37] As Donoho concludes, "the scope and impact of data science will continue to expand enormously in coming decades as scientific data and data about science itself become ubiquitously available."[37]

Data Scientist Jobs: Where does the big data talent gap lie?

Data science is one area of the digital sector that is desperately short of talent. In fact, IBM thinks data science will account for 28% of all digital jobs by 2020, but worryingly, the same report revealed that on average, each of these places remains unfilled for up to 45 days due to a lack of talent equipped with the necessary skills.

"Machine learning, big data and data science skills are the most challenging to recruit for, and can potentially create

the greatest disruption if not filled," according to IBM's The Quant Crunch report.

The Royal Society found that demand for workers with specialist data skills in the UK has more than tripled over the last five years to 231%, comparable to a general increase of regular workers of 36%.

And with the European Commission forecasting that 100,000 new data-related jobs will be created in the region by 2020, the fact there aren't enough people with the right skills to fill the role is certainly worrying.

What has triggered the data scientist skills gap?

Simply put, there aren't enough data scientists to go around - the demand for data analysis has grown exponentially over the last few years, and there aren't enough people being trained to meet the demand.

This growing gap between demand and available talent has meant that almost half of all European companies are thought to be struggling to fill their data scientist positions, according to recent data published by O'Reilly Media.

The report found that European organizations are heavily investing in products that help improve the accessibility and usability of data. Specifically, 59% are building platforms that allow for data integration and extract, transform and load (ETL) processes. Similarly, 53% said they were already working on data science platforms, while 53% said they were prioritizing data preparation and cleaning, something that is core to the development of AI and analytics services.

However, the lack of talent is holding back progress. As much as 47% of the region's organizations are currently struggling to fill their data science roles, with 37% also struggling with data engineering positions. That's compared to just 25% for security-based roles.

The findings are broadly in line with data published by Gartner, which found that organizations are not only unable to source talent for full-time positions, but also do not have enough data scientists consistently available throughout the business.

The amount of data businesses are generating compared to even just a few years ago is staggering, and it's growing fast, with Forbes stating that 90% of data present in the

world today was produced in the last two years alone. IDC thinks a mammoth 163 zettabytes of data a year will be generated by business by 2025.

There are more and more devices coming onto the market that scrape data, including wearables, smart home/office devices and that doesn't even include the rising number of people using the internet to shop, interact with brands and more, generating business insights that are powering business decisions.

Although the amount of data is impressive and is a positive thing, what isn't so positive is that data is useless without it first being analyzed and then used to inform business transformation. Without the manpower to work out what all of the information means, analytical insights which add value to the business can't be deduced. Rather, if an organization lacks the tools to coerce information from harnessed data, they may even be better placed abolishing data initiatives and cost-saving, to a degree. Simply, an investment in data tools must be matched by an investment in personnel.

How can data scientists benefit businesses?

Data scientists transform masses of data into analytical insights which help businesses make faster, more informed decisions.

Approximately 80% of the data generated over the next five years will be unstructured, highlighting the need for analytical tools operated by skilled personnel in order to extract insights. Using such tools, data scientists drag numbers and statistics into tables to form predictive models that can simulate a variety of possibilities.

Armed with the knowledge of the potential outcomes that each course of action can result in, businesses can select whichever solution aligns best with their initiatives. Logical, best-case scenario actions can be prescribed that improve overall performance. Moreover, as businesses build a record of performance metrics, an internal database of business-insights is formed, with which businesses are able to base decisions on recurring trends.

Boiled-down, the skill-set of a data scientist consists primarily of mathematics and statistics, programming knowledge and analytical thinking. Automated programs

can replicate these skills in the modern day; however a significant advantage of the data scientist is that they can think creatively. In this emerging field, it is certain that data science can be used in presently undiscovered ways to create further insight.

The possibilities are endless. A creative data scientist may forge new methods of gathering, interpreting and analyzing to produce a profitable data strategy. According to Bernard Marr, "the corporate data superstars of the future will be people who can find new data to solve business problems and come up with new and innovative methods of applying data statistics".

In this, it can be said that the impact of data science is only as good as the data scientist. And with the skills gap only widening, sourcing such talent comes at a cost.

What is industry doing to help close the skills gap?

The European Commission has made a commitment to try and fill the data scientist skills gap across the whole of Europe by running sessions exploring what businesses can do to try and make the problem less severe, addressing the demands of organizations and using the

data they collect to gain key insights about their customers and how to grow their business.

Businesses are also starting to react to the country's data scientist shortage and are collaborating with other firms and educational establishments to try and close the gap before it becomes too large to manage.

For example, data management and analytics firm SAS has teamed up with HSBC and the Data Lab to introduce an MSc course in Data Science for Business. The course will run at the University of Stirling and has been designed for those looking to start their career in data analytics. It teaches students how to use advanced analytics and apply these skills to real-life scenarios.

"There is a shortage of graduates emerging with the skills to apply the technical aspects of data science and use analytics to make sound business decisions," Dr Kepa Mendibil, course leader of the MSc in Data Science at the University of Stirling's School of Management said.

"Through this course, we have focused on the practical challenges that organizations are experiencing by merging disciplines to develop a teaching Program-me that makes

the link between business, management and data analytics."

The Office of National Statistics is also innovating to ensure the country has the resources it needs, collaborating with universities across the country to implement data science courses, with both MSc and short courses.

Although closing the data scientist skills gap is not something that can be solved immediately, it's important businesses and the government work together to ensure they are making steps towards creating a better-equipped workforce.

The Royal Society's report recommends a fundamental change in education. To provide young people with the necessary platform to pursue data-specialist roles, transferable skills such as communication and problem-solving should be embedded into the national curriculum. And to ensure the skills gap does not completely spiral out of control, the change is needed within the next ten years.

A two-pronged approach is needed - teaching young people about the merits and opportunities of data -

starting as early as possible and training existing staff how to analyze data to ensure the true meaning of information companies' hold isn't simply lost forever.

What does the data scientist role entail?

While data scientists are in high demand, the role can differ quite substantially from company to company, explains data scientist Jessica Kirkpatrick, at tech recruitment firm Hired.

"In simple terms, data science is taking the scientific method and applying it to a business' data set to derive insights that help the company make strategic decisions," she says. "However, that can mean a wide variety of things in terms of day-to-day duties."

While one data scientist might be tasked with taking raw data from an application or website, for instance and organizing it in a database in a way that makes it easy to analyze, another might find themselves running A/B tests on a website to see which version performs better with users.

"For example at Hired, we might pilot a new interview scheduling tool with a small group of clients and see how

this new feature affects their ability to hire candidates," Kirkpatrick adds.

"We then compare how efficiently this test group is able to hire candidates with the rest of the clients who do not have access to the tool. This allows us to understand if the tool helps our clients or not and if we should add it as a permanent feature to our product."

But the role is increasingly focused not just on the data, but on company strategy, Kirkpatrick explains, helping executives identify the most efficient areas of the business and which areas require extra resources, or where companies should scale down.

"We have been seeing an increase in [the number of] Chief Data Officer Roles in recent years, because companies are realizing that it's important to have someone with deep analytical knowledge of the business in the boardroom," Kirkpatrick says.

Chapter 8:

DATA ENGINEER

As the field of data science matures, a distinct specialization is emerging: **data engineering**. Tech giants like Facebook, Amazon, and Google recognize the value of data engineers relative to data scientists. That's why they're targeting candidates with skills to build critical infrastructure like data pipelines and warehouses.

The best computer science degrees keep up with this trend by helping graduate students develop high-level data engineering skills. The University of Illinois Online Master of Computer Science in Data Science (MCS-DS) degree is one of these degrees: it offers a comprehensive full-stack data sciences education to add this fast-growing career path to the employment opportunities of its graduates.

Why Data Engineering Matters

Data science is one of the most in-demand career fields in computer science according to LinkedIn, job openings increased 56% over the past year. These big data detectives unearth valuable insights through analysis of massive datasets. At the highest level, their skills are essential for developing machine learning algorithms and artificial intelligence (AI) applications.

However, in order to work their magic, data scientists need data. And not just any data — they need a *clean* dataset. That means they need raw and messy data converted to a consistent format that can be used with the data scientist's analytic tools. As computer science students know well, this simple-sounding task becomes increasingly challenging and time-consuming as a dataset grows in scale. In fact, some data scientists spend as much as 80% of their time "wrangling" or "managing" data before it's ready to be analyzed.

That's where data engineering comes in. Data engineers evaluate, parse, and clean datasets, using programming languages like Python and R to build data pipelines and

warehouses. This infrastructure efficiently delivers clean datasets at scale for data science to produce big data products. The data engineer's specialized expertise becomes crucial to a company's success as it grows; a startup that can only afford to hire one data scientist might have no choice but to direct 80% of their hours to data engineering. This inefficiency becomes a crippling as the company scales up.

Just as data scientists and data engineers can sometimes be distinct roles within a company, top professionals in these fields can sometimes come from distinct educational backgrounds. While a data scientist typically might focus on math and statistical analysis, data engineers are often system thinkers and programmers at heart. As the data industry continues to develop, it's becoming apparent that specializing in data engineering early in your education is a significant advantage for your career.

A Data Science Education for Data Engineers

The University of Illinois is one of the top-ranked computer science schools in the country, with an incredible history of pioneering research dating back to the 1940s. Its Online Masters of Computer Science in Data Science (MCS-DS) degree provides a top-tier

advanced education for data scientists, and students that want to pursue a career in data engineering have the flexibility to choose courses that prepare them for that direction. The degree requires graduate-level coursework in data mining, cloud computing, data visualization and machine learning for all students, through courses such as Introduction to Data Mining, taught by Jiawei Han, author of the well-known textbook “Data Mining: Concepts and Techniques.”

Computer science graduate students that want to become data engineers have plenty of opportunities to dive deeper as they fulfill their degree requirements. Advanced courses include data engineering-centric classes such as Theory and Practice of Data Cleaning and Data Creation.

MCS-DS students with a data engineering focus can complete their education with a Data Mining Capstone Project. In this hands-on course, students learn the latest data mining research techniques in an online seminar. They also complete a major project that applies data mining techniques to solve a real-world challenge. This kind of interactive, face-to-face learning experience puts the MCS-DS on a higher level than typical online computer science programs.

Getting Your Degree from an Industry Leader

The University of Illinois MCS-DS degree gives you excellent preparation for a data engineering career. Illinois' highly-ranked computer science program is known for its track record of excellence, and Illinois alumni and faculty are responsible for companies that have created entirely new industries.

Want to be a part of this legacy? Learn more about the University of Illinois MCS-DS degree and gain access to the computational and statistical knowledge needed to turn big data into meaningful insights

Who Is a Data Engineer & How to Become a Data Engineer?

A simple guide on how to ride the waves of Data Engineering and not let them pull you under.

It seems like these days everybody wants to be a Data Scientist. But what about Data Engineering? In its heart, it is a hybrid of sorts between a data analyst and a data scientist; Data Engineer is typically in charge of managing data workflows, pipelines, and ETL processes. In view of these important functions, it is the next hottest buzzword nowadays that is actively gaining momentum.

High salary and huge demand — this is only a small part of what makes this job to be hot, hot, hot! If you want to be such a hero, it's never too late to start learning. In this post, I have put together all the needed information to help you in taking the first steps.

So, let's get started!

What is Data Engineering?

Frankly speaking, there is no better explanation than this:

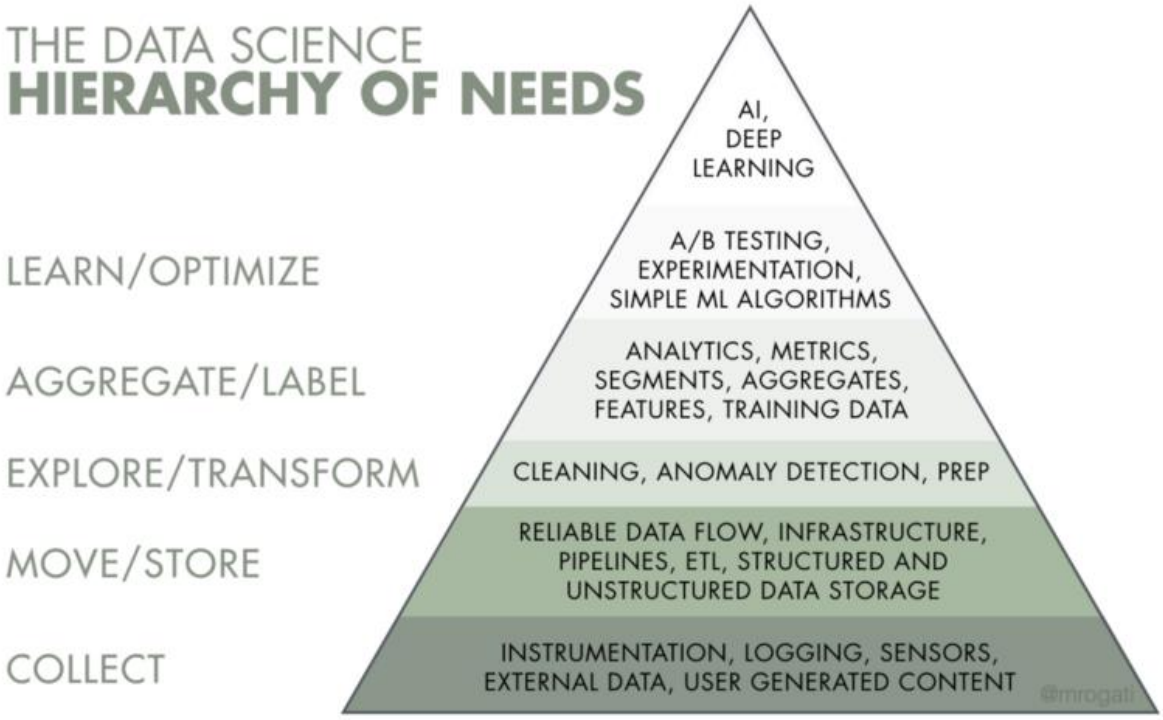
"A scientist can discover a new star, but he cannot make one. He would

have to ask an engineer to do it for him.”
–Gordon Lindsay Glegg

So, the role of the Data Engineer is really valuable.

It follows from the title the data engineering is associated with data, namely, their delivery, storage, and processing. Accordingly, the main task of engineers is to provide a reliable infrastructure for data. If we look at the AI Hierarchy of Needs, data engineering takes the first 2–3 stages in it: **Collect, Move & Store and Data Preparation.**

Hence, for any data-driven organization, it is vital to employ data engineer to be on the top.



What does a data engineer do?

With the advent of "big data," the area of responsibility has changed dramatically. If earlier these experts wrote large SQL queries and overtook data using tools such as Informatica ETL, Pentaho ETL, Talend, now the requirements for data engineers have advanced.

Most companies with open positions for the Data Engineer role have the following requirements:

- Excellent knowledge of SQL and Python

- Experience with cloud platforms, in particular, Amazon Web Services
- Preferred knowledge of Java / Scala
- Good understanding of SQL and NoSQL databases (data modeling, data warehousing)

Keep in mind, it's only essentials. From this list, we can assume the data engineers are specialists from the field of software engineering and backend development.

For example, if a company starts generating a large amount of data from different sources, your task, as a Data Engineer, is to organize the collection of information, it's processing and storage.

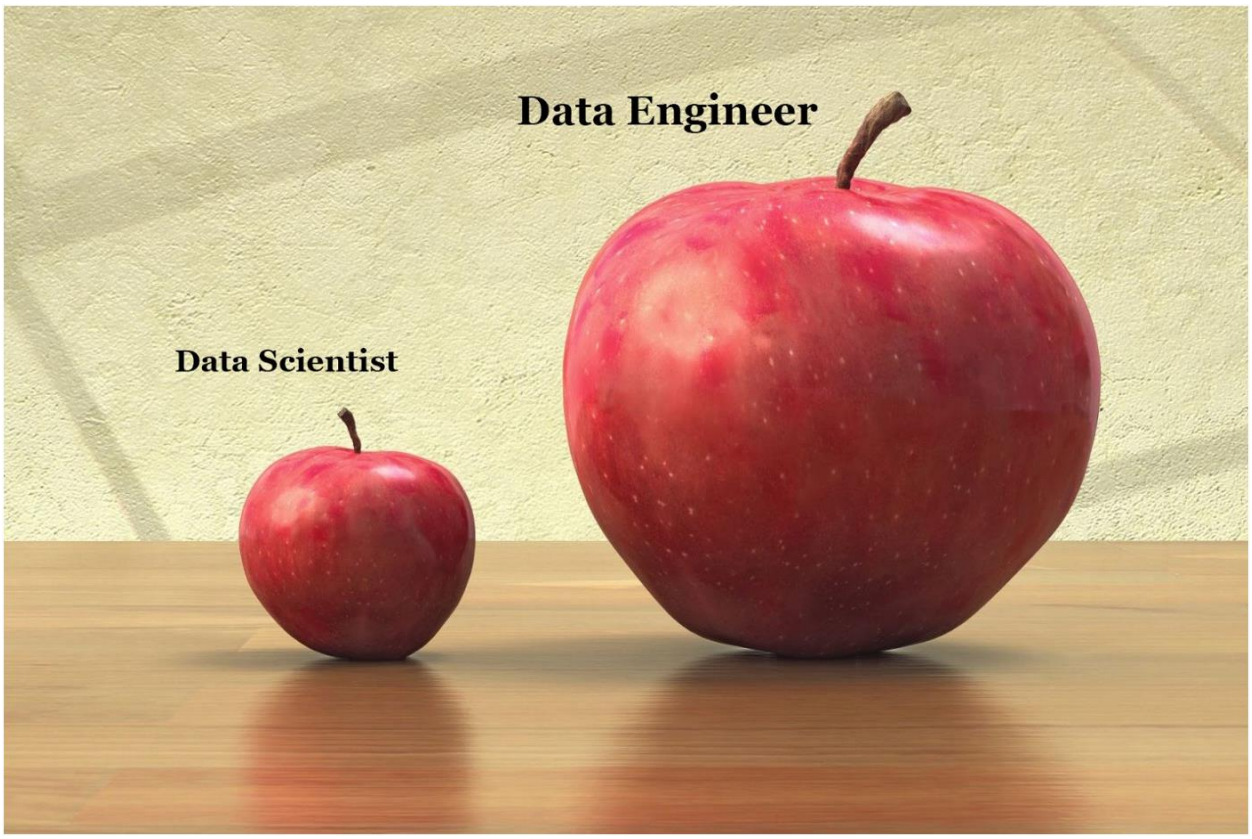
The list of tools used in this case may differ; everything depends on the volume of this data, the speed of their arrival and heterogeneity. Majority of companies have no big data at all, therefore, as a centralized repository that is so-called Data Warehouse, you can use SQL database (PostgreSQL, MySQL, etc.) with a small number of scripts that drive data into the repository.

IT giants like Google, Amazon, Facebook or Dropbox have higher requirements:

- Knowledge of Python, Java or Scala
- Experience with big data: Hadoop, Spark, Kafka
- Knowledge of algorithms and data structures
- Understanding the basics of distributed systems
- Experience with data visualization tools like Tableau or ElasticSearch will be a big plus

That is, there is clearly a bias in the big data, namely their processing under high loads. These companies have increased requirements for system resiliency.

Data Engineers vs. Data Scientists



Okay, that was a simple and humorous explanation (nothing personal), but the truth is, everything is much more complicated.

You should know first there is really much ambiguity between data science and data engineer roles and skills. So, you can easily get baffled about what skills are essentially required to be a successful data engineer. Of course, there are certain skills that overlap for both the roles. But also, there is a whole slew of diametrically different skills.

Data science is a real thing — but the world is moving to a functional data science world where practitioners can do their own analytics. You need data engineers, more than data scientists, to enable the data pipelines and integrated data structures.

Is a data engineer more in demand than data scientists?

- Yes, because before you can make a carrot cake you first need to harvest, clean and store the carrots!

Data Engineer understands programming better than any data scientist, but when it comes to statistics, everything is exactly the opposite.

But here the advantage of the data engineer:

without him/her, the value of this prototype model, most often consisting of a piece of code in a Python file of terrible quality, which came from a data scientist and somehow gives result is tending toward zero.

Without the data engineer, this code will never become a project and no business problem will be solved effectively. A data engineer is trying to turn this into a product.

Essential Things Data Engineer Should Know

So, if this job sparks a light in you and you are full of enthusiasm, you can learn it, you can master all the needed skills and became a real data

engineering rock-star. And, yes, you can do it even without programming or other tech backgrounds. It's hard, but it's possible!

What are the first steps?

You should have a general understanding of what is what.

First of all, Data Engineering is primarily related to computer science. To be more specific, you should have an understanding of efficient algorithms and data structures. Secondly, since data engineers deal with data, an understanding of the operation of databases and the structures underlying them is a necessity.

For example, the usual B-tree SQL databases are based on the B-Tree structure and in the modern distributed repositories LSM-Tree and other hash table modifications.

**These steps are based on a great article by Adil Khashtamov. So, if you know Russian, please support this writer and read his post too.*

1. Algorithms and Data Structures

Using the right data structure can drastically improve the performance of an algorithm. Ideally, we should all learn data structures and algorithms in our schools, but it's rarely ever covered. Anyway, it's never too late.

So, here are my favorite free courses to learn data structures and algorithms:

- Easy to Advanced Data Structures (Udemy)
- Algorithms, Part I (Coursera)
- Algorithms, Part II (Coursera)

Plus, do not forget about the classic work on the algorithms by Thomas Cormen — Introduction to Algorithms. This is the perfect reference when you need to refresh your memory.

- To improve your skills, use Leetcode.

You can also dive into the world of the database by dint of awesome videos by Carnegie Mellon University on Youtube:

- Intro to Database Systems
- Advanced Database Systems

2. Learn SQL

Our whole life is data. And in order to extract this data from the database, you need to “speak” with it in the same language.

SQL (Structured Query Language) is the lingua franca in the data area. No matter what anyone says, SQL lives, it is alive and will live for a very long time.

If you have been in development for a long time, you probably noticed that rumors about the imminent death of SQL appear periodically. The language was developed in the early 70s and is still wildly popular among analysts, developers, and just enthusiasts.

There is nothing to do without SQL knowledge in data engineering since you will inevitably have to construct queries to extract data. All modern big data warehouse support SQL:

- Amazon Redshift
- HP Vertica
- Oracle
- SQL Server

...and many others.

To analyze a large layer of data stored in distributed systems like HDFS, SQL engines were invented: Apache Hive, Impala, etc. See, no going anywhere.

How to learn SQL? Just do it on practice.

For this purpose, I would recommend getting acquainted with an excellent tutorial, which is free by the way, from Mode Analytics.

- Intermediate SQL
- Joining Data in SQL

A distinctive feature of these courses is the presence of an interactive environment where you can write and execute SQL queries directly in the browser. The Modern SQL resource will not be superfluous. And you can apply this knowledge on Leetcode tasks in the Databases section.

3. Programming in Python and Java / Scala

Why it is worth learning the Python programming language, I already wrote in the article Python vs R. Choosing the Best Tool for AI, ML & Data Science. As for Java and Scala, most of the tools for storing and processing huge amounts of data are written in these languages. For example:

- Apache Kafka (Scala)
- Hadoop, HDFS (Java)
- Apache Spark (Scala)
- Apache Cassandra (Java)
- HBase (Java)
- Apache Hive (Java)

To understand how these tools work you need to know the languages in which they are written. The functional approach of Scala allows you to effectively solve problems of parallel data processing. Python, unfortunately, cannot boast of speed and parallel processing. On the whole, knowledge of several languages and programming paradigms has a good effect on the breadth of approaches to solving problems.

For plunging into the Scala language, you can read Programming in Scala by the author of the language. Also, the company Twitter has published a good introductory guide — Scala School.

As for Python, I consider Fluent Python to be the best intermediate level book.

4. Big Data Tools

Here is a list of the most popular tools in the big data world:

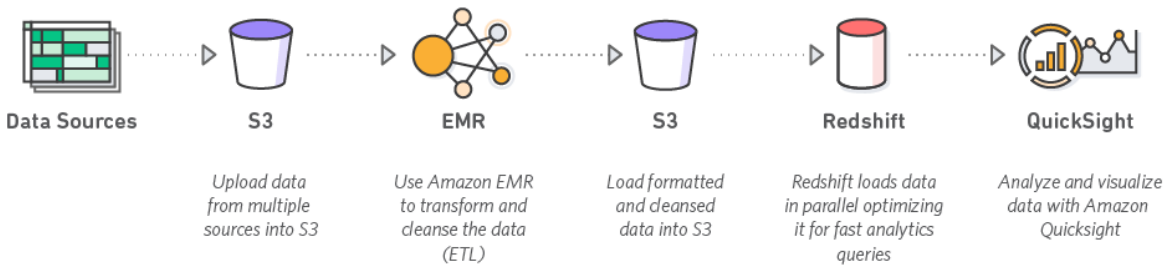
- Apache Spark
- Apache Kafka
- Apache Hadoop (HDFS, HBase, Hive)

- Apache Cassandra

More information on big data building blocks you can find in this awesome interactive environment. The most popular tools are Spark and Kafka. They are definitely worth exploring, preferably understanding how they work from the inside. Jay Kreps (co-author Kafka) in 2013 published a monumental work of The Log: What every software engineer should know about real-time data's unifying abstraction, core ideas from this boob, by the way, was used for the creation of Apache Kafka.

- An introduction to Hadoop can be A Complete Guide to Mastering Hadoop (free).
- The most comprehensive guide to Apache Spark for me is Spark: The Definitive Guide.

5. Cloud Platforms



Knowledge of at least one cloud platform is in the nest requirements for the position of Data Engineer. Employers give preference to Amazon Web Services, in the second place is the Google Cloud Platform, and ends with the top three Microsoft Azure leaders.

You should be well-oriented in Amazon EC2, AWS Lambda, Amazon S3, DynamoDB.

6. Distributed Systems

Working with big data implies the presence of clusters of independently working computers, the communication between which takes place over the network. The larger the cluster, the greater the likelihood of failure of its member nodes. To become a cool data expert, you need to understand the problems and existing solutions for distributed systems. This area is old and complex.

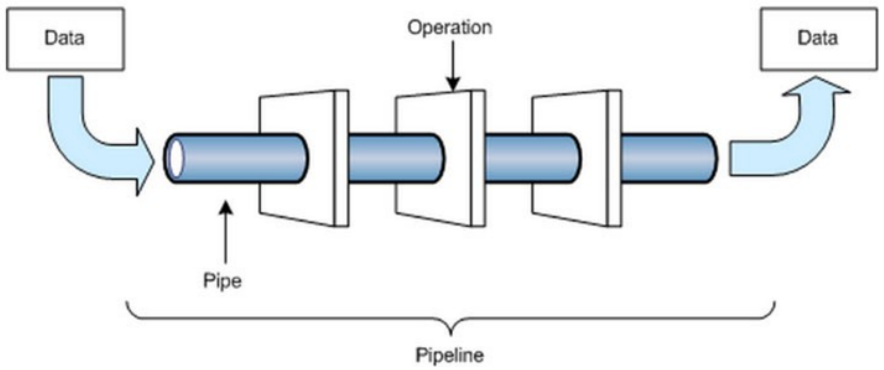
Andrew Tanenbaum is considered to be a pioneer in this realm. For those who don't afraid theory, I recommend his book Distributed Systems, for

beginners it may seem difficult, but it will really help you to brush your skills up.

I consider Designing Data-Intensive Applications from Martin Kleppmann to be the best introductory book. By the way, Martin has a wonderful blog. His work will help to systematize knowledge about building a modern infrastructure for storing and processing big data.

For those who like watching videos, there is a course Distributed Computer Systems on Youtube.

7. Data Pipelines



Data pipelines are something you can't live without as a Data Engineer.

Much of the time data engineer builds a so-called. Pipeline date, that is, builds the process of delivering data from one place to another. These can be custom scripts that go to the external service API or make a SQL query, enrich the data and put it into centralized storage (data warehouse) or storage of unstructured data (data lakes).

Summing it up: the Ultimate Data Engineer Checklist

Resuming, take a good understanding of:

- Take a good understanding of information systems;

- Software Engineering (Agile, DevOps, Design Techniques, SOA)
- Distributed Systems and Parallel Programming;
- Database Fundamentals — Plan, Design, Operation, and Troubleshooting;
- Experiments design — A/B test for putting into “Proof of Concept” systems reliability, performance, as well as designing good streamlines for delivering good solutions “on-the-fly”

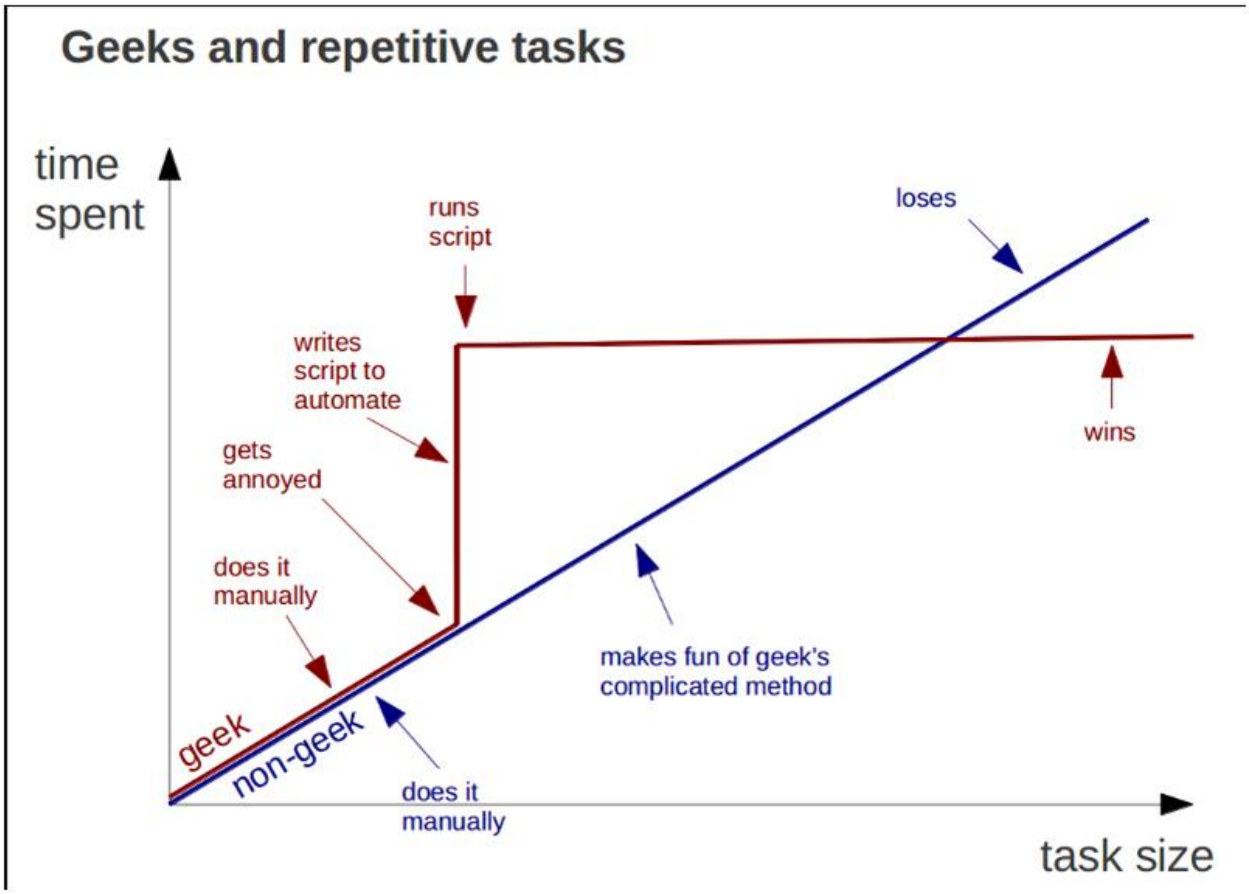
This is some of few requirements to become a data engineer, so, as much your study and can understand about data systems, information systems, continuous delivery/deployment/integration, programming languages and other topics on computer science (not all domain areas ofc), the better you gather your skill set.

And finally, the last but very important thing I want to say:

“The journey of becoming Data Engineering is not as easy as it might seem. It is unforgiving, frustrating and you have to be ready for this. Some moments on this journey will push you to throw everything in the towel. But, this is a true work and learning process”!

Just doesn’t sugar coat it right from the beginning? The whole point of the journey is to learn as much as you can and be prepared for new challenges.

Here’s a great visual I came across that illustrates this point really well:



Hey Guys,

And yes, don't forget to avoid burnout and get rest. It is also very essential. Good luck, Wish you all success!

Chapter 9:

American Universities and Institutions Certified You in the Major of "Big Data"!

BEST BIG DAT DEGREES

Remember the movie *The Social Network*?

Yes, it's about Facebook. Yes, it's about Mark Zuckerberg, who is currently one of the ten wealthiest people in the world. And yes, it gets ugly at times. Underneath the drama, however, was a very simple but profound insight: Data can change the world.

A better way to put that insight might be: *what we do with data* **can change the world . . . for better or worse!** In the case of Facebook, it had some very humble beginnings involving

writing an algorithm that used existing data to create an entirely new way of interacting with others in the digital age.

What kind of Big Data degree should you get?

You can study this field in depth with a degree in big data. If you want the flexibility of an online course of study, take a look at our ranking of the Best Online Big Data Degrees.

But that's just one example. Imagine how much other data is out there about any number of things, and imagine the possibilities of what can be done with it all . . . yes, both good and bad! **We are in the infancy stages of even knowing how to collect and store all this data, much less how to interpret it and use it.**

Enter the desperate need for technical expertise in this rapidly growing area that has left no industry untouched. Schools are quickly mobilizing to create highly specific programs tailored to the needs we have now for professionals who know how to wrangle terabytes of information, interpret it, and explain to the rest of us how it effects a company's needs on the ground.

How much money can you make with a Big Data degree?

Many schools have responded to this need through Master of Science degrees in a variety of different areas–some in business analytics, some in data mining, and others still in analytics.

But not only is the need for big data analyst in demand, and exponentially so, the field is incredibly lucrative. PayScale, for example, places the average salary for Data Scientists above $96,000 a year. But for Directors of Analytics that number leaps way above $134,000 annually.

What can you do with a Big Data degree?

Maybe Facebook changed the world "for better," maybe it didn't, or maybe it's a bit of both. Whatever your opinion happens to be, there's no escaping the fact that what Mark Zuckerberg did with some calculus chops and some basic data has completely and irreversibly changed the way we interact with each other.

So how will a degree in big data help you? One of these degrees will not only set you up for a lucrative, stable career. It will provide you access to one of the newest, most exciting industries. For example, graduates from this program have gone on to become:

- Statisticians
- Business Intelligence Consultants
- Data Analysts
- Data Miners and Data Engineers

What are the requirements for a Big Data degree?

They often go by different names, and many have different emphases, but what they all have in common is a **commitment to training professionals in big data; who can help us all figure out what to do with all the information at our disposal.** These programs ensure you graduate an expert in big data. They do that through rigorous curriculums as well as hands-on experience. Indeed, these programs require practices and internships

What are the best Big Data degrees?

So where are these programs? And where do you start? That's where we come in. We've compiled a list of the top 50 degrees in big data with your needs in mind. Some are online, some not, and some offer flexibility in this particular respect.

The first thing we want to know, like you, is the bottom line: **how much is this going to set me back?** So we've included the annual tuition for each school right up front for you. We've also included our College Choice Score, which is computed based on each school's reputation in the field and its return on investment.

The end result is a list of schools that balance cost, reputation, and, ultimately, what they can do for you and your unique goals.

The rankings you're about to read are based on a few important sources. The first source is actual college freshman polled during a nationwide survey published by the Higher Education Research Institute at UCLA.

These students rated academic reputation, financial aid offerings, overall cost of school, and the survey also took into account graduate success rates on the post-college job market. These factors were weighed equally alongside data from other publicly available sources, including U.S. News & World Report, the National Center for Education Statistics, and PayScale.com.

The Social Network was a big hit. Could a *Social Network 2* be in the works . . . and could it be about you?

2018 Ranking

#1

CARNEGIE MELLON UNIVERSITY

Overview

Carnegie Mellon has a long history of excellence in technology, engineering, and business education and research. It offers a Master of Science in Information Technology, which can be earned online with a specialization in Business Intelligence & Data Analytics. Working adults and traditional students alike can complete the program at their own pace through Carnegie's cutting-edge content management tools.

Features

Coursework in the program includes training in economics, business, finances and accounting, geographic systems, database management, statistics, digital privacy, data warehousing, exploring and visualizing data, and applied data mining. Carnegie's reputation precedes it, and its MS program in Information Technology with a specialization in Business Intelligence and Data Analytics is no exception. Carnegie's big data program emphasizes:

- Data analytics
- Management strategy
- and Information technology

Notables

CMU's academic community is comprised of Nobel Prize laureates, Turing Award winners, and American Association for the Advancement of Science fellows, and much more. As a whole, the University is ranked among the top 25 best schools in the nation. And its information technology department is often considered *the* best.

#2

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Overview

Missouri University of Science and Technology is one of the most respected STEM-focused public universities in the Midwest. While it is known for its engineering programs, it also offers recognized business programs and online education. Missouri S&T offers an online graduate certificate in Business Analytics

and Data Science, which can be applied toward a masters degree in the field for students who want to continue their education.

Features

Students will benefit from coursework in the areas of big data analytics, information visualization, business intelligence, cloud computing, machine learning, applied time series analysis, and regression analysis. Areas of concentration in Missouri S&T's program include:

- Business analytics and data science
- Big data and security
- or big data management and analytics

Notables

Missouri S&T University has several rankings and awards to its name. It is home to some of the "World's Coolest Labs." It is nationally and globally ranked. And its engineering degrees are among the top 10 nationally, according to *USA Today*.

#3

JOHNS HOPKINS UNIVERSITY

Overview

Johns Hopkins University is located in Baltimore, Maryland. It is considered the site of vibrant active research according to the Carnegie Foundation. JHU is strong in many areas, but is most known for its STEM and health programs.

Featurse

Johns Hopkins University offers students interested in big data two options: an MS in Data Science and a post-master's certificate in Data Science. Both programs can be pursued completely or in part, which provides options for today's working professional. Emphases of the program include:

- Applied mathematics
- Statistics
- and Computer science

Notables

Through its programs, JHU trains students not just to manage data but to interpret and apply it. Its faculty sets a strong precedent for this, being among the leaders in this emerging field. On-site students can also benefit from JHU's state-of-the-art facilities. Coursework in the program includes the areas of data visualization, computer programming, mathematics, probability, statistics, and simulation.

#4

UNIVERSITY OF CALIFORNIA BERKELEY

Overview

Berkeley needs little introduction. It is the flagship institution of the UC system, and it has been one of the country's leaders in research and innovation for decades. Berkeley is home to the School of Information, one of the first and only centers of its kind.

Features

The School of Information offers an online Master of Information and Data Science program through which students have access to the best resources in the field as well as those of the social sciences, computer science, statistics, management, and law. The program is designed to train students in the following areas:

- Research design
- Cleansing
- Storage and retrieval
- Mining and exploring, and much more

Notables

While Berkeley's program is offered both on campus and online, online students do have a brief residency requirement as part of the program. Graduates of Berkeley's program have obtained employment in a wide variety of industries.

#5

VILLANOVA UNIVERSITY

Overview

Villanova offers a Master of Science in Analytics degree through the Villanova School of Business. The program's location in the Business School makes it a program with a practical emphasis, which will be appealing to students who are either already in the workforce or anticipate pursuing a career in business.

Features

The degree can be completed fully online, and its graduates are in high demand. Villanova's program is 33 credit hours, and its coursework includes training in everything from data models to structured analysis, multivariate data analysis to enterprise data management. Indeed, students can become experts in:

- Analytic methods for optimization and simulation
- Business intelligence
- Advanced applications
- And more

Notables

Villanova is regularly recognized for excellence in any number of areas. It is a top-ranked regional university, and it has developed a reputation in graduate programs, business, and online education.

#6

FAIRFIELD UNIVERSITY

Overview

Fairfield University is based in Fairfield, Connecticut. Fairfield offers a Master of Science in Business Analytics, a 30-credit degree that is no exception to the rule—graduates of the program are in high demand, and starting salaries are comparatively high.

Features

In the MS in Business analytics program, students will take courses in information systems, database management, data mining, business statistics and analytics, and forecasting and predictive analytics. Electives provide additional opportunities to pursue further interests. The program also offers a marketing analytics concentration, which includes coursework in:

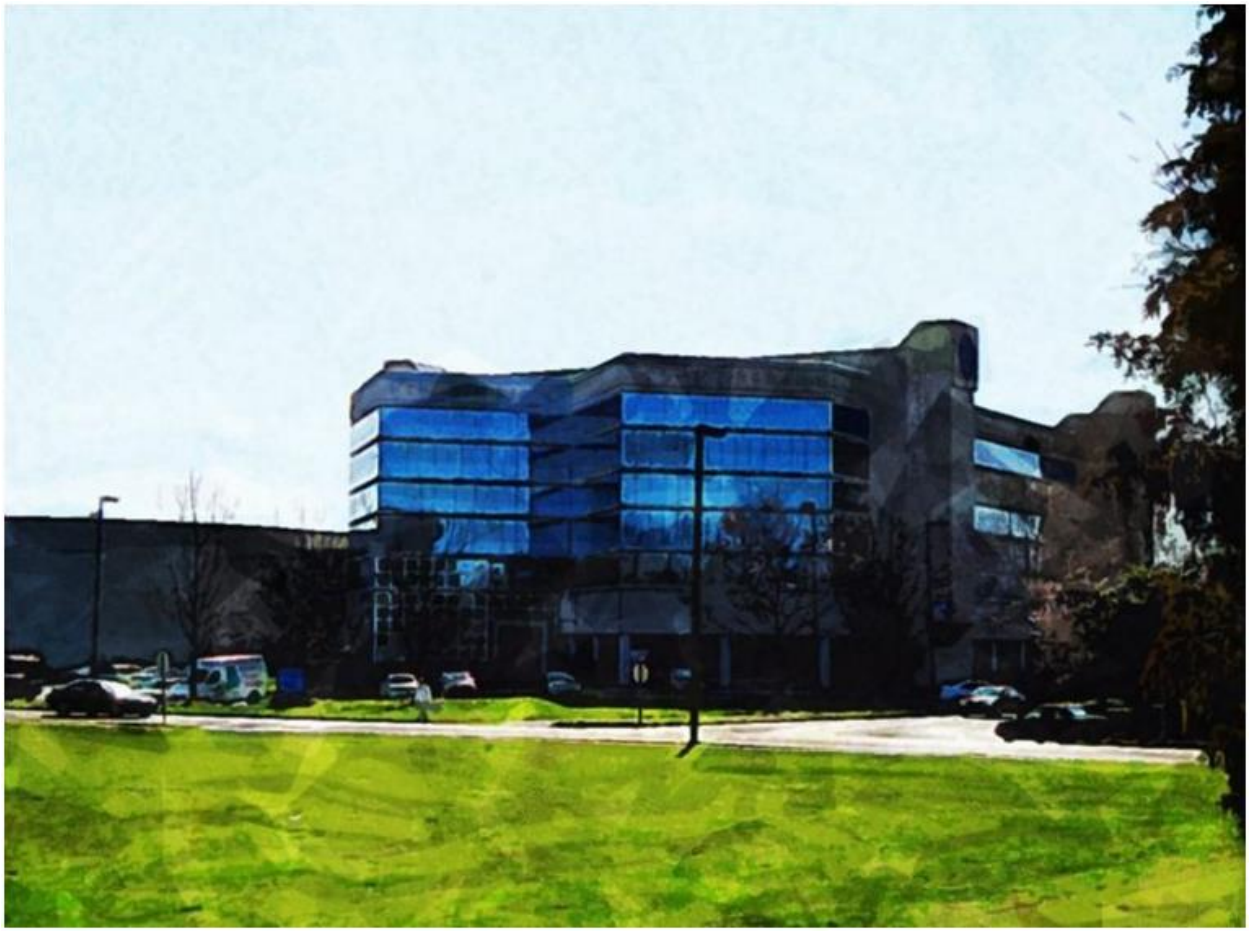
- Marketing management and research
- Multivariate data analysis for decision making
- Experimental research in marketing
- Contemporary topics in marketing.

Notables

Fairfield is one of the most well-respected regional schools in all of New England. Its graduates enjoy a high starting salary average, making it a good return on investment. It has a growing number of online programs, including data analytics.

#7

UNIVERSITY OF ALABAMA IN HUNTSVILLE



Overview

The University of Alabama in Huntsville is a major center for technology and research in Alabama and one of the region's leading schools. UAH offers a Master of Science in Management Science with a specialization in Business Analytics, which can be completed on campus or online.

Features

UAH's MSMS-BA provides training in advanced analytic tools and techniques to meet the needs of organizations. Students will receive training in detecting patterns and relationships in data, predicting trends, developing forecasts, and improving decision-making within a business environment. They will also learn skills such as data mining, predictive modeling, design of experiments, and statistical analysis. Examples of how these skills can be applied include the ability to:

- Better understand customer buying patterns
- Improve customer retention
- Assess the impact of marketing programs
- Manage new product launches, and more

Notables

University of Alabama in Huntsville has a long tradition of offering technical education in response to demand. Indeed, when a need arose for engineering education for the military, the University of Alabama system opened its Huntsville campus to meet the need.

#8

PENNSYLVANIA STATE UNIVERSITY

Overview

Penn State's programs in big data draw on the resources of four colleges and departments for the best, broadest, and most thorough education in big data mining, analysis, and communication possible. It offers three programs in analytics: a general Master in Data Analytics, a graduate certificate in business analytics, and a Master of Professional Studies in Business Analytics.

Features

Penn State places an emphasis on data mining, modeling, and architecture as well as extraction, transformation, and loading (ETL) development. Its program can be completed online or residentially, and it is one of the first to offer training from multiple departments and schools. Students can expect to receive training in the areas of:

- Applied statistics
- Predictive analysis
- Data-driven decision making
- Database design, and more

Notables

Penn State remains one of the world's undisputed leaders in online education, having been one of the original institutions to establish online offerings nearly twenty years ago. It is also one of the largest, with tens of thousands of enrolled distance students.

#9

BOSTON UNIVERSITY

Overview

Boston University is a nationally ranked university. It has a growing offering of online programs through its Metropolitan College Online, a high-quality distance education system that provides degree programs for working adults. Through this college, BU offers a Master of Science in Computer Information systems, which focuses on data mining, analysis, and visualization.

Features

BU’s program is shorter than those of some schools, and it also seeks to strike a balance between theory and practice. In addition to its primary objectives of analytics, data analysis and visualization, web analytics and mining, and data mining, the program also offers training in:

- Applied probability and statistics
- Data visualization techniques
- Web analytics and metrics

Notables

Boston University is no stranger to our rankings or those of others. It has world-renowned programs in business, law, and medicine. Graduates of the program enjoy an excellent placement record in a wide variety of careers in computer information systems.

#10

OREGON STATE UNIVERSITY

Overview

Oregon State University is a large public institution known for its STEM programs. It has made great strides in recent years with its online offerings. It now offers an array of degrees, certifications, and professional programs online.

Features

For working professionals in data analytics, those looking to change careers, or students entering the field fresh, OSU offers two graduate options: a Master of Science or a Graduate Certificate in Data Analytics. The programs can be completed online through OSU's College of Science and Department of Statistics. Both programs are intended to train students in:

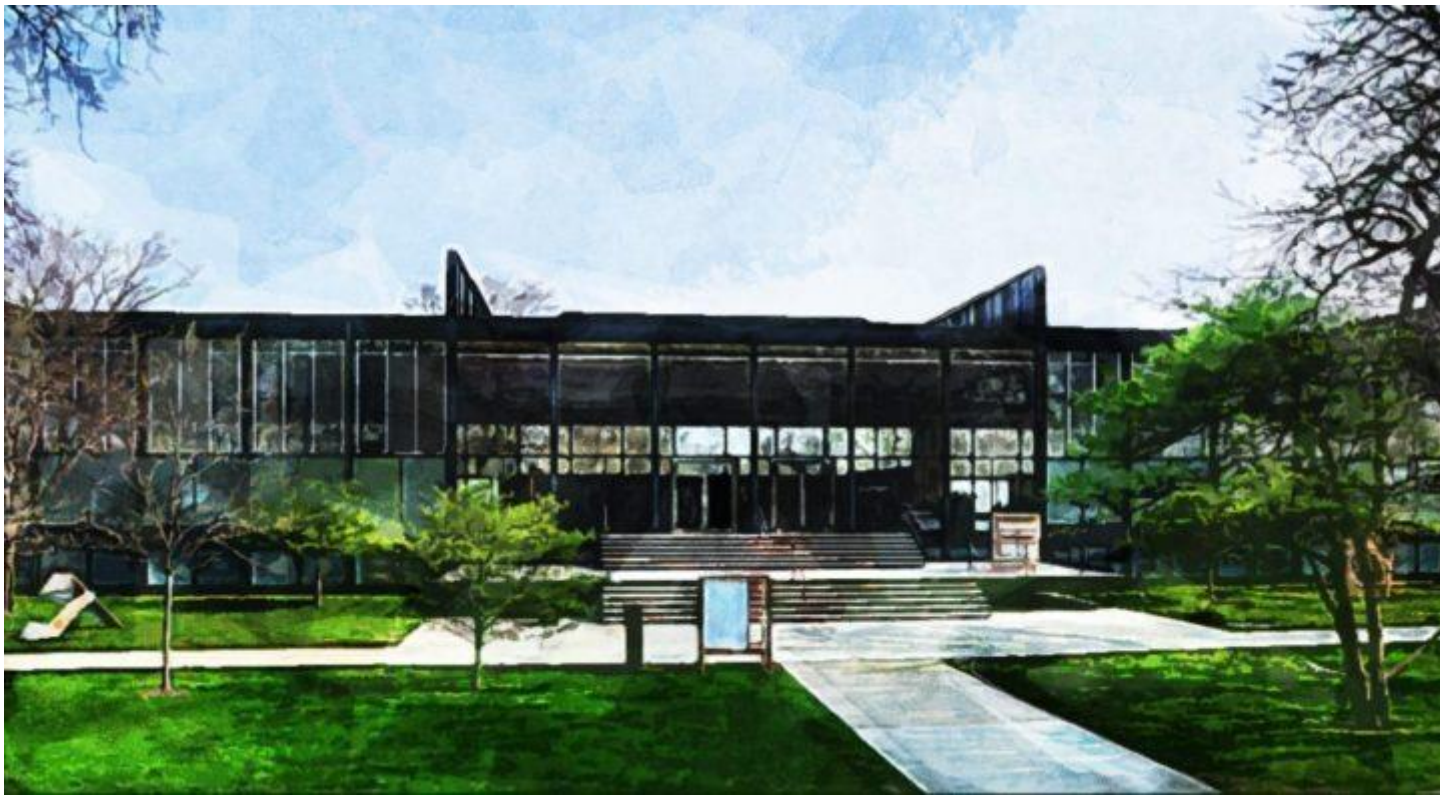
- Statistical and analytical skills
- Data analysis, interpretation, and visualization
- And real-world data problems

Notables

The program places an emphasis on quantitative methods, and students can expect to be well prepared for any number of scenarios in statistics, computer science, mathematics, policy, and applied sciences.

#11

ILLINOIS INSTITUTE OF TECHNOLOGY



Overview

Illinois Tech is based in Chicago, Illinois. It offers a Master of Data Science program that can be completed in sixteen months on a full-time basis or two years on a part-time basis. The program emphasizes the proper application of data research, and it prepares data professionals for responsible, ethical design-making through the study of mathematics, statistics, and computer science.

Features

Students will become experts in analyzing data, visualizing results, and articulating discoveries. The program also encourages critical thinking, communication, and good decision-making skills, which are essential to successful job performance. Students will learn:

- To approach data using the scientific method
- Question premises and think structurally
- And find real-world solutions to big data problems

Notables

Illinois Tech is one of the foremost polytechnic institutes in the Midwest. It was ranked among the top 25 “Most Rigorous” universities in the country. And it is, of course, considered one of the best as well.

#12

WEST VIRGINIA UNIVERSITY

Overview

West Virginia University is located in Morgantown, West Virginia. It has a long history of providing practical education to the residents of the state. It has recently expanded that mission to online offerings, including a Master of Science in Business Data Analytics through its College of Business and Economics.

Features

Classwork for this program includes the areas of business intelligence, ethics and data collection, business statistical methods, data mining, simulation modeling, visualization, decision science, management, and a business analytics practicum. Outcomes of the program include facility in:

- Statistical techniques
- Data mining strategies
- Databases
- and Sustainable and effective analytic tools

Notables

This degree can be completed in as little as 12 months and requires two brief residencies on campus. Because of the hybrid model, this is an ideal program for those who want the flexibility of online education as well as

#13

CITY UNIVERSITY OF NEW YORK

Overview

The School of Professional Studies at the City University of New York offers ten bachelor's degrees, eight bachelor's degrees, and numerous undergraduate and advanced certificates and professional non-degree programs. It is designed to provide online and on campus programs that meet the needs of adults who are looking for a seamless way to finish or transition into a bachelor's

degree, earn a master's degree or certificate in a specialized field, advance in the workplace, or change careers.

Features

The SPS offers an online Master of Science in Data Analytics. Through CUNY's program students will build portfolios of increasingly complex projects using popular programming languages such as R and Python, popular languages in IT. Students will also gain hands-on experience:

- Building predictive and prescriptive models
- Practicing presentations
- Collaborating with teams on research projects

Notables

The MS in Data Analytics program culminates with a capstone project that represents highly sophisticated, but practical, solutions to address real world problems. Though a huge requirement, this capstone project is a great way to impress future employers.

#14

UNIVERSITY OF WISCONSIN

Overview

The University of Wisconsin features a growing catalogue of online courses as well, and it remains one of the well-respected public universities in the country in areas as diverse as linguistics, business, and the sciences.

Features

UW offers a Master of Science in Data Science. Students are trained in statistical and predictive models for business analytics as well as forecasting strategic methods of management in finances and operations. UW students are also trained in:

- Business intelligence solutions
- Data warehousing
- Mining techniques

Notables

UW also offers a Graduate Certificate in Business Analytics, which requires fifteen credit hours. Electives in UW's program include business forecasting methods, web mining and analytics, marketing analytics, database marketing, and business intelligence technologies and solutions.

#15

REGIS UNIVERSITY

Overview

Regis University is a private Catholic institution located in Denver, Colorado. It is regularly featured in rankings of regional universities and online programs. Its College of Computer and

Information Sciences offers an online Master of Science in Data Science.

Features

This 36-credit graduate degree that brings together technical skills with the Jesuit commitment to social impact, challenging students to consider the implications of big data on business and society. It specializes in the areas of:

- Data analytics
- Data engineering
- Statistics
- Predictive analytics

Notables

A unique feature of Regis's program is that students are required to take courses in the ethics, privacy, and social justice in data science in addition to standard courses in the field. Students in Regis's program will be prepared to assemble data, identify trends and patterns, implement data to effectively communicate, analyze data to identify new areas of revenue, appraise data for ethical issues, investigate how data science affects society, and more.

#16

GEORGIA SOUTHERN UNIVERSITY

Overview

Georgia Southern University (GSU) is part of the wider University of Georgia system. GSU offers a Master of Science in Computer Science degrees with optional specializations in Data Mining or Data Warehousing.

Features

GSU's program is a 30-hour program, and it prepares students primarily for careers in the IT industry. The MSCS seeks to educate students in both theoretical and practical skills, for instance, both data mining and business application. Students can expect to be prepared in areas such as databases, artificial

intelligence, Web systems, data mining, data warehousing, and distributed database systems. Areas of practical application include:

- Software engineering
- Wireless and mobile systems

Notables

GSU has been garnering national attention in recent years for a number of initiatives, perhaps most significantly for achieving status as a national rather than regional institution. In addition to its academic accolades, GSU has been acknowledged by many rankings as “best value” school.

#17

TEXAS A&M UNIVERSITY

Overview

Texas A&M University is located in College Station, Texas. It needs little introduction, particularly in a list of rankings in STEM fields. It is not only one of the largest universities in the country but also one of the leaders in higher education in the sciences and technology.

Features

Coursework for the MSA is offered through the Mays Business School and welcomes working professionals with strong quantitative skills, for example bachelor's degree holders in the sciences, mathematics, business, and engineering. Students can expect to be prepared to serve in a number of careers, such as:

- Manufacturing,
- Natural resources
- Finance and insurance
- and many other sectors

Notables

The flagship school of the Texas A&M system, it has also expanded its online offerings in recent years. A&M offers a Master of Science in Analytics through its Department of Statistics. This program specifically is a 30-six hour degree, making it one of the more demanding programs on our list.

#18

UNIVERSITY OF ARKANSAS

Overview

The University of Arkansas is located in Fayetteville, Arkansas. It is home to the Walton College of Business, an increasingly well-respected fixture on the national management education landscape. Students interested in big data will be interested in two of U of A's programs: a twelve-hour online graduate certification in Business Analytics, or a Professional Master of Information Systems degree.

Features

Students who complete the graduate certificate can transfer to the MSIS program. Coursework for the program includes training in data management systems, decision support and analytics, and business intelligence and knowledge management, IT ERP fundamentals, e-business development, and/or systems development. Graduates have gone on to become:

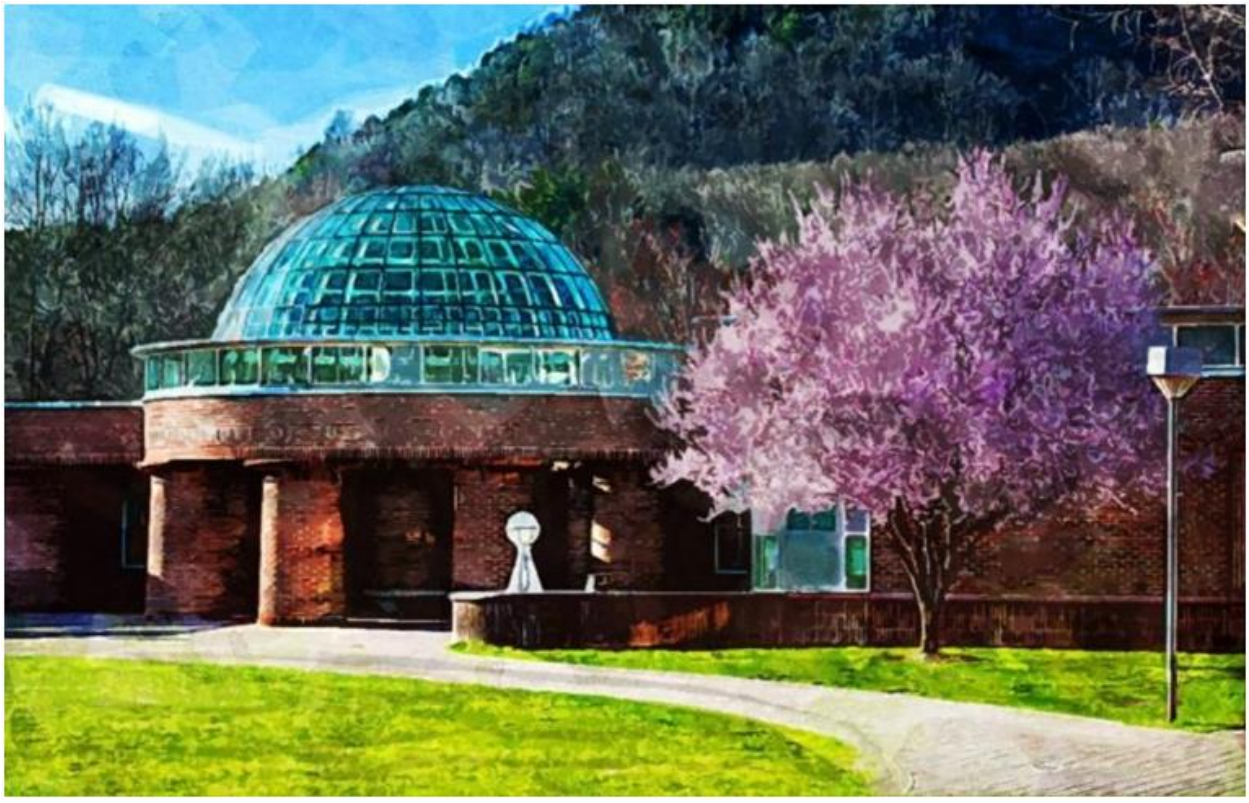
- Data Analysts
- Statisticians
- And much more

Notables

Arkansas has also increased its number of online offerings in recent years.An additional benefit of U of A’s program is its partnerships with companies such as IBM, Microsoft, SAP, and others, which offer students opportunities to acquire experience with problems real companies face.

#19

QUINNIPIAC UNIVERSITY



Overview

Located in Connecticut, Quinnipiac has come into the national spotlight in recent years by investing in online education. Quinnipiac offers a Master of Science in Business Analytics through its online campus, and it is designed with working professionals in mind.

Features

Quinnipiac is dedicated to turning its graduates into assets, and to this end, it offers coursework in a wide range of areas, such as data visualization, mining, management and warehousing, statistical analysis, information technology project management, and predictive modeling. Its graduates are in high demand in the data professions, and they have been placed in careers with:

- Marketing agencies
- Financial services
- Insurance
- Manufacturing, and more

Notables

Does the name “Quinnipiac” sound familiar? If so, you probably heard the school’s name during the 2016 U.S. Presidential race, during which Quinnipiac offered its expertise in data management through a number of well-publicized election polls.

#20

UNIVERSITY OF OKLAHOMA

Overview

This is the flagship school within the University of Oklahoma System. U of O has a long history of excellence as a large, public institution in areas as diverse as health care, business, and the arts. It is also known for being one of the first public institutions to incorporate digital technology into higher education.

Features

U of O offers a Master of Science degree in Data Science and Analytics, an intensive fourteen-month program offered through the School of Computer Science and the School of Industrial and Systems Engineering. Oklahoma's program integrates both computer science and industrial and systems engineering. Coursework in the program includes:

- Data analysis
- Data analytics
- Systems engineering

Notables

Oklahoma's program is designed to be flexible—it offers both online and on-campus options as well as full- and part-time enrollment, and students can opt for one of two completion tracks: coursework only or research-thesis.

#21

AUBURN UNIVERSITY

Overview

Auburn University has a storied history both in the South and, more recently, as a national-level research university that has attracted attention for its design, engineering, and business programs. Auburn offers a Master of Science in Information Systems through its Harbert College of Business.

Features

This is a 30-credit degree that places a heavy emphasis on information technology. Graduates of Auburn's program can expect to be prepared in an array of areas, from business intelligence and analytics to database analysis. Further, you will gain experience in:

- Administration
- Systems architecture
- Information security
- Software engineering

Notables

For those who want a more interdisciplinary degree, this is a great route to go. Students will also explore big data issues through courses such as predictive modeling, data communication, quantitative methods, and advanced database administration and development. Students can complete the degree fully online if they wish.

#22

NORTHWESTERN UNIVERSITY

Overview

Northwestern University is located in North Chicago, Illinois and needs little introduction. It is one of the best schools in the Midwest and across the nation. Northwestern offers a Master of

Science in Predictive Analytics, one of the very first online big data degree programs in the country.

Features

One of the features that sets Northwestern's program apart from others is the areas of application they prepare students for, such as sports management and performance analytics, web and network data science, advanced modeling techniques, marketing and risk analytics, and data visualization. Students can expect to be prepared in:

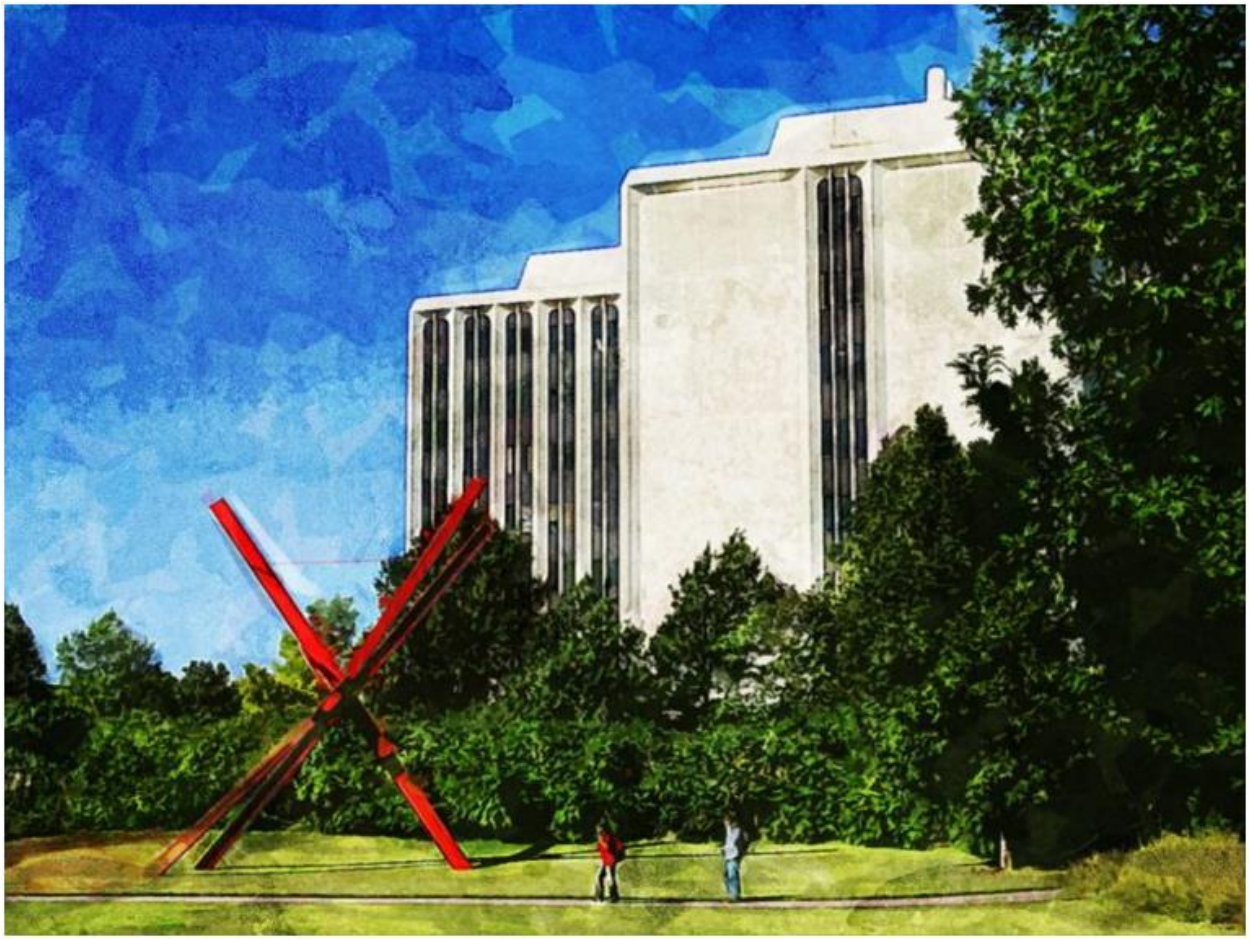
- Database administration
- Predictive modeling and quantitative analysis
- Communication of analytics
- And leadership development

Notables

Northwestern is a globally and nationally ranked university with a vibrant research culture. It is a pacesetter not only in national research universities but internationally as well. It leads the pack in any number of disciplines, from business and law to the arts and sciences.

#23

UNIVERSITY OF NEBRASKA



Overview

The University of Nebraska is located in Lincoln. It offers students interested in big data a Master of Business Administration degree with a concentration in Business Analytics. It will appeal particularly to students who have an interest in applying data solutions to business challenges.

Features

Whether you are looking for a brief, specialized program in analytics or a more comprehensive program, Nebraska will appeal to a wide variety of students. Nebraska also offers a twelve-credit graduate certificate in business analytics, which can be completed in a year or less. Both programs offer training in:

- Information technology and statistics
- Business analytics, econometrics
- Strategic database marketing
- And data mining applications

Notables

University of Nebraska Lincoln is the flagship school of the University of Nebraska system. It is one of the leading public universities not only in the Midwest but in the country also. Equally, its online offerings are robust.

#24

ARIZONA STATE UNIVERSITY

Overview

Arizona State University has locations in both Phoenix and Tempe. It is one of the nation's most active research universities. ASU offers a Master of Science in Business Analytics, a sixteen-month program that can be completed on a part-time basis.

Features

Each course in ASU's program lasts five weeks, and students take classes one at a time. The program requires ten courses total, but students can work at their own pace according to their goals and capabilities. Coursework in the program includes training in everything from enterprise analytics to applied analytics, data mining to data-driven quality management. You'll also gain a foundation in:

- Analytical decision-making tools
- Business analytics strategy
- Marketing analytics
- And more

Notables

ASU is another national-level leader in both on-campus degree programs and online education. It has quickly established itself as one of the standards in online higher education. Its Carey School of Business, one of the largest in the country, in particular has been a pacesetter in distance education for business students.

#25

DEPAUL UNIVERSITY

Overview

DePaul University, located in Chicago, Illinois, is a Catholic university with a mandate to provide education access to lower-income and minority students. It has been recognized as the most diverse university in the country by the Princeton Review. DePaul offers a Master of Science in Predictive Analytics through its College of Computing and Digital Media. The MSPA is also affiliated with DePaul's Data Mining and Predictive Analytics Center.

Features

Students in both programs can expect exposure to a wide array of subjects relevant to big data, such as programming, tools and techniques for computational analysis, and statistics and data analysis. Additional subjects include:

- Mining big data
- Social network analysis
- Programming machine learning applications
- Intelligent information retrieval, and more

Notables

DePaul also offers a Graduate Analytics Certificate, a five-course program that is less involved compared to the MS degree. This is a great way to gain advanced experience and deepen your education without breaking the bank. It will still appeal to employers and take less time to complete. You can always use the credits later to fulfill the full degree if you want.

#26

CENTRAL CONNECTICUT STATE UNIVERSITY

Overview

Central Connecticut State is an affordable school located in New Britain, Connecticut. CCSU offers a Master of Science in Data Mining, a degree directly relevant to students interested in careers in big data. Coursework can be completed online, but students should be prepared to travel to campus at the completion of their degree to present a final thesis project.

Features

CCSU's program offers students cutting-edge training with the latest in data modeling technology, and students take courses they might not have the opportunity to take at other institutions, such as multivariate analysis, clustering and affinity analysis, predictive analytics, genomics and proteomics, and text mining. There is also room for particular areas of specialization—the program offers concentrations in:

- Experimental design
- Artificial intelligence
- Nonparametric statistics
- Applied stochastic processes

Notables

Recently CCSU has doubled down on its commitment to making education affordable and accessible. It has set itself up for success by expanding its online offerings in recent years. And it puts excellent higher education in reach for a larger public.

#27

OKLAHOMA STATE UNIVERSITY

Overview

Oklahoma State University (OSU) is located in Stillwater. It offers a Master of Business Analytics, a 33-credit program designed for working professionals, through its Spears School of Business and Watson Graduate School of Management. Students can opt to specialize in Business Data Mining or in Marketing Analytics.

Features

In addition to the Master of Business Analytics, students can also earn certificates in a Data Mining, Predictive Analytics, and Marketing Data Science. Whatever course students happen to opt for, students will take a diverse range of courses, including in:

- Programming languages (such as Python)

- Analytics
- Marketing and business
- Statistics, and more

Notables

Oklahoma State University has been investing heavily in online education in recent years and has managed to offer some of the most affordable distance education programs on the market. Indeed, OSU's programs are some of the most diverse and affordable on our list.

#28

THOMAS EDISON STATE UNIVERSITY

Overview

Thomas Edison State University is a public university. It is located in Trenton, NJ. TESU offers a Master of Science in Information Technology, which is available online and can be completed in two years.

Features

TESU is known for its capability in technological education, and its MSIT program is geared toward IT. Graduates of the MSIT program will acquire facility in technical skills of the principles and theories of information technology, analyze the societal implications of the technical, ethical and regulatory issues of information technology, compare and contrast major IT components, and apply their knowledge to marketing, product, project and personnel management for an IT business. Students will also gain proficiency in:

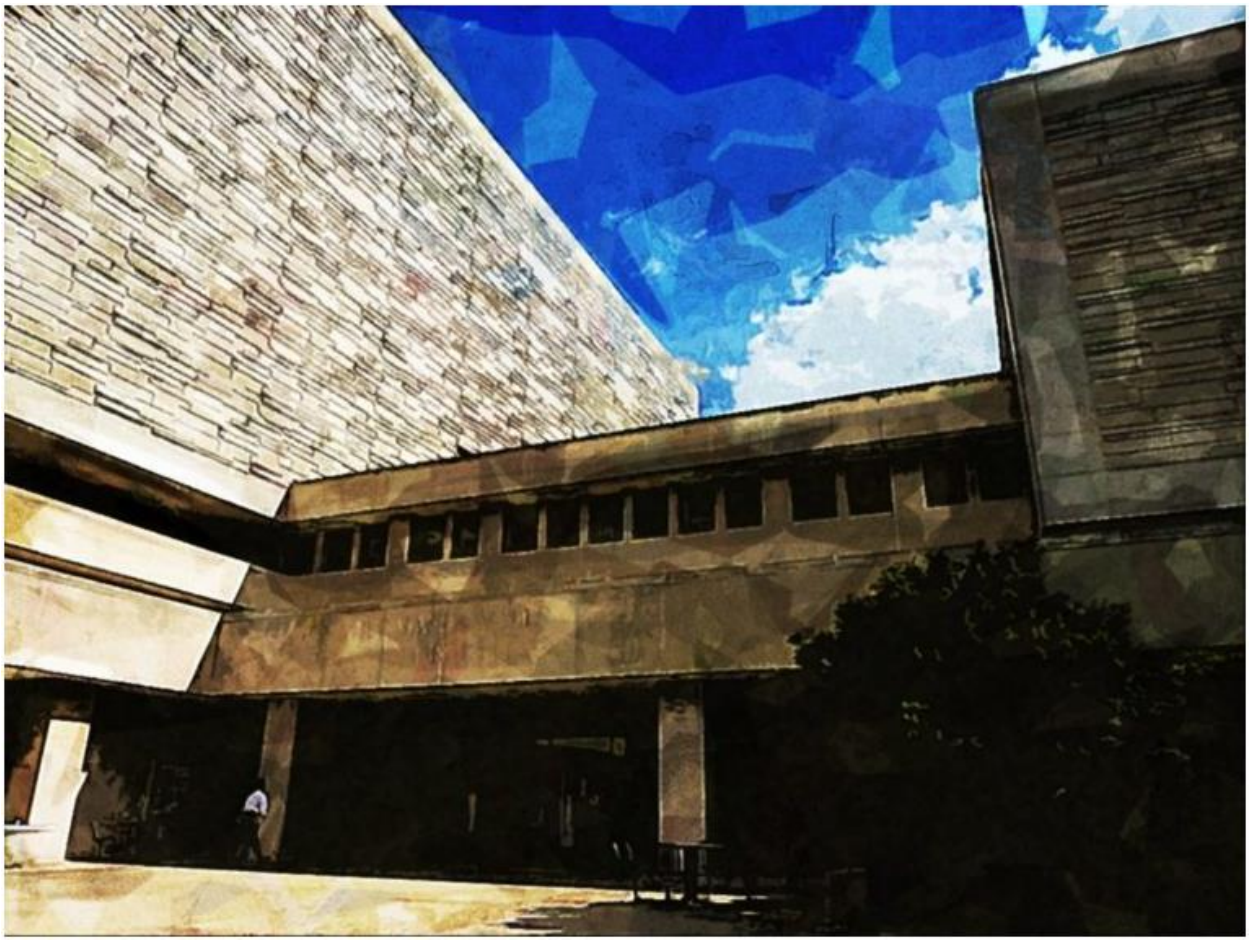
- Computer networks
- Data communications
- Distributed database management systems

Notables

TESU specializes in addressing the needs of working adults and nontraditional students. It seeks to help adult learners, working professionals, and military personnel achieve their goals of obtaining high-quality post-secondary education.

#29

INDIANA UNIVERSITY



Overview

Indiana University in Bloomington, the flagship school of the Indiana University system, is a nationwide-leader in public education in a wide variety of fields, including business and STEM fields. It also features a highly active research community. IU offers a Master of Science in Data Science through its School

of Informatics and Computing, a thirty-credit program that prepares students in data collection, analysis, and use. Students can choose either the Technical Path or the Decision-Maker Path, depending on whether their primary interest is in computing and statistics or in management and leadership.

Features

Coursework in the MSDS program includes business analytics, Python, real world data science, data semantics, systems and protocol security, applied machine learning, and the management, access, and use of big and complex data. IU offers a number of opportunities outside the coursework component of the degree through such programs as:

- The BEST Aspiring Entrepreneur Program
- ICE (Informatics and Computing Entrepreneurs) group
- The Velocity Conference for graduate students
- VisionTech Entrepreneur Advisory Sessions

Notables

In addition to the MSDS degree, IU also offers a twelve-credit graduate certification in Data Science. Again, this is a great way to dip your toes in without having to commit to several years of education. IU Bloomington is nationally and globally ranked.

#30

CALIFORNIA STATE UNIVERSITY FULLERTON

Overview

Cal State Fullerton is part of the California State University system. It is located in Orange County but has begun to increase its online offerings significantly in recent years. For the purposes of this list, CSF offers not one but two options: the Master of Science in Information Science and the Master of Science in Statistics.

Features

Students in the Master of Science programs typically take two courses per term, and courses are often interactive. Coursework in the MS degrees includes areas such as business intelligence, regression analysis, forecasting, statistics, project management, database design and processing, and data mining. Students will develop expertise:

- Identifying, designing, and managing data storage and sets
- Recognizing strategies for analysis, methodology, and decision making
- And much more

Notables

It also offers an accelerated online graduate certification in Data Science, a short, five-course program that will allow working adults to get the qualifications they need in the big data field quickly. Cal State Fullerton is regularly featured in national, regional, and state rankings.

#31

SOUTHERN METHODIST UNIVERSITY

Overview

Southern Methodist University is located in Dallas, Texas. It is one of the larger and wealthier private research institutions in the country, and it boasts a well-rounded offering of undergraduate and graduate programs in both the arts and sciences.

Features

It has also invested heavily in its online offerings. SMU offers a Master of Science in Data Science, an interdisciplinary program offered collaboratively by the Dedman College of Humanities and Sciences, the Lyle School of Engineering, and the Meadows School of the Arts. Recent graduates have been placed in:

Education

- Government
- Large corporations
- Nonprofit
- Healthcare and more

Notables

SMU’s approach to education in data is project-based, which reflects how students might encounter problems in real-world scenarios. Its curriculum is intentionally interdisciplinary, and it offers an ideal balance between the flexibility of online education and a traditional graduate school experience.

#32

UNIVERSITY OF MARYLAND UNIVERSITY COLLEGE

Overview

The University of Maryland University College is based in Adelphi, Maryland. It offers two options for students interested in big data: a Master of Science or graduate certification in Data Analytics. Students have a choice whether to pursue their degree online, on campus, or a combination of both.

Features

UMUC encourages students to combine technical expertise and business acumen from the outset of the program. Coursework in the program is exhaustive and includes everything from decision management systems to data management and visualization. You'll also learn the ins and outs of:

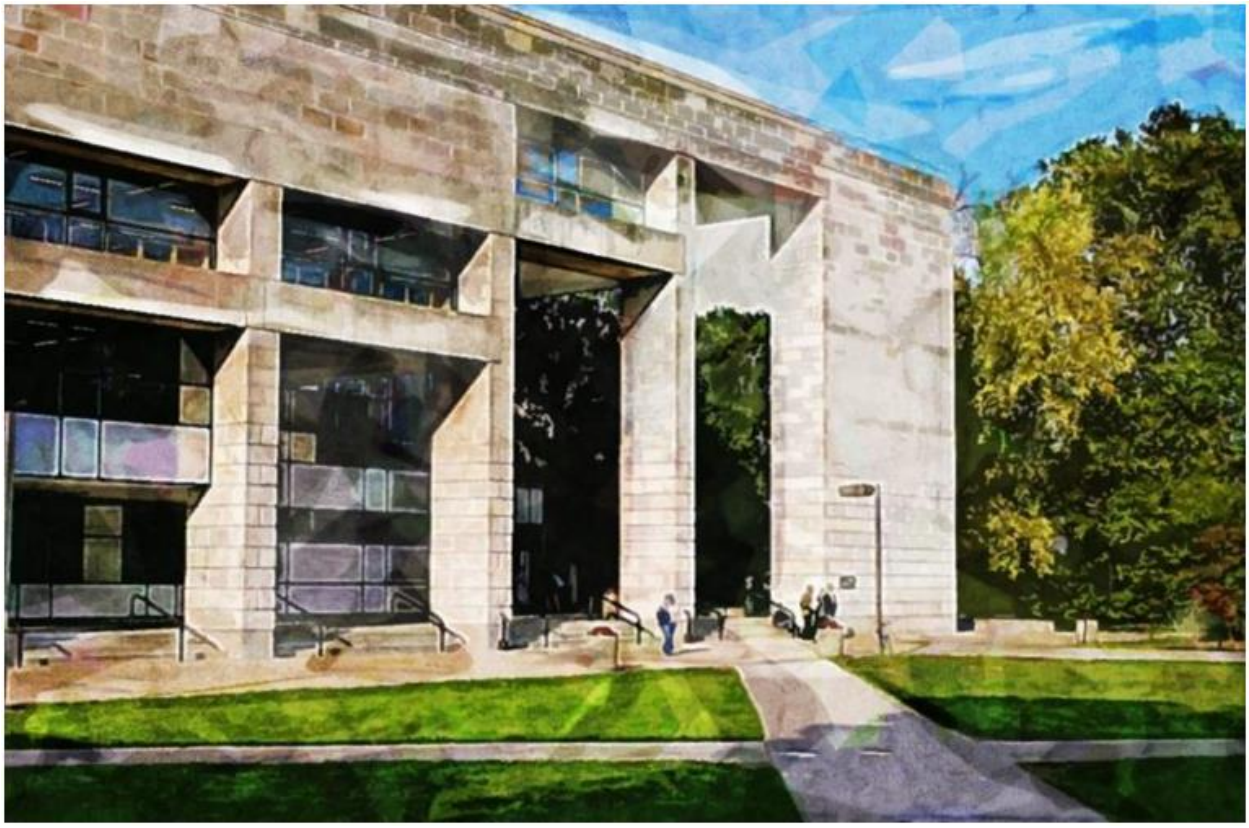
- Data mining
- Predictive modeling
- Big data analytics, and more

Notables

The University of Maryland University College occupies a unique and needful niche in higher education: it was founded specifically to meet the educational needs of working adults and non-traditional students. UMUC has therefore placed a heavy emphasis on online and distance education, which appeals to military personnel, working professionals, and others.

#33

IOWA STATE UNIVERSITY



Overview

Iowa State is a well-known public institution that has made its mark in affordable and accessible education for as many students as possible. It is one of the leading research universities in Iowa, and it is particularly well-known for its engineering and business programs.

Features

ISU offers a Master of Business Analytics, which reflects a combination of their areas of strength and utilizes faculty from computer science, electrical and computer engineering, statistics, and industrial and manufacturing systems engineering departments. Students can expect to encounter a wide variety of subjects in their coursework, including:

- Project management
- Statistical and predictive modeling
- consumer sentiment analysis
- Knowledge discovery, and more

Notables

Students can complete their degree from Iowa State online, but brief residencies are required. Students are also required to finish their program with a capstone project. Both of these requirements encourage students to have a hands-on learning experience.

#34

SAINT MARY'S COLLEGE

Overview

Saint Mary’s College is located in Notre Dame, Indiana. It is a Catholic liberal arts college that offers a Master of Science in Data Science degree designed for students with strong mathematical backgrounds and interests. The program is designed to be completed primarily online.

Features

Saint Mary's program emphasizes the mathematical and computational aspects of data science, and it encourages expertise in the tools necessary to work with today's big data problems. One student in the program had this to say about his experience: "I'm able to dig beyond what a simple analyst can do, and not only tell them what happened, but why it happened, and why we came up with the answers we did." Students will learn:

- A variety of programming languages
- Big data analysis and methodology
- And much more

Notables

Saint Mary's boasts a stellar academic reputation. This is owed in no small part to its extraordinarily high retention and graduation rate, generous financial aid, and strong placement records.

#35

BELLEVUE UNIVERSITY

Overview

Another school with a mandate to provide education for non-traditional and working adult students, Bellevue University is in Bellevue, Nebraska. It has likewise invested heavily in its online offerings. Its efforts have paid off—it is now recognized as one of the highest quality online educators in the nation.

Features

Bellevue offers a Master of Science in Business Analytics that is designed to prepare students for careers in managing the big data problems of today's large corporations. Coursework in the program covers:

- Business administration fundamentals
- Project management
- Computer information systems
- Management systems, and much more

Notables

Bellevue's degree is designed to be flexible, affordable, expedient, supportive, and real-world focused. It is also designed to cover as many disciplines as possible, from business to IT. Bellevue students enjoy an excellent placement record in a wide variety of industries.

#36

NOVA SOUTHEASTERN UNIVERSITY

Overview

Nova Southeastern University is located in Davie, Florida. It is one of the most diverse campuses in the United States. Nova offers a Master of Science in Information Technology, a 36-credit graduate degree focused on the systems and software of big data.

Features

Nova also offers a Graduate Certificate in Business Intelligence and Analytics. Both programs are offered through the College of Engineering and Computing, and they offer training in business intelligence, data mining, and organizational decision making.

Coursework in the program includes decision support systems or business modeling, database management and applications, data warehousing, data mining, and a capstone project in business intelligence. Graduates of the program work in:

- Business intelligence analytics or architecture
- Data warehousing
- Information technology business analysis

Notables

This school occupies another unique niche in higher education—it was founded as a graduate school from the beginning rather than primarily an undergraduate institution. It emphasizes education in the social science and sciences, and is also well-known for education in technology and distance education.

#37

ELMHURST COLLEGE

Overview

Elmhurst College is a small liberal arts college in Elmhurst, Illinois. It is located just outside Chicago and offers a Master of Science in Data Science that can be completed in as quickly as

two years, even at a part-time pace. Elmhurst also offers a Graduate Certificate in Data Science.

Features

Elmhurst's program emphasizes both the theory behind big data management as well as the applied skills involved in real-world scenarios, such as analytical modeling and decision making based on data sets, as well as the software skills involved in these applications. Coursework for the program includes:

- Quantitative methods
- Data warehousing
- Data mining project
- Programming languages and environments, and more

Notables

Elmhurst regularly appears in rankings of best regional schools. Its degrees can be completed online or on campus, or a mix of both. Choose your own adventure!

#38

SAINT LOUIS UNIVERSITY

Overview

Saint Louis University offers programs up to the doctoral level and also a number of online programs. SLU offers a Master of Science in Applied Analytics through its School of Professional Studies, which is designed specifically for working professionals. The program is structured in accelerated eight-week courses for both full-time and part-time students alike.

Features

The program requires 36 credit hours to complete, and students are required to complete a capstone project in analytics to finish the program. The program prepares students in the diverse areas of data science and information technology. In addition to studying data mining and research methodology, you will also study:

- Business analytics
- Programming and problem solving
- Information retrieval and visualization
- And much more

Notables

Saint Louis University is the oldest university west of the Mississippi River! It regularly receives attention as one of the top national research institutions in the United States. And it is regionally ranked by an array of sources, including yours truly.

#39

SAINT JOSEPH'S UNIVERSITY

Overview

Located in Philadelphia, Pennsylvania, Saint Joseph’s University is a private, Catholic university. It has invested heavily in its online programs in recent years. Through its Haub School of Business, it offers a Master of Science in Business Intelligence and Analytics.

Features

Coursework for Saint Joseph's program includes subjects such as data warehousing, DSS modeling, enterprise data, business intelligence, critical performance management, and business process modeling. Graduates of Saint Joseph's program have secured employment in a range of industries, including:

- Information Technology
- Healthcare Management
- Education
- And more

Notables

One of the unique features of Saint Joseph's program is that it includes an SAS certification, which will provide graduates a leg up on the competition in the job market. Furthermore, this degree is available fully online for those who are looking primarily for a distance program.

#40

BRANDEIS UNIVERSITY

Overview

Brandeis, a relatively younger university, is located in Waltham, Massachusetts. Brandeis offers a Master of Science in Strategic Analysis, a degree that prepares professionals specifically for decision-making positions in big data. The degree is available fully online.

Features

Skills that graduates of the program at Brandeis can expect to acquire include identifying, interpreting, and communicating patterns in large data sets and then creating actionable and prioritized items for whatever organization they happen to be working for. Coursework in the program includes business intelligence, marketing analytics, data governance, and visualization. Graduates of the Brandeis program have secured employment in a wide range of fields, including but not limited to:

- Project Consultants
- Compliance Managers
- Data Architects or Scientists
- Marketing Consultants, and more

Notables

Though young, Brandeis has wasted no time establishing a world-class reputation in New England and beyond. Indeed, it is regularly featured in rankings of the top universities of any sort in the country.

#41

COLORADO TECHNICAL UNIVERSITY

Overview

Colorado Technical University is primarily an online university. More than 90 percent of its students study from a distance. CTU offers the only Doctor of Computer Science on our list with a specialization in Big Data Analytics. It is a 96-credit hour program designed for students who are not just interested in big data but want to pursue careers as leaders in the field.

Features

Courses in CTU's program include database systems, business intelligence, analytics for big data, principles of research methods, qualitative and quantitative research methods. Doctoral students are expected to complete a dissertation, and the degree is typically completed in three years. Graduates of CTU's program will be equipped to pursue careers in senior level leadership positions, in:

- Consulting
- Higher education
- And in the technological sectors of business, government, and nonprofit

Notables

While a for-profit university, Colorado Technical University has nevertheless worked its way into national recognition by outlets such as, *U.S. News & World Report*, NSA, and Homeland Security. CTU has established a reputation for its online degree programs, which are both diverse and growing.

#42

LEWIS UNIVERSITY

Overview

Lewis University is a small, private Catholic liberal arts university. Lewis offers a Master of Science in Data Science, which prepares students for all of the technical and analytical demands of big data. It is a thirty-six credit hour degree that offers

two specializations: computer science and computational biology and bioinformatics.

Features

Coursework in Lewis's program includes data mining, visualization, predictive analysis, and data management. Students who opt for the computational biology specialization will take courses in data systems in the life sciences, mathematics for data scientists, and the healthcare industry. Faculty and students work closely together in Lewis's program, and students can expect to receive training:

- In large-scale data storage systems
- Encryption and authentication systems
- Statistical programming
- And languages such as Python and Java

Notables

Lewis regularly receives attention as a leading school in the Midwest region. It has also increased its online offerings to great effect in recent years. Not surprising then, Lewis is ranked for both its traditional and online programs.

#43

SLIPPERY ROCK UNIVERSITY

Overview

Slippery Rock is part of the Pennsylvania State System of Higher Education. Has grown out of its humble beginnings as a teacher's college and now offers undergraduate and graduate programs in technology, business, education, and more in addition to a variety of online programs.

Features

Slippery Rock offers a Master of Science in Data Analytics through its Mathematics Department, a thirty-three credit program that can be completed in as little as ten months or in two years on a part-time basis. Coursework in SLU's program includes training in:

- Statistics
- Data mining
- Regression methods
- Big data analysis and forecasting

Notables

One of the benefits of Slippery Rock's program that is shares with only one or two others on our list is that students will graduate not only with their degree but also with a joint certificate in statistical applications and data analytics from SRU and the SAS (Statistical Analysis System) Institute for advanced analytics.

#44

CAPELLA UNIVERSITY

Overview

Capella is another for-profit university that has managed to garner the respect of the higher education world through its consistent quality and regional accreditation. Students interested in big data will find a home in two different programs: Capella's MBA program, which offers a specialization in Business Intelligence, which also offers opportunities for students to earn additional graduate certifications, and their Master of Science in Analytics.

Features

Capella's MSA program prepares data analytics professionals to work with, understand, and transform data. Students will also learn to apply their knowledge to real-world problems and communicate their results to whatever organization they happen to work for. Students are introduced to tools such as SAS, R, Python, and Tableau. They will also develop expertise:

- Data mining and statistics
- Applied analytics
- Modeling and forecasting
- Visualization, and more

Notables

Capella is unique as the first university to receive U.S. Department of Education approval for an innovative competency-based curriculum model, in which students demonstrate accomplishment in skills areas rather than working through a traditional credit-hour system.

#45

UNIVERSITY OF TENNESSEE AT CHATTANOOGA

Overview

The University of Tennessee at Chattanooga is part of the larger University of Tennessee system. Well known for its business curriculum, UTC offers an MBA degree with a concentration in Business Analytics, which prepares students not just for the technical know-how necessary for data professionals but also the professional business training necessary to thrive in management.

Overview

Students who concentrate in UTC's MBA Business Analytics focus will receive training in data storage, reporting, and analysis. For full-time students, the program can be completed in as few as 16-24 months, while part-time students can expect to take longer. Coursework for UTC's program covers:

- Database and data warehouses
- Prescriptive analytics
- Data mining and data analytics
- Advanced data analysis

Notables

It has a long history of offering a wide variety of practical education opportunities at an affordable rate for a diverse student body. That has continued to be true as it has expanded into the online education market. UT Chattanooga offers an array of alternative learning experiences.

#46

UNIVERSITY OF NORTH CAROLINA GREENSBORO

Overview

The Greensboro member within the wider University of North Carolina System is a mid-sized, regional, public research university. It offers a Master of Science in Information Technology with a concentration in Business Analytics. It also

offers an online post-bachelor's certificate in Business Analytics that has been developed in partnership with SAS.

Features

Coursework for UNC Greensboro's program includes data management, systems analysis and design, information systems management, project management, models and methods in business analytics, and more. The program is designed to provide the most up-to-date information, knowledge, and skills possible in the field, and students will be prepared to:

- Work with large amounts of data
- Interpret and present that data
- Apply that knowledge across industries

Notables

The University of North Carolina Greensboro has been hard at work appealing to students in need of advanced education in business and technology, particularly non-traditional students and part-time learners. It has put numerous services and resources in place to ensure academic success to alternative learners.

#47

CITY UNIVERSITY OF SEATTLE

Overview

The City University of Seattle is a small, private university in Seattle, Washington. It offers a range of degree opportunities across both academic and professional levels. One of those programs is the Master of Science in Computer Science degree with a concentration in Data Management/Big Data. The degree requires 45 credit hours to complete.

Features

CUS's program is designed to offer both technical education and business knowledge, that is, both theory and practice. The program is flexible but also rigorous to meet the demands of the highly competitive job market. Coursework in the program includes computer architecture, software engineering, software testing, information security, and technical project management. Students can also specialize in:

- Programming
- Technology management
- Development management
- Cloud development, more

Notables

CUS has garnered quite a bit of attention in recent years for its online education efforts from outlets such as *U.S. News & World Report* and the *Princeton Review*. Indeed, CUS programs are especially designed to provide not just a theoretical foundation, but to develop communication skills in its students, to give them hands-on experience, and more.

#48

SOUTHERN NEW HAMPSHIRE UNIVERSITY

Overview

Southern New Hampshire University (SNHU) is a private research university. It was founded to provide accessible, affordable education to returning learners. SNHU also offers a Master of Science in Data Analytics, a program designed as a Professional Science Master's for employment in scientific and technology-oriented industries.

Features

SNHU's program is recognized by the Professional Science Master's (PSM) national office as a Professional Science Master's program, which is any degree designed to allow students to pursue advanced training in science or mathematics while developing

workplace skills valued by employers. Coursework in SNHU's program includes:

- Foundations of data and decision methods and modeling
- Presentation and visualization of data
- Optimization and risk assessment
- Predictive analytics

Notables

SNHU is the standard-bearer of online education. It is regularly recognized for its innovation in the field of distance learning. Its business programs in particular receive regular attention for excellence.

#49

AMERICAN UNIVERSITY

Overview

Located in Washington, DC, American University is a top-ranked university. It comes as little surprise that this university has strong programs in politics and business. American University offers a

Master of Science in Analytics, which can be completed between one and two years.

Features

American University places equal emphasis on the theoretical and practical dimensions of big data. Students will be able, for instance, to both identify data-based issues and communicate them—and their solutions—effectively and clearly. The program also provides students with numerous opportunities for hands-on experience through local organizations, career services, networking, and more. Students at American University's MSA have the option of specializing in seven different areas, among them:

- Accounting forensics
- Biostatistics
- Business fundamentals
- Computational statistics, and more

Notables

American University is known for its political activity, public service, and commitment to traditional American values. It is recognized for national-level excellence, particularly in the business and management fields, by outlets such as the *Wall Street Journal*, the *Princeton Review*, and more.

#50

DAKOTA STATE UNIVERSITY

Overview

Dakota State University (DSU) has established its reputation as a leader in research and education in the areas of technology and STEM fields. DSU offers a Master of Science in Analytics, which offers the opportunity to specialize in either Healthcare Analytics or Information Systems. The program prides itself on being both flexible and affordable.

Features

This is the ideal program for those looking for an exhaustive introduction to big data. Indeed, there will be few things DSU students aren't prepared to face. Possible career outcomes for DSU students include

- Data analyst, scientist, or architect
- Governance specialist
- Visualization specialist
- Predictive modeler, and so much more

Notables

DSU has also invested heavily in various types of online, distance, hybrid, and on-ground education models to great success. But perhaps one of the most unique aspects of DSU's program is its optional emphasis on healthcare, which offers courses in healthcare information, healthcare informatics and systems technology, data analytics in healthcare, and health information lifecycle governance.

IBM's Big Data University: Free Online Learning with Over 400,000 Students

Bernard Marr Contributor

Enterprise & Cloud

This article is more than 2 years old.

It is widely forecasted that a shortage of skills in data science and analytics will mean a great deal of money is wasted through missed opportunities in coming years.

Traditional academic establishments have begun to move to fill the gap. However, most courses teaching the hot topic skill sets such as machine learning and predictive modeling are graduate level, with a high barrier to entry. This has created an

opportunity for online distance learning providers, which can move far more nimbly to adapt to demand.

Although there are a large number of startups in the field, one old name which is once again making waves is IBM IBM +0% . Its Big Data University initiative has now signed up over 400,000 students. Although it doesn't charge for any of its courses, it takes a business-centric approach to delivery which sets it apart from many other online open course providers.

Students can sign up independently at home and work at their own speed and on their own initiative, just as they can with Coursera and Futurelearn or other similar services.

However, IBM also works with partners to tailor course packages to fill the needs of individual organizations. And it allows (actually, encourages) third party vendors to sell on their free course material, while adding their own value such as custom delivery and support packages.

This parallels the vendor ecosystem which has emerged around distribution of open source

software, platforms and services, often said to be "democratizing" access to Big Data technology. Will this democratize access to Big Data education and training in the same way? It appears that's what IBM is banking on.

"The number one question I get asked," Leon Katsnelson, IBM's director and CTO for analytics platform emerging technologies told me, is "Why are you doing this for free? What's the real reason?"

"I tell them IBM has not turned into a communist or anti-capitalist entity. We are in the business of

selling tools and implements for people to use. If people learn from us, they will buy from us."

So although Big Data University is owned and administered by IBM it is considered a "community" rather than a corporate division and its courses are designed to be fully platform agnostic.

"We very early on discovered that people want to learn technology. They don't necessarily want to come to us to learn about a specific implementation by a particular vendor.

"We teach people, we help people get skills, we are not there to teach how to use IBM products," Katsnelson says.

Time and money are seen as the obstacles barring the way to education for many people. Making course material and accreditation free is the solution taken by IBM and the open course industry to the first problem.

Structuring learning in a way that it can take place at any speed, in bite-size chunks, is the solution that IBM is taking to the second. Partner organizations all see their own portal which can be used to create

learning paths for their staff by picking modules from across all of BDU's courses.

Another innovation in IBM's model of delivery is a focus on applying the skills which are taught. Students can all download the open source Big Data Workbench toolset, and are given practical real—world problems to solve.

This gives the student a sense of real-world achievement, seen by the university as vital in motivating self-learners and distance-learners.

Worldwide the Big Data University has partnered with organizations including Tata Consulting, Trans Neuron Technologies and Cognizant. It has also just this year launched in China and announced a partnership with Kaitan-Weibo, an educational division of the social media giant Weibo.

As well as opening up new fields of study, Big Data has kick-started a wave of changes in the ways that more traditional subjects are taught. Distance and online learning provides opportunities for more in-depth investigation into how, and why, we learn efficiently.

Currently this is an area where a lot of data is available but it still needs more innovative analysis. BDU is working on this though.

For example while course providers may easily be able to tell that a large number of students are rewinding and Re-watching a video lecture from a certain point, it's difficult to get actionable insights on this without knowing whether it's because the instructor slurs their speech, or because a point was not explained well. Online course providers – including IBM - are working on ways to extract meaning from this data.

Katsnelson said "Education has really gone high tech and become data driven. What we are discussing here is sort of the next frontier. We are able to measure and affect people's learning."

This means that we are likely to see continued innovation in the way that courses are delivered.

The changing needs of employers mean they are likely to want to integrate analytically trained staff across an increasingly wide segment of their business. The need for these "citizen data scientists" – non data-scientists with data science skills -has

been forecasted to grow five times more quickly than the need for people specifically trained as data scientists. The Big Data University is clearly an interesting attempt to meet that need.

Bernard Marr

Bernard Marr is an internationally best-selling author, popular keynote speaker, futurist, and a strategic business & technology advisor to governments and companies...

How Big Data Is Taking Teachers Out of the Lecturing Business?

“Schools and universities are embracing technology that tailors content to students' abilities and takes teachers out of the lecturing business. But is it an improvement?”

When Arnecia Hawkins enrolled at Arizona State University last fall, she did not realize she was volunteering as a test subject in an experimental reinvention of American higher education. Yet here she was, near the end of her spring semester, learning math from a machine. In a well-appointed computer lab in Tempe, on Arizona State's desert resort of a campus, she and a sophomore named Jessica were practicing calculating annuities. Through a software dashboard, they could click and scroll among videos, text, quizzes and practice problems at their own pace. As they worked, their answers, along with reams of data on the ways in which they arrived at those answers, were beamed to distant servers. Predictive algorithms developed by a team of data scientists compared their stats with data gathered from tens of thousands of other students, looking for clues as to what Hawkins was learning, what she was struggling with, what she should learn next and how, exactly, she should learn it.

Having a computer for an instructor was a change for Hawkins. “I'm not gonna lie—at first I was really annoyed with it,” she says. The arrangement was a switch for her

professor, too. David Heckman, a mathematician, was accustomed to lecturing to the class, but he had to take on the role of a roving mentor, responding to raised hands and coaching students when they got stumped. Soon, though, both began to see some benefits. Hawkins liked the self-pacing, which allowed her to work ahead on her own time, either from her laptop or from the computer lab. For Heckman, the program allowed him to more easily track his students' performance. He could open a dashboard that told him, in granular detail, how each student was doing—not only who was on track and who was not but who was working on any given concept. Heckman says he likes lecturing better, but he seems to be adjusting. One definite perk for instuctors: the software does most of the grading for them.

At the end of the term, Hawkins will have completed the last college math class she will probably ever have to take. She will think back on this data-driven course model—so new and controversial right now—as the “normal” college experience. “Do we even have regular math classes here?” she asks.

Arizona State's decision to move to computerized learning was born, at least in part, of necessity. With more than 70,000 students, Arizona State is the largest public university in the U.S. Like institutions at every level of American education, it is going through some wrenching changes. The university has lost 50 percent of its state funding over the past five years. Meanwhile enrollment is

rising, with alarmingly high numbers of students showing up on campus unprepared to do college-level work. “There is a sea of people we're trying to educate that we've never tried to educate before,” says Al Boggess, director of the Arizona State math department. “The politicians are saying, ‘Educate them. Remediation? Figure it out. And we want them to graduate in four years. And your funding is going down, too.’”

Two years ago Arizona State administrators went looking for a more efficient way to shepherd students through basic general-education requirements—particularly those courses, such as college math, that disproportionately cause students to drop out. A few months after hearing a pitch by Jose Ferreira, the founder and CEO of the New York City adaptive-learning start-up Knewton, Arizona State made a big move. That fall, with little debate or warning, it placed 4,700 students into computerized math courses. Last year some 50 instructors coached 7,600 Arizona State students through three entry-level math courses running on Knewton software. By the fall of 2014 ASU aims to adapt six more courses, adding another 19,000 students a year to the adaptive-learning ranks. (In May, Knewton announced a partnership with Macmillan Education, a sister company to *Scientific American*.)

Arizona State is one of the earliest, most aggressive adopters of data-driven, personalized learning. Yet educational institutions at all levels are pursuing similar options as a way to cope with rising enrollments, falling budgets and more stringent requirements for student achievement. Public primary and secondary schools in 45 states and the District of Columbia are rushing to implement new, higher

standards in English-language arts and mathematics known as the Common Core state standards, and those schools need new instructional materials and tests to make that happen. Around half of those tests will be online and adaptive, meaning that a computer will tailor questions to each student's ability and calculate each student's score [see “Why We Need High-Speed Schools,” on page 69]. School systems are experimenting with a range of other adaptive programs, from math and reading lessons for elementary school students to “quizzing engines” that help high school students prepare for Advanced Placement exams. The technology is also catching on overseas. The 2015 edition of the Organization for Economic Co-operation and Development's Program for International Student Assessment (PISA) test, which is given to 15-year-olds (in more than 70 nations and economies so far) every three years, will include adaptive components to evaluate hard-to-measure skills such as collaborative problem solving.

Proponents of adaptive learning say that technology has finally made it possible to deliver individualized instruction to every student at an affordable cost—to discard the factory model that has dominated Western education for the past two centuries. Critics say it is data-driven learning, not traditional learning, that threatens to turn schools into factories. They see this increasing digitization as yet another unnecessary sellout to for-profit companies that push their products on teachers and students in the name of “reform.” The supposedly advanced tasks that computers can now barely pull off—diagnosing a student's strengths and weaknesses and adjusting materials and approaches to suit individual learners—are things human teachers have been

doing well for hundreds of years. Instead of delegating these tasks to computers, opponents say, we should be spending more on training, hiring and retaining good teachers.

And while adaptive-learning companies claim to have nothing but the future of America's children in mind, there is no denying the potential for profit. Dozens of them are rushing to get in on the burgeoning market for instructional technology, which is now a multibillion-dollar industry [*see box at left*]. As much as 20 percent of instructional content in K–12 schools is already delivered digitally, says Adam Newman, a founding partner of the market-analysis firm Education Growth Advisors. Although adaptive-learning software makes up only a small slice of the digital-instruction pie—around $50 million for the K–12 market—it could grow quickly. Newman says the concept of adaptivity is already very much in the water in K–12 schools. “In K–12, the focus has been on differentiating instruction for years,” he says. “Differentiating instruction, even without technology, is really a form of adaptation.”

Higher-education administrators are warming up to adaptivity, too. In a recent Inside Higher Ed/Gallup poll, 66 percent of college presidents said they found adaptive-learning and testing technologies promising. The Bill & Melinda Gates Foundation has launched the Adaptive Learning Market Acceleration Program, which will issue 10 $100,000 grants to U.S. colleges and universities to develop adaptive courses that enroll at least 500 students over three semesters. “In the long term—20 years out—I would expect virtually every course to have an adaptive component of

some kind," says Peter Stokes, an expert on digital education at Northeastern University. That will be a good thing, he says—an opportunity to apply empirical study and cognitive science to education in a way that has never been done. In higher education in particular, "very, very, very few instructors have a formal education in how to teach," he says. "We do things, and we think they work. But when you start doing scientific measurement, you realize that some of our ways of doing things have no empirical basis."

In general, "adaptive" refers to a computerized-learning interface that constantly assesses a student's thinking habits and automatically customizes material for him or her. Not surprisingly, though, competitors argue ferociously about who can claim the title of true adaptivity. Some say that a test that does nothing more than choose your next question based on whether you get the item in front of you correct—a test that steers itself according to the logic of binary branching—does not, in 2013, count as fully adaptive. In this view, adaptivity requires the creation of a psychometric profile of each user, plus the continuous adjustment of the experience based on that person's progress.

To make this happen, adaptive-software makers must first map the connections among every concept in a piece of learning material. Once that is done, every time a student watches a video, reads an explanation, solves a practice problem or takes a quiz, data on the student's performance, the effectiveness of the content, and more flow to a server. Then the algorithms take over, comparing that student with

thousands or even millions of others. Patterns should emerge. It could turn out that a particular student is struggling with the same concept as students who share a specific psychometric profile. The software will know what works well for that type of student and will adjust the material accordingly. With billions of data points from millions of students and given enough processing power and experience, these algorithms should be able to do all kinds of prognostication, down to telling you that you will learn exponents best between 9:42 and 10:03 a.m.

They should also be able to predict the best way to get you to remember the material you are learning. Ulrik Juul Christensen, CEO of Area9, the developer of the data-analysis software underpinning McGraw-Hill's adaptive LearnSmart products, emphasizes his company's use of the concept of memory decay. More than two million students currently use LearnSmart's adaptive software to study dozens of topics, either on their own or as part of a course. Research has shown that those students (all of us, really) remember a new word or fact best when they learn it and then relearn it when they are just on the cusp of forgetting it. Area9's instructional software uses algorithms to predict each user's unique memory-decay curve so that it can remind a student of something learned last week at the moment it is about to slip out of his or her brain forever.

Few human instructors can claim that sort of prescience. Nevertheless, Christensen dismisses the idea that computers

could ever replace teachers. “I don't think we are so stupid that we would let computers take over teaching our kids,” he says.

Backlash

In March, Gerald J. Conti, a social studies teacher at Westhill High School in Syracuse, N.Y., posted a scathing retirement letter to his Facebook page that quickly became a viral sensation. “In their pursuit of Federal tax dollars,” he wrote, “our legislators have failed us by selling children out to private industries such as Pearson Education,” the educational-publishing giant, which has partnered with Knewton to develop products. “My profession is being demeaned by a pervasive atmosphere of distrust, dictating that teachers cannot be permitted to develop and administer their own quizzes and tests (now titled as generic ‘assessments’) or grade their own students' examinations.” Conti sees big data leading not to personalized learning for all but to an educational monoculture: “STEM [science, technology, engineering and mathematics] rules the day, and ‘data driven’ education seeks only conformity, standardization, testing and a zombie-like adherence to the shallow and generic Common Core.”

Conti's letter is only one example of the backlash building against tech-oriented, testing-focused education reform. In January teachers at Garfield High School in Seattle voted to boycott the Measures of Academic Progress (MAP) test, administered in school districts across the country to assess student performance. After tangling with their district's superintendent and school board, the teachers continued

the boycott, which soon spread to other Seattle schools. Educators in Chicago and elsewhere held protests to show solidarity. In mid-May it was announced that Seattle high schools would be allowed to opt out of MAP, as long as they replaced it with some other evaluation.

It would be easy for proponents of data-driven learning to counter these protests if they could definitely prove that their methods work better than the status quo. But they cannot do that, at least not yet. Empirical evidence about effectiveness is, as Darrell M. West, an adaptive-learning proponent and founder of the Brookings Institution's Center for Technology Innovation, has written, “preliminary and impressionistic.” Any accurate evaluation of adaptive-learning technology would have to isolate and account for all variables: increases or decreases in a class's size; whether the classroom was “flipped” (meaning homework was done in class and lectures were delivered via video on the students' own time); whether the material was delivered via video, text or game; and so on. Arizona State says 78 percent of students taking the Knewton-ized developmental math course passed, up from 56 percent before. Yet it is always possible that more students are passing not because of technology but because of a change in policy: the university now lets students retake developmental math or stretch it over two semesters without paying tuition twice.

Even if proponents of adaptive technology prove that it works wonderfully, they will still have to contend with privacy concerns. It turns out that plenty of people find

pervasive psychometric-data gathering unnerving. Witness the fury that greeted inBloom earlier this year. InBloom essentially offers off-site digital storage for student data—names, addresses, phone numbers, attendance, test scores, health records—formatted in a way that enables third-party education applications to use it. When inBloom was launched in February, the company announced partnerships with school districts in nine states, and parents were outraged. Fears of a “national database” of student information spread. Critics said that school districts, through inBloom, were giving their children's confidential data away to companies who sought to profit by proposing a solution to a problem that does not exist. Since then, all but three of those nine states have backed out.

This might all seem like overreaction, but to be fair, adaptive-education proponents already talk about a student's data-generated profile following them throughout their educational career and even beyond. Last fall the education-reform campaign Digital Learning Now released a paper arguing for the creation of “data backpacks” for pre-K–12 students—electronic transcripts that kids would carry with them from grade to grade so that they will show up on the first day of school with “data about their learning preferences, motivations, personal accomplishments, and an expanded record of their achievement over time.” Once it comes time to apply for college or look for a job, why not use the scores stored in their data backpacks as credentials? Something similar is already happening in Japan, where it is common for managers who have studied English with the adaptive-learning software iKnow to list their iKnow scores on their resumes.

This Is Not a Test

It is far from clear whether concerned parents and scorned instructors are enough to stop the march of big data on education. “The reality is that it's going to be done,” says Eva Baker, director of the Center for the Study of Evaluation at the University of California, Los Angeles. “It's not going to be a little part. It's going to be a big part. And it's going to be put in place partly because it's going to be less expensive than doing professional development.”

That does not mean teachers are going away. Nor does it mean that schools will become increasingly test-obsessed. It could mean the opposite. Sufficiently advanced testing is indistinguishable from instruction. In a fully adaptive classroom, students will be continually assessed, with every keystroke and mouse click feeding a learner profile. High-stakes exams could eventually disappear, replaced by the calculus of perpetual monitoring.

Long before that happens, generational turnover could make these computerized methods of instruction and testing, so foreign now, unremarkable, as they are for Arizona State's Hawkins and her classmates. Teachers could come around, too. Arizona State's executive vice provost Phil Regier believes they will, at least: “I think a good majority of the instructors would say this was a good move. And by the way, in three years 80 percent of them aren't going to know anything else.”

Top 50 Best Values Big Data Graduate

Institute called Big Data “the Next Frontier.” That was then; now, Big Data is so much a part of life, it’s not frontier – **it’s everyday**. Professionals in Big Data may be called data scientists, data analysts, information engineers, or any one of a number of different titles. The U.S. Bureau of Labor Statistics (BLS) classifies these workers as statisticians, computer programmers, or in other occupations, depending on their tasks. In fact, the work of the BLS is in itself a fine example of Big Data at work!

Big decisions rely on Big Data.

Also note, Linkedin, the massive professional network, ranked business intelligence, data mining, and data engineering – all Big Data skillsets – among the 25 hottest skills of 2015.

These trends are only growing as the needs for data scientists in the fields of business, finance, government, healthcare, social networking, and technology, continue to evolve.

Technology is a technique or artifact that has the effect of centralizing control, power or wealth. – *Innovation Economics*

In 2012 alone, the BLS data for numbers of employed statisticians and computer programmers were 25,570 and 316,790 respectively, with a median annual wage of $75,560 and $74,280, respectively. These numbers are more than double the median annual wage of $34,750 for all workers in 2012. And that was five years ago, when no one had heard of Big Data outside specialist circles.

The following top 50 graduate degree programs in Big Data were ranked according to Value Colleges' tried and true definition of "value" – the programs that provide the best balance of cost, quality, and return on investment. That includes four evenly weighted metrics (totaling 100%) for this ranking:

- Reputation (U.S. News)
- ROI (a median was derived between Payscale 2017-2018 College Salary Report and the US Dept. of Education's Scorecard data)
- Cost (NR total program cost reported from each institution's published rates)
- Big Data Pioneer (Schools on previous VC Big Data ranking)

While these degrees are highly specialized, they are also **highly sought after**. As the schools profiled in these rankings describe, many of them are all-inclusive, providing tools, knowledge, and networking opportunities for students to secure a very lucrative career in many fields. So bring your problem-solving, communication, teamwork, and curiosity and

put them to work for you, with this small but growing group of mathematicians, statisticians, and computer scientists. **Together, you will help solve the world's data storage, access and marketing needs, one data set at a time.**

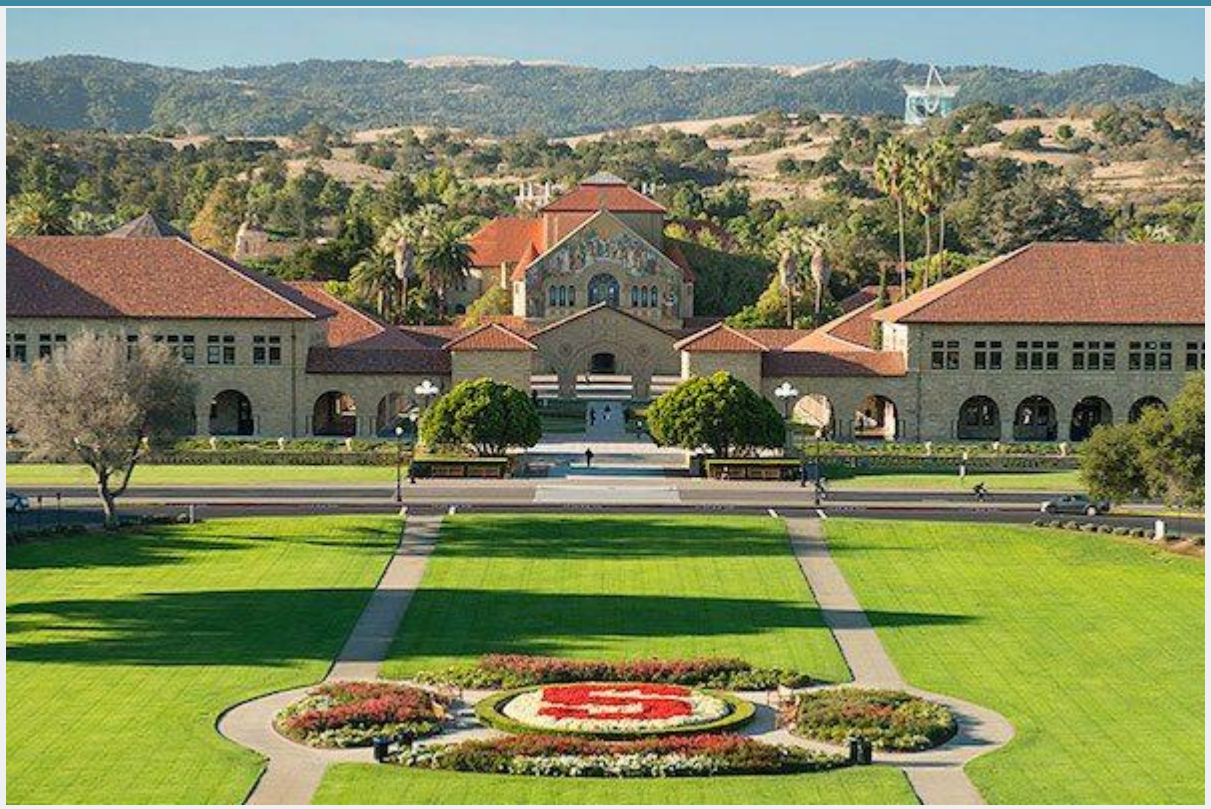
If you need the flexibility of an online degree, take a look at our ranking of the Best Value Top 50 Online Big Data Programs for 2018.

Recommended Online Master's Programs in Big Data Analytics

1. M.S. in Business Intelligence – Full Sail University
2. MBA in Market Research & Social Analytics – Saint Leo University
3. MSTIM in Data Science – Northcentral University

1

Stanford University



One of the most revered post-secondary institutions in the United States, Stanford University is officially known as Leland Stanford Junior University, and colloquially named The Farm. Established in 1885 by Leland and James Stanford, in memory of their only child Leland Jr, who died in the typhoid fever epidemic at age 15. Currently, Stanford is known as the most highly selective undergraduate program in the nation, and has often been called the world's most prestigious university. Some of the forces that strengthen the draw to this highly acclaimed school is the close proximity to the SIlicon Valley, as well as their academic contribution.

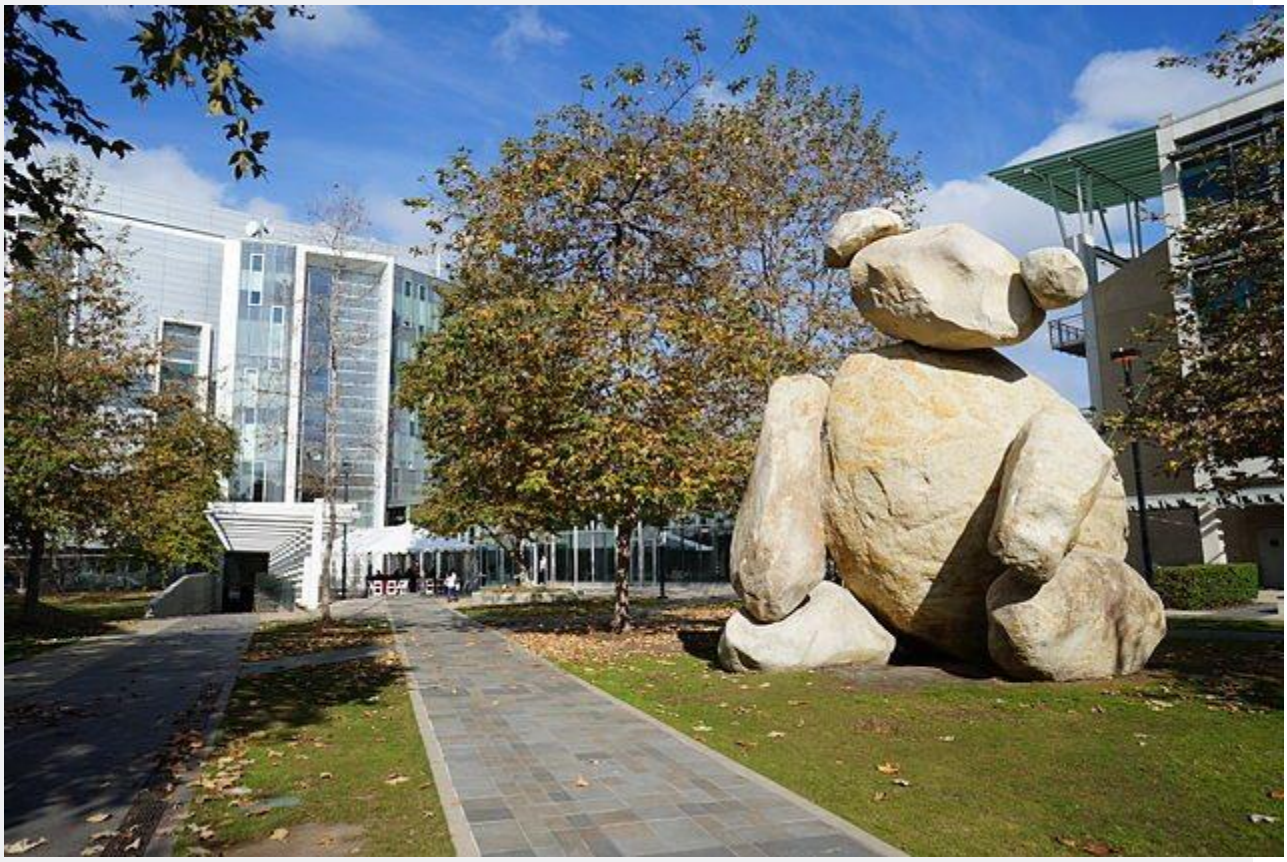
Stanford University offers some of the most prestigious graduate programming in the country, including Stanford's residential Master's of Science in Statistics: Data Science,

which is a response to the escalation of Big Data’s influence on the disciplines of engineering and the applied sciences. The program is offered through the Department of Statistics and ICME (Institute for Computational and Mathematical Engineering), as they have teamed up to create this ground-breaking computationally informed program to build experts in the the fields of Big Data. Graduates of the program are well positioned to enter the corresponding P.h.D. in Statistics, ICME, MS&E, or Computer Science.

Degree Name: M.S. in Statistics: Data Science // M.S. in Statistics
Program Duration: 33 Credit Hours // 45 credits

2

University of California, San Diego



The University of California, San Diego, which is also referred to as UC San Diego, and is a public research institution situated in the La Jolla neighborhood of San Diego on 2,141 acres beside the Pacific Ocean. Originally they were founded in 1960, beside the Scripps Institute of Oceanography. UC San Diego is the seventh oldest of the ten University of California school, and has an enrollment of about 28,000 undergraduate, and 7,000 graduate students. One exciting fact about UC San Diego, is that between the faculty, alumni and researchers, twenty Nobel prizes have been awarded cumulatively, as well as eight National Medals of Science, eight MacArthur Fellowships, two Pulitzer prizes, and three Field medals.

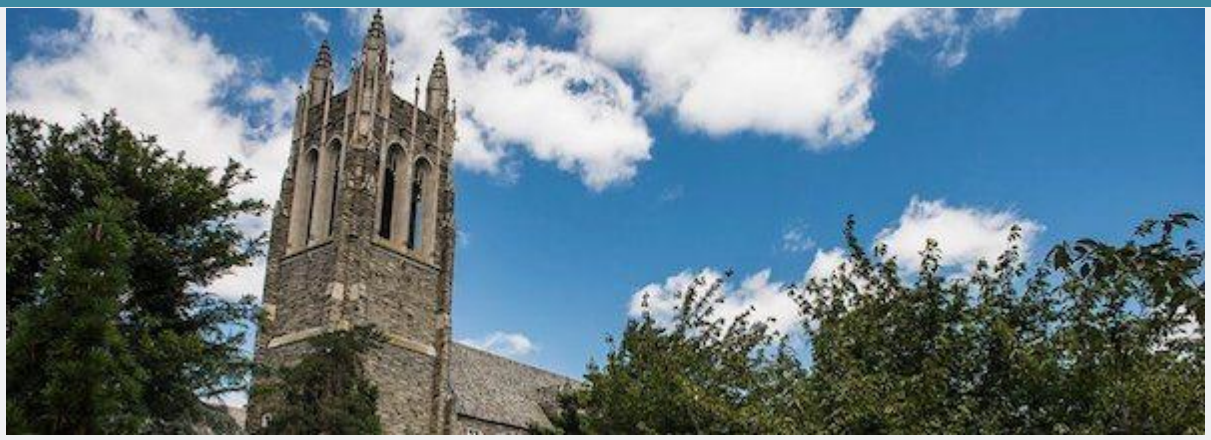
The University of California, San Diego residential graduate degree programs span many fields and casts a wide net professionally. The Master of Advanced Study Degree, in Data Science & Engineering at UCSD is an interdisciplinary course of study that converges database management, software programming, and statistics to create efficient and effective ways to identify date-driven trends, and create computational models of data. The other key component that the program embeds are the skills to present findings in visual and accessible ways, making the research digestible by larger populations. This program is taught by expert UC San Diego faculty, and is offered on an alternating weekend schedule Fridays and/or Saturdays, for student accessibility. The classes are intimate settings, allowing for deeper collaboration within cohorts, and can be completed in as little as two years. The course of study culminates with a capstone design project done in teams.

The University of California, San Diego offers a Master's of Science in Business Analytics Degree program, through the Rady School of Management. This residential graduate program offers students the critical skills for successful engagement in today's world business strategy, and analytics. There are many opportunities in this course of study for diligent experiential learning that gives students the skills to identify solutions to large scale business related issues and generate groundbreaking insights that are essential to the success business operations. Coursework also prioritizes effective and ethical communication skill building, and team management. Graduate receive extensive career support from Career Connections, with personal coaching and programs and resources.

Degree Name: Master of Advanced Study in Data Science and Engineering // M.S. in Business Analytics
Program Duration:10 courses // 50 units

3

Saint Joseph's University



Saint Joseph's University, also known as SJU, or St. Joe's is a private university a Roman Catholic, Jesuit foundation, and is situated in the Greater Philadelphia Metropolitan area, on the historic Philadelphia Main line, in Lower Merion Township. Founded by the Society of Jesus in 1851, with the original name of Saint Joseph's College, it is the seventh largest Jesuit college in the country. Enrolled are over 9200 undergraduate students, graduate and doctoral students through the Erivan K. Haub School of Business, the College of Arts and Sciences, The Program of Professional & Liberal Studies, and Haub Degree Completion Program.

The Haub School of Business at St. Joe's offers a well respected residential Business Intelligence and Analytics Master's of Science program, and has made a name for itself by creating leaders in Organizational data analysis, strategy, and decision making. The program is STEM designated with an

option of being completed on a full or part-time basis. Recently, the program was rated one of seven of the best master's programs "for the tech inclined," by inc.com. They have also consistently been ranked by the Princeton Review and the US News and World Report ranking. Applicants must submit undergraduate transcripts, a personal statement, a resume, standard letters of recommendation, and GMAT or GRE scores.

Degree Name: M.S. in Business Intelligence and Analytics
Program Duration: 30 credits

4

Texas A&M University

Texas A&M University, also known as TAMU, or A&M is a public research institution located in College Station, Texas, and is a member of the Texas A&M University System, with an endowment that remains one of the 10 biggest in the country. They are designated a land, sea, and space grant university; the only school in Texas to have all three designations. They're student population is one of the largest in Texas, and also one of the largest in the country. The main campus also makes the record of largest, as it is on 5200 acres, making breaking

national records for campus size. The main campus is home to the George Bush Presidential Library.

Texas A&M offers what is known as an extremely competitive residential Master's in Analytics program. They seek only the strongest candidates who apply with competitive GPAs, and/or GRE or GMAT Scores. They are also looking for candidates with a statistics background, and have successfully completed at least one statistics course with an A or B grade. They are also seeking students for their program who have relevant work experience under their belt, with a minimum of three years at a full-time capacity, and with organizational support from an employer that can carry from the workforce back into the graduate program.

Degree Name: M.S. in Analytics
Program Duration: 36 credits

5

Boston University, Metropolitan College

Boston University, Metropolitan College, also known as MET, is one of the seven four-year degree granting schools and colleges under the Boston University umbrella. Since 1965, when it was established, they have offered a variety of part-time programming formats, from evening study, to today's popular online and hybrid programs. Of the Metropolitan College campus full-time faculty, 94% have earned doctoral degrees. US News and World Report Ranked This institution as #4 in Best Online Graduate Criminal Justice Programs, #4 in Best Online Graduate Information Technology Programs, and #10 in Best Online Business Programs (excluding MBAs). The United States Distance Learning Association (USDLA) awarded MET the 21st Century Award for Best Practices in Distance Learning.

Boston University Metropolitan College offers a residential Master of Science in Computer Science with a concentration in Data Analytics that addresses some of the foundational Big

Data questions and innovations of our time. Students will be given the skills and resources to investigate and communicate data processing, analysis, and visualization. Graduates of the program will have cutting-edge data analysis and relevant application skills, have the necessary practice applying a variety of data visualization techniques to the explanation of real world data sets, and a firm background in web analytics and metrics. There are three program start dates in September, January, and May of each year.

Degree Name: M.S. in Computer Science with a concentration in Data Analytics
Program Duration: 40 semester credits

6

University of Virginia

The University of Virginia, which is often referred to as UVA or simply, Virginia, is a public research University that is also the flagship for the Commonwealth of Virginia. It was established in 1819 by Thomas Jefferson, the author of the constitution. There is lore attached to student culture at UVA, as they are the home of secret societies and a student-run honor code. They are also the host of several historical foundations. UVA was designated as the first and only collegiate World Heritage Site in 1987, by UNESCO. Located in Charlottesville, students come from all 50 states and 148 countries to attend.

Among UVA's exciting graduate offerings is a residential Master's of Science program in Data Science. The Data Science institute at UVA holds this exceptional program geared towards lovers of organizing and putting to use large quantities of data, while applying them to real-world scenarios. This professional program spans 11 months, starting in July and coming to completion in May, and therefore, is as rigorous as it is fast-paced. The interdisciplinary coursework is taught by core faculty from the Statistics, Computer Science, and Systems and Information Engineering Departments. There are many opportunities for experiential learning, and the conclusion of a capstone project that brings students into professional environments ripe for real-world skill building.

Degree Name: M.S. in Data Science
Program Duration: 32 semester credits

7

University of Michigan at Dearborn

The University of Michigan at Dearborn, also known as UM-Dearborn, or UM-D is a public research institution situated in Dearborn, Michigan. There are three University of Michigan campuses, the other two being Ann Arbor, and Flint. Graduates of UM are part of the largest alumni organization of it's kind in the entire world, The University of Michigan Alumni Association. There is a great deal of collaboration among students, faculty and researchers between the three campuses. UM-D offers 90 undergraduate academic majors, 28 Master's programs, and 3 doctoral degree programs. UM-D is a known active member in the Coalition of Urban and Metropolitan Universities.

The University of Michigan-Dearborn's College of Business offers a well-respected Master of Science-Business Analytics program, designed to give students a holistic skillset to create effective and accountable business plans, by interpreting and applying the outcomes of in-depth data analysis. The Financial Engineer just rated this program #28 in

the US. This residential graduate program can be taken with full or part-time status, during the fall and winter semesters, with the possibility of coursework in the summer months. Admission is granted on a rolling basis, and students have the opportunity to begin in September or May, with the possibility of a January start date for part-time students.

Degree Name: M.S. in Business Analytics
Program Duration: 30 credits

8

University of Iowa

The University of Iowa, also known as UI, or U of I, or even just Iowa, is a flagship public research University, located in Iowa City, Iowa. Established in 1847, it is officially the oldest university in the state, and is currently the second largest university in the state, the largest being Iowa State University. The campus is on an urban 1,700 acres, on the banks of the Iowa River. Carnegie Classifications has designated UI as a Doctoral University, with Highest Research Activity. An interesting fact is that IU was the founder of the Master of Fine

Arts degree, and it is known for being the home of the internationally renown Iowa Writer's Workshop. They were also the first to recognize an LGBT organization of students on campus.

The University of Iowa offers a residential graduate program in Business Analytics. The program is driven by the premise that they are offering the most currently in demand skills, as the fields requiring skilled data interpreters are growing exponentially. They offer both a certificate program, as well as the Master's of Science degree. A GRE is not required for either program. The course of study requires 30 credit hours by way of ten, 12 week classes, with no Friday or weekend in-person requirements. The program is to be completed in a three year period. University of Iowa graduate students are waived from paying the one-time university fee.

Degree Name: M.S. in Business Analytics
Program Duration: 30 credits

9

University of Florida

The University of Florida, also known as UF, or simply Florida, is a public research institution classified as very high research with public land, sea, and space grant status. The breathtaking campus spans 2,000 acres of land in Gainesville, Florida. Measured by student population, UF is the third largest campus in the state. Established in 1853, governor Thomas Brown signed a bill that accessed public funding for the health of the university. In 2016 the US News and World Report ranked UF's Engineering School 38th nationally, and in 2017, they ranked the University of Florida as tied for the ninth best public university in the US, and tied for 42nd overall, among all national public and private universities.

University of Florida offers a residential graduate program, with a degree outcome of Master's of Science in Information Systems, and Operation Management (ISOM). This professional program prepares students to flourish in the world of Big Data. An exciting fact is that 100% of domestic graduates of the program in 2016, were immediately given full-time positions, with salaries beginning at 55-110K annually. The program is STEM accredited, which excites employers. Apparently, it is not uncommon to find recruiters from consulting firms and government agencies, as well as industry firms engaging students in the on campus building Hough Hall in search of talented new employees.

Degree Name: M.S. in Information Systems and Operations Management (ISOM)
Program Duration: 40 credits

10

University of Minnesota - Twin Cities

The University of Minnesota-Twin Cities is often referred to as The University of Minnesota, the U of M, or casually just the U. This university is a public research university located on two campuses, one in Saint Paul and the other Minneapolis. The campuses are a mere 3 miles apart. The Saint Paul campus is the oldest and largest within the University of Minnesota system, and has the sixth largest student body in the US, with total enrollment of 51,147 in 2013-2014. An exciting fact is the faculty, alumni, and researchers have won a total of 25 Nobel prizes, and 3 Pulitzer prizes.

The University of Minnesota offers a Master's of Science in Data Science program, considered a "New Degree for the Modern Digital Age." This residential graduate program gives

students the fundamentals of data analytics, management, and communication skill and strategy. The traditional full-time student will complete the program in two years, with 31 credit hours. Each student will complete a comprehensive project, showing that they have grasped concepts relating the applicable use of big data in a real world professional scenario that can be applied to a myriad of disciplines. The students of the program recently participated in the inaugural Data Science poster fair.

The Carlson School of Management at the University of Minnesota Twin Cities offers a residential Master's of Science in Business Analytics. This is a full-time graduate that pairs the concepts of data analysis, with critical business information. Offered are foundational business literacy classes with embedded comprehensive data analysis skills, and the necessary skillset development of problem solving and analytic methodology. The program consists of a 45 credit hour requirement, which can be accomplished in three fast-paced rigorous semesters. Students are given hands-on opportunities to work in the context of real world problem solving, which may include consulting with actual business clients.

Degree Name: M.S. in Data Science // M.S. in Business Analytics
Program Duration: 31 credits // 45 credits

11

Bentley University

Bentley University is a private institution with a business focus. It was established in 1917, as a with a distinct focus on finance and accounting. Bentley is situated nine miles west of Boston, in Waltham, Massachusetts, though when it first opened it's doors, they were located in the Back Bay neighborhood. They offer high quality focused degree programs, including 11 distinct Business related Bachelor of Science programs, and six Bachelor of Arts degrees. The US News and World Report has ranked Bentley University second in the Top 10 Master's Universities in the North, and Top 50 Undergraduate Business Programs in 2017.

Bentley University's MCcallum Graduate School of Business offers the highly revered Master's of Science degree in Business Analytics. This program was designed as a training ground for the prospective 2 million data analytics jobs that are and will be created in the coming years. This residential graduate program answers the call of the industry by giving students the complex skills to break down Big Data, into manageable information by building competencies in statistics, economics, data science, management marketing and finance, and information management. There is a full time and part time enrollment option, and candidates are chosen based on their whole package including: work experience, learning objectives, and test scores, as well as the general application.

Degree Name: M.S. in Business Analytics
Program Duration: 33 semester hours

12

Seattle University

Seattle University, colloquially known as SU, is located in the First Hill Neighborhood of Seattle, Washington. It is a private Jesuit Catholic School and is one of 28 member institutions of the Association of Jesuit Colleges and Universities. Enrolled are almost 7,500 students within nine schools and colleges. *The Wall Street Journal/Times Higher Education* College Rankings 2018 placed SU #2 in the Northwest overall and as the top-ranked private university in the Northwest. Seattle University was ranked #7 among the West's best regional universities, according to the *2018 U.S. News & World Report "*Best Colleges" rankings.

One of the exciting residential graduate programs that Seattle University has to offer is the Albers School of Business and Economics Master's of Science in Business Analytics. Full-time students can complete the program in just one year's time. Entrance is offered once each fall. Prospective students must complete the following prerequisites prior to the start of the program: Statistics, Business Calculus, and Computer Programming, all of which can be waived with proof that a similar class has been taken. Graduates will have key competencies in describing complex questions related to business with the use of analytic models and data visualization. Students will also learn the ins and outs of business ethics related to decision making.

Degree Name: M.S. in Business Analytics
Program Duration: 45 credits

13

Texas Tech University

Texas Tech University is often called Texas Tech, Tech, or TTU and is a public university, situated in the town of Lubbock Texas. They first materialized as an university in February 10, 1923, and were known as Texas Technological College. It is the flagship location for the Texas Tech University system, which is comprised of three additional schools. The Carnegie Foundation has classified TU as The Highest Level of Research, and the fields of research are vast in reach, with topics including epidemiology, grid computing, atmospheric sciences, and wind energy. They are known for having a gorgeous campus, and were awarded the Grand Award for Excellence in Grounds Keeping, and having one of the best public art collections in the US.

Texas Tech University is the home to Rawls College of Business, the entity that hosts the residential Master of Science in Data Science. This STEM designated graduate program is distinctly one year in length, consisting of 12 classes. Using the foundations of statistics, business, and technology, students will learn how to interpret big data, and use their understandings to contribute strategy solutions to businesses. For applicants that have a minimum of five years of comprehensive and relevant work experience the GRE/ GMAT scores can be waived. There no essay requirement for the Rawls College of Business, but students will still have to complete the general graduate school personal essay.

Degree Name: M.S. in Data Science
Program Duration: 36 semester credits

14

Worcester Polytechnic Institute

1006

Worcester Polytechnic Institute, also known as WPI, is a private research University in Worcester, Massachusetts. They have an institutional focus on the instruction and research of the technical arts and applied sciences. WPI was one of the early universities in the US to focus on engineering and technology, and was incepted in 1865. Currently, they offer 14 distinct academic departments including 50 undergraduate and graduate degree programs. They are not only known for research, but the way in which it is conducted. Many faculty members work with students on extensive research projects in a myriad of disciplines, including biotechnology and information security.

Worcester Polytechnic Institute offers an exceptional residential graduate program in Data Science. As per WPI's legacy in experiential learning, students in this cutting-edge program work together, and with faculty who are experts in the field, to decipher large quantities of data, and grow the skills to apply findings to strategy building efforts, and gleaning solutions to complex problems. The program consists of 33 credit hours, and comes together at the end with a Graduate Qualifying Project (GQP), which is a semester long project, done in teams, and in tandem with an industry partner. WPI alumni and students are waived from the GRE requirement.

Degree Name: M.S. in Data Science
Program Duration: 33 semester credits

15

University of Illinois at Chicago

The University of Illinois at Chicago is often referred to as UIC, and is a state funded research-intensive university, in the category of “high research.” The campus is situated in the Near West Side Community area, just facing the Chicago Loop. It is the largest post-secondary institution in the Chicago Metro Area, with enrollment at close-to 29,000 students, enrolled in all 15 UIC Colleges. UIC is known for being the home of the largest medical school in the nation, with over $412 million in research consumption. On campus is the UIC Pavilion, which is where many events occur including concerts and basketball games.

The University of Illinois at Chicago offers the innovative Liautaud Master’s of Science in Business Analytics. This residential graduate program is bringing students up to speed in the evolving industries that require Big Data experts. Students of the program will receive an in depth education in data interpretation and organization, technology that functions to integrate the underpinnings of financial accounting and

marketing, and visual interpretation methods. This rigorous curriculum is fast-paced, as the 32 required credit hours can be completed in one year with a full-time status. Students will gain hands-on technological skills, and will work with tools such as Python and SAS.

Degree Name: M.S. in Business Analytics
Program Duration: 32 credits

16

Santa Clara University

Santa Clara University, which is also referred to as simply Santa Clara, is a private Jesuit institution located in Santa Clara, California. Established in 1851, they are the oldest operating post-secondary institution in the state of California, and has been in the same location for 166 years. Their scenic and historical campus surrounds Mission Santa Clara de Asis, which was founded in 1776, and Santa Clara has emulated it's architectural style. They are known for being a top producer in Fulbright Scholars, and have also produced Rhode's Scholars. Their decked out in red and white sports teams are called the Broncos.

Santa Clara University is the home of the Leavey School of Business, who parents their well-respected residential Business

Analytics program. This Master's in Science degree program is a response to the ever growing fields that are demanding skilled data analysts, and what better place for the university to be, but the Silicon Valley. Full time students in the program are on the fast track to completing their degree in only 15 months, and the part time program, designed professionals in the field can be completed in as little as 24 months. The coursework includes a 5 quarter STEM approved practicum, giving students hands-on professional experience.

Degree Name: M.S. in Business Analytics
Program Duration: 45 credits

17

University of Oklahoma

The University of Oklahoma, colloquially referred to as UO is public research institution located in Norman, Oklahoma. Established in 1890, prior to the existence of the state. At the time of it's inception it was built beside Indian Territory, for 17 years prior to Oklahoma's statehood. It was calculated in the fall of 2016 that there were over 31,250 students enrolled at the university, primarily at the main Norman campus. They have nearly 3,000 faculty members to accommodate the high level of

student enrollment and research activity. UO is ranked number 1 among public universities for enrollment of National Merit Scholars.

The University of Oklahoma's Gallogly College of Engineering offers a residential graduate program in Data Science and Analytics. This Master's of Science degree program is demanding and fast-paced, as students become graduates in as little as 14 months with full-time status. Part-time options of study are also available. The classes are meant to meet the student where they are at, and therefore, they offer online and hybrid options as well. Following the completion of all core courses, students are then expected to participate in a comprehensive practicum, as they expect students to employ course competencies to the organizations for which they are placed.

Degree Name: M.S. in Data Science and Analytics
Program Duration: 30-33 semester hours

18

University of Chicago

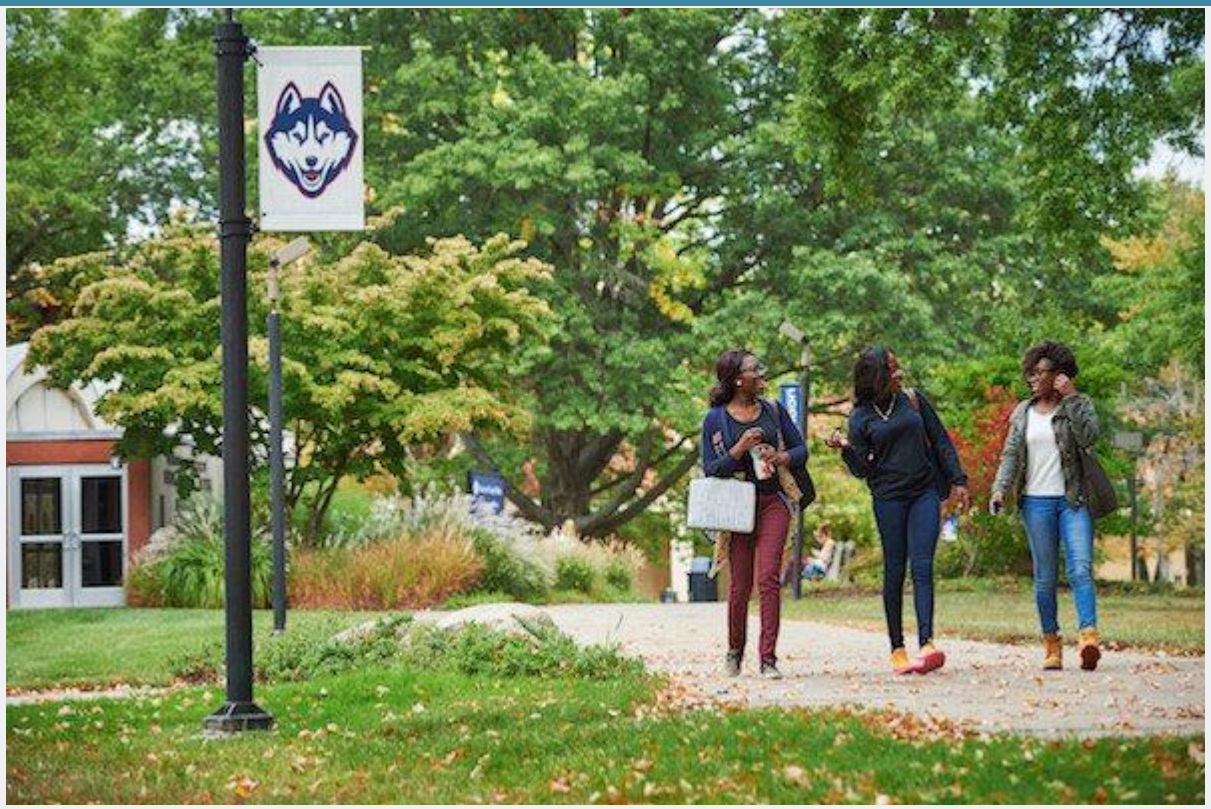
The University of Chicago, also known as U of C, Chicago, or UChicago, is a private research institution, founded in 1890. It holds many top 10 places in national and international rankings. Located in the epicenter of bustling Chicago, students have all of the resources that this international city has to offer at arms length. Not only do they offer extensive undergraduate and graduate programming, they are also known for being the home of many professional schools, including the Pritzker School of Medicine, the Booth School of Business, the Law School, the School of Social Service Administration, and the Harris School of Public Policy Studies.

The University of Chicago's Graham School for Continuing Liberal and Professional Studies is the home to the residential MSc Program in Analytics. This apropo graduate program takes students to the cutting edge of big data management and analysis, and imparts the skills to apply research to business solutions, with analytic tools. Another critical skill that is covered in the program is ethical communications and leadership strategies. The program is designed for professionals who already have experience in a variety of fields including information technology and computer science as as well as social or biological sciences. This could also be a good fit for recent college graduates who have studied social or physical sciences, and are interested in entering the field of data analytics.

Degree Name: M.S. in Analytics
Program Duration: 35 credits

19

University of Connecticut - Hartford Campus



The University of Connecticut Hartford Campus, is also called (UConn Hartford), sits in the downtown region of the state's capital. They are known for the many academic and and hands-on service and educational opportunities, many of which take advantage all that the city of Hartford has to offer. This fairly new location is integrated into the city's resources, as they hold connections to the public library system, the Wadsworth Atheneum, and the Connecticut Science Center. They have a total of 3,100 students enrolled, and 300 faculty and staff. Their close 32 mile proximity to the Storrs campus gives students easy access to more inter-campus resources.

One of the exciting offerings that the University of Connecticut Hartford Campus has to offer is the residential graduate program in Business Analytics and Project Management. This Master's in Science Program degree program is offered through

the School of Business, and can be taken on a full time or part time basis. This course of study provides students with the necessary skills to be make data driven interventions for and with small businesses and large corporations to solve complex problems, and create new opportunities. Project Management is a key component to this program, giving students a broad foundation in the discipline and connecting it to data interpretation and analysis.

Degree Name: M.S. in Business Analytics and Project Management
Program Duration: 37 semester credits

20

Georgia Tech

Georgia Tech, as this iconic institution called by those who know it well, was founded in 1885 as a part of plans for reconstruction of Atlanta after the Civil War. Originally named Georgia School of Technology, most things about Georgia Tech

have evolved with the times. Since their inception, Georgia Tech has gone from offering only mechanical engineering degrees, to at the turn of the 20th century offering electrical, civil, and chemical engineering, to finally offering the plethora of degrees and focuses that have made it consistently ranked in the top 10 universities in the country. Whether it is their exemplary and highly regarded academics, or their consistently top ranked collegiate athletics– the Georgia Tech Ramblin' Wrecks–people know this institution well, and for well-deserved reasons.

The Georgia Tech Master of Science in Analytics will take students through a graduate program that will prepare them for an illustrious career working in Big Data, with an degree that is designed to be completed in one year, with a total of 36 credit-hours required for each student. Georgia Tech's world class faculty offer residential graduate students three tracks through the program, the Analytical Tools Track, the Business Analytics Track, and the Computational Data Analytics. The differences in these three being the areas of focus the student hopes to achieve upon graduation, so students should be selective in deciding between these tracks, as the curriculum is very different.

Degree Name: M.S. in Analytics
Program Duration: 30 semester credits

21

Rensselaer Polytechnic Institute

The Rensselaer Polytechnic Institute, also known as RPI, is a private research and space grant university, located in Troy New York, with additional campuses in Hartford and Groton, Connecticut. Founded in 1824, by Stephen Van Rensselaer and Amos Eaton for explicit purpose of “the application of science to the common purposes of life.” RPI is thought to be the best technological university in the english speaking world. As of 2017 Rensselaer faculty and alumni include 6 members of the National Inventors Hall of Fame, and six National Medal of Technology winners, as well as eight Fulbright scholars and one Nobel Prize winner in Physics.

The Rensselaer Polytechnic Institute is responding to the fact that all successful businesses and corporations in the current technologically advanced era require highly skilled data analysts to be on hand. Students of this well received program will gain skills to be leaders in the field with foundational

statistics, business concepts, and data management. Students are given the opportunity to select electives tailored to their professional goals. RPI has the luxury of small class sizes, giving students the opportunity to build strong peer relationships, and have access to faculty members. One of the crowning jewels of the program is the capstone course, engaging with industry partners. There are also many embedded opportunities for mentorship from the active board of executives in the business analytics industry.

Degree Name: M.S. in Business Analytics
Program Duration: 36 credits/12 classes

Northwestern University

Northwestern University, which is often referred to as NU or simply Northwestern, is a private research institution in Evanston, Illinois, and has additional campuses in neighboring Chicago, and Doha, Qatar. They also offer academic programming with facilities in Washington D.C. and San Francisco, CA. The school was founded by John Evans, whose vision was to serve the Northwest Territory, which today includes Ohio, Indiana, Illinois, Michigan, Wisconsin, and parts of Michigan. NU is known for being highly competitive, as illustrated by the 9.0% acceptance rate in 2017, having received

over 37,000 applications. One of the schools for which they are known is their Kellogg School of Management.

Northwestern has one outstanding residential Big Data Graduate program in Analytics, that is housed in the McCormick School of Engineering. Northwestern recognizes the hyper-competitive nature of this field, and encourages and supports making industry connections while still learning the skills to become an industry leader. With a blend of instruction from renowned professors and lecturers from major corporations, students of Analytics graduate program at Northwestern can trust that not only will the instruction they receive be on the cutting edge of industry practices, they will also have the opportunity to apply their skills to current industry-inspired projects, solving problems alongside professionals in the field. After all, nothing says “hire me!” like helping solve problems in a company before you even finish school.

Degree Name: M.S. in Analytics
Program Duration: 15 months (5 Quarters)

23

University of Massachusetts - Amherst

1020

The University of Massachusetts at Amherst is also known as UMass Amherst, UMass, or simply Massachusetts. It is the flagship campus for the UMass system, and is a public land grant and research institution. They are the largest public university in the state, with more than 29,000 students enrolled, and close to 1,300 faculty members. This research University falls under the category of Highest Research Activity, ranked by the Carnegie Foundation for the Advancement of Teaching. The US News and World Report ranked Amherst in the Top Ten Best College Towns in the United States. Their sports teams are called the Minutemen and Minutewoman, with maroon, black and white colors.

UMAss Amherst's College of Information and Computer Sciences offers a well respected residential graduate degree program in Data Science. This Master's In Science degree program can be completed with 30 course credits many of which fall under the Data Science Course Requirements with a B grade or better. Some of the course offerings include: Natural Language Processing, Machine Learning, Data Visualizations, and Advanced Algorithms. There are many elective options for students to take in order to fulfill the elective requirement, and outside courses will be considered, though they must be pre-approved. In order to graduate, students must maintain a GPA of 3.0 in the 30 credit hours.

Degree Name: M.S. IN Computer Science with a concentration in Data Science
Program Duration: 30 credits

24

Michigan State University

Michigan State University, also referred to as MSU, is a public research university situated in East Lansing, Michigan. It was founded in 1855, and became a model for land-grant universities created under the Morrill Act of 1862. Originally MSU was established as Agricultural College of the State of Michigan, one of the first institutions in the country to teach scientific agriculture. They were also one of the first institutions to offer programming in packaging, hospitality business, supply chain management, and communication sciences. They have one of the largest numbers of enrolled students, approximately 55,580, from all disciplines and degree denominations.

Michigan State University's Eli Broad College of Business offers a residential graduate program in Business Analytics. This Master's in Science degree program has the tagline "Building Intelligence in a Data Driven World." It is a rigorous three semester course of study designed to train students to be leaders in the growing fields of data management process and

foundations in business analytics. Students are given hands-on project management opportunities, with access to live data sets and opportunities for application of analysis. Students will also have opportunities to build expertise in are such as: Data Mining, Applied Statistical Approaches, How to Interpret and Communicate data Analysis, and Social Network Analytics and Marketing Technology.

Degree Name: M.S. in Business Analytics
Program Duration: 30 credits

25

Case Western Reserve University

Case Western Reserve University is also known as Case Western, Case, and CWRU. It is a private institution situated in Cleveland, Ohio, and was established in 1967 by the Federation of Case Institute of Technology in partnership with Western Reserve University. Time Magazine commented on the merger,

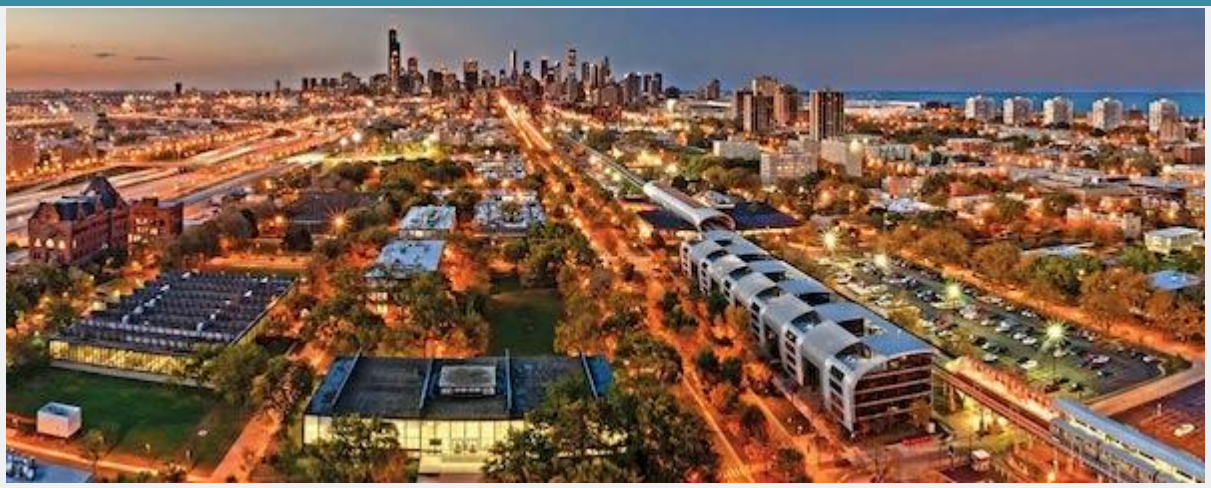
by calling it the creation of “Cleveland’s Big Leaguer University.” The US News and World Report ranked CWRU’s undergraduate program 37th among national universities. The Wall Street Journal/ Times Higher Education ranked CWRU as 32nd among all universities, and 29ths among private universities. An interesting fact is that the institution is associated with 17 Nobel laureates.

Case Western Reserve University’s Weatherhead School of Management offers a residential graduate program in Management and Business Analytics, and it is this highly acclaimed institutions newest program. This Master’s in Science degree program qualifies as a STEM-eligible program, making way for international students in a technical field to apply for a 24-month extension to their optional training program (OTP). The course of the 33 credit hour program can be completed in 11 months with full-time status. This program stands out from others as it highlights the highly in demand professional track of analytic skill mastery, as applied to marketing and interdisciplinary training.

Degree Name: M.S.M. in Business Analytics
Program Duration: 33 semester credits

26

Illinois Tech



Illinois Institute of Technology, also known as Illinois Tech, or IIT, is a private research doctoral institution located in Chicago. It's programs include: engineering, science, architecture, business, communications, industrial technology, information technology, law, and design. IIT follows the legacy of several 19th century engineering and professional education institutions in the US. What is now known as "the Million Dollar Sermon," delivered by minister Frank Wakely Gunsaulus in 1890, who referred to the fact that advanced education was generally reserved only for the most societally elite, and he said that with one million dollars we could build a technical school that would offer high quality technical educations to students from all backgrounds, in an effort to change industrial society. This sermon was the foreshadowed event of the founding of Illinois Tech.

Illinois Institute of Technology offers a residential graduate program in Data Science, driven by their College of Science. This Master's of Science program prepares students for leadership roles in the many STEM related fields involving "Big Data." This is a mathematics and statistics centered course of study with high level data analytics incorporated into the

coursework geared at solving complex business related questions of our times. Students are also given opportunities to practice high level data visualization techniques aimed at communicating data issues with real world language and scenarios. The full-time program is delivered at the Miels Campus, only minutes away from the Chicago Loop. Part-time students are given the opportunity to take their coursework online.

Degree Name: Master of Data Science
Program Duration: 33 credit

27

North Carolina State University

North Carolina State University, also known as NCSU, NC State, or simply State, is a public research institution situated in North Carolina's Capital city, Raleigh. This institution is a land, sea, and space grant institution. NC State makes up one of the

corners of what is known as “the research triangle,” with the other corners being Duke University, and The University of North Carolina, at Chapel Hill. Established in 1887, as a land-grant college, NC State was founded as the North Carolina College of Agriculture and Mechanical Arts. More than 34,000 students are currently enrolled, making it the largest post-secondary institution in the Carolinas.

It is no surprise that North Carolina State University’s Institute for Advanced Analytics offers a powerful residential graduate program in Analytics. This Master’s of Science degree granting program is dedicated to creating leaders in the professional leaders in the STEM related fields that require data analytics experts to be successful. An interesting fact is that this degree program offered at NC State is the first of it’s kind in the nation. The Harvard Business Review named the Institute for Advanced Analytics, among the best, and one of the only few sources of talent and strength in the business analytics field, along with MIT, Stanford, Berkeley, Harvard, and Carnegie Mellon.

Degree Name: M.S. in Analytics
Program Duration: 30 credits/10 months

28

Purdue University

Purdue University is touted as Indiana's Land Grant University. Founded in 1869 by West Lafayette businessman John Purdue, the university has been a touchstone in research and engineering for many international, governmental, and private enterprises for a century and a half. They now are known far and wide for their research, their contributions to the military and national security research, their athletics teams known as the boilermakers, their online writing lab (OWL), which was launched in 1994 and is available freely online to all. The OWL and includes handouts, information on subject areas, resources geared towards students in grades 7-12, and citation formatting help for MLA, APA and more.

The Master of Science in Business Analytics and Information Management at Purdue's Krannert School of Management is a full-time, 11-month program beginning in June every year, and is focuses on developing the skills of information technologies and analytical techniques. Students with benefit greatly from the

Business Information and Analytics Center within the Krannert school, a data-focused project center that centers the acquisition and analysis of Big Data initiatives. It is these and other bold plans to give students a cutting edge educational experience that makes Purdue a leader in STEM-based business education and research. The residential Masters in Business Analytics and Information Management program takes students with a strong STEM background and gives them the skills to excel in the field of Big Data. The Krannert School of Management is one of the best business schools in the nation, public or private, according to U.S. News & World Report, with an incredible job recruitment record for graduates.

Degree Name: M.S. in Business Analytics and Information Management
Program Duration: 36 semester credits

29

Louisiana State University at Baton Rouge

Louisiana State University at Baton Rouge is also known as LSU, and is technically called Louisiana State University and Agricultural and Mechanical College. This public, co-educational university is located in Baton Rouge, and was founded 1853. Originally called Louisiana State Seminary of Learning and Military Academy, LSU is now the flagship school in the Louisiana State University system. In 2017, they had over 25,000 undergraduate, and over 5,000 graduate students enrolled in their 14 schools and colleges. Many of their graduate internal institutions have received national recognition for their various disciplines. They have also become known for their exemplary research facilities.

The Louisiana State University's E.J. Ourso College of Business offers a residential graduate program in Analytics. This Master's in Science degree program's parent organization is the Stephenson Department of Entrepreneurship and Information Systems, in partnership with the Department of Experimental Statistics, as well as the Department of Marketing. This multi-faceted course of study imparts the knowledge of data driven methodology as a critical contribution to organizational decision making and effectiveness. Business analytics and business intelligence are key competencies that are explored, as well as employing the tools of information technology to solve large scale problems that impact business operations.

Degree Name: M.S. in Analytics
Program Duration: 36 semester credits

30

University of Texas at Dallas

The University of Texas at Dallas, also known as UTD, or UT Dallas, is one of the outstanding institutions that make up the University of Texas System. Their main campus is located in Richardson, Texas, in what is known as the Telecom Corridor, just 18 miles from downtown Dallas. Founded in 1961, as the Graduate Resource Center of the Southwest, and later named The Southwest Center for Advanced Studies (SCAS), which began as a research arm of Texas Instruments. The Carnegie Foundation Classifies UTD as the offering the Highest Level of Research, as well as considering undergraduate admissions "More Selective." Notable faculty and members of the alumni have received the Nobel Prize.

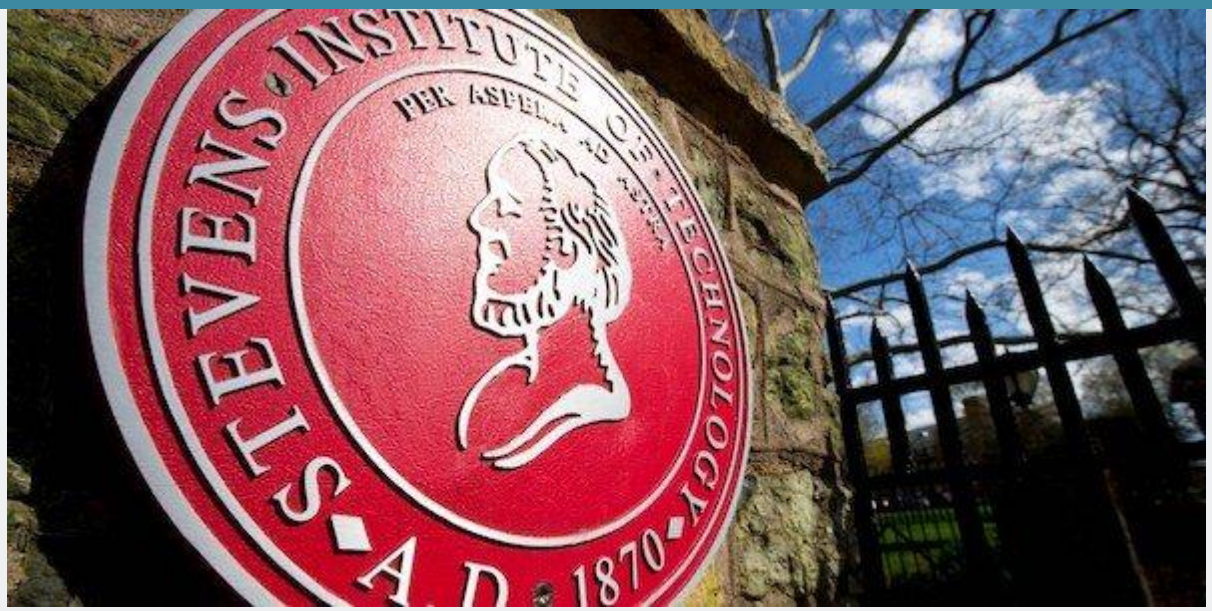
UT Dallas offers a residential graduate degree in Business Analytics, a Master's in Science degree granting course of study. This MS BUAN is a STEM-degree program that can be completed in 18-24 months at the Naveen Jindal School of Management. Students that participate in this course of study achieve access to the most cutting-edge "Big Data" education, with core courses that carve out a foundation in successful business analytical techniques, as well as data visualization. Graduates are prepared to have careers at the front lines of Data Science and Marketing. Students have many elective options including Decisions and Operations Analytics, Data

Science, Financial Analytics, Healthcare Analytics, and IT Analytics.

Degree Name: M.S. in Business Analytics
Program Duration: 36 semester credits

31

Stevens Institute of Technology



Stevens Institute of Technology, also known as SIT, is a private research university located in Hoboken New Jersey, with a satellite location in Washington D.C. Established in 1870, it is one of the earliest technological universities in the nation, and was one of the first colleges in the United States with a single focus on mechanical engineering. Current enrollment consists of more than 5,000 undergraduate and graduate students, which represent 47 of the 50 states, and 60 countries including the continents of Asia, Europe, and within Latin America. SIT is also home to national centers of the US Department of Defense, and the US Department of Homeland Security.

Stevens Institute of Technology offers a residential Business Intelligence and Analytics Master's of Science Program geared towards innovative practices of data analysis and interpretation of large amounts of complex data. Students are given the foundational skills to use data to create organizational transformation. The curriculum is said to promote the skills needed to the "big data revolution," and includes key competencies such as: data warehousing, data and text mining, social network analytics, optimization, risk analytics, and technological platforms, such as Hadoop and Data Stream Analytics. This program was ranked number seven of its kind in the US, in the TFE Times, this second year of which it placed in the top 10!

Degree Name: Business Intelligence and Analytics
Program Duration: 36 credits

32

University of San Francisco

The University of San Francisco, also referred to as USF, is a Jesuit Catholic university in the heart of San Francisco. The main campus offers a 55 acre setting, between the Golden Gate Bridge and gorgeous Golden Gate Park. The nickname for this main campus is “The Hilltop,” and a portion of this main campus is located on one of San Francisco’s major geographical features called Lone Mountain. This renowned institution offers classes at several other locations, and actually started as a

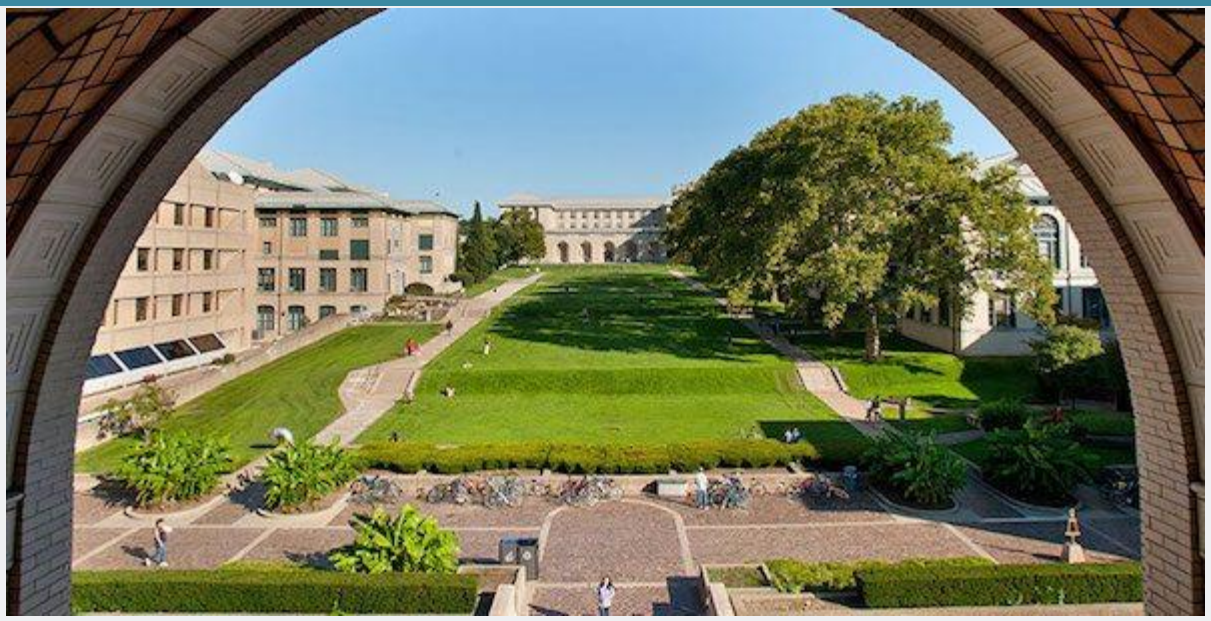
small one room schoolhouse on Market Street, in 1855, before that area became the city's downtown. 1859, is when they were accredited to start issuing degrees.

The University of San Francisco's College of Arts and Sciences offers a residential graduate program in Data Science. This Master's of Science in Data Science (MSDS) degree granting program can be completed in as little as one year with it's fast-paced curriculum that focuses on computation and mathematical fundamentals, that are critical components in the growing data driven disciplines. 98% of the MSDS graduates are gainfully employed within three months of graduation, many of which work for Google, Amazon, Williams-Sonoma, Mozilla, Capital One Labs, and Eventbrite. One of the exciting aspects about the program is the practicum requirement giving students valuable work experience, for approximately 15 hours per week, over the course of 9 months, where they tackle big world data science and analytics with companies notable in the San Francisco Bay Area.

Degree Name: Master of Science in Data Science
Program Duration: 35 units

33

Carnegie Mellon University



Carnegie Mellon University, is also referred to as CMU, and is a notable research university, located in Pittsburgh, PA. Established in 1900, by founder Andrew Carnegie, with the original name the Carnegie Technical School, in 1912 it became known as the Carnegie Institute of Technology in 1912; from that point on they began granting four year degrees. CMU has a 140 acre main campus, just 3 miles from the Pittsburgh's center. They have seven colleges and independent schools, which include the College of Engineering, the College of Fine Arts, and Dietrich College of Humanities and Social Sciences, Mellon College of Science, Tepper School of Business, H John Heinz lll College of Information Sciences and Public Policy.

Offered by Carnegie Mellon University's Heinz College is the residential Master's in Computational Data Science/ Business Intelligence and Data Science. The course of study is comprised of a synthesis of technology and business related classes, framed as a vehicle for a successful career in information systems management. This graduate program is a rigorous three semesters in length, with on campus coursework

that brings together the interdependencies of strategy, technology, and management. After 16-months, students will be eligible for graduation, and ready to re-enter the workforce as leaders in the field. Participants in the program are required to take a projects course revolving around one of the programs specializations, which will prepare students to take their ideas from academia to the workforce.

Degree Name: Masters in Computational Data Science (MCDS) // MISM Business Intelligence & Data Analytics
Program Duration: 144 eligible units of study // 162 units

34

Oklahoma State University

Oklahoma State University, which is also referred to as OKstate and OSU is a land and sun grant institution situated in the town of Stillwater, Oklahoma. Established in 1890, under the Morrill Act, it was originally referred to as the Oklahoma Agricultural and Mechanical College (Oklahoma A&M). It is the flagship institution for the Oklahoma state University system. The freshman class of 2012 was the largest class on record with 4,298 students. The Carnegie Foundation has classified OSU

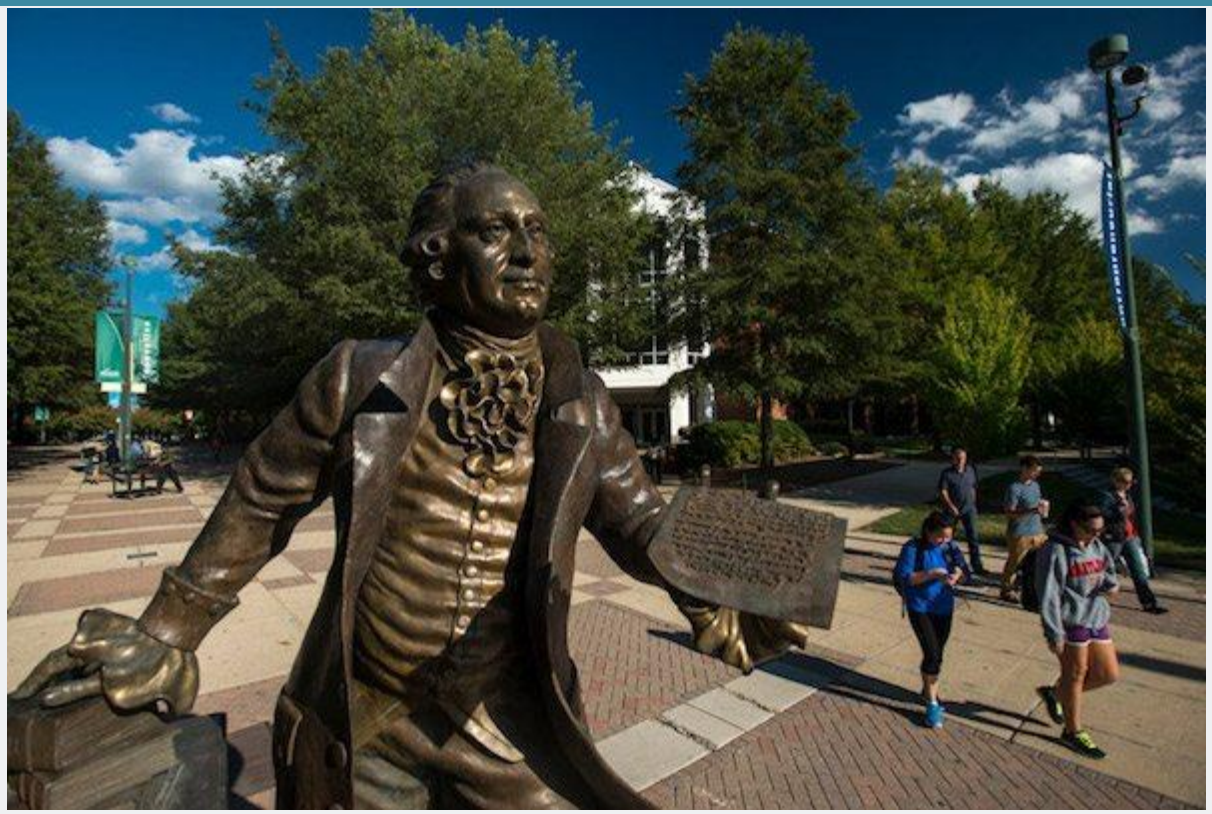
with High Research Activity. The Princeton Review has ranked OSU as one of the 120 Best Western Colleges for 2011.

Oklahoma State University's Spears School of Business, offers a residential graduate program in Business Analytics. This hands-on Master's of Science program offers experiential learning opportunities for students to grasp data analysis in a multi-platform environment, including access to SAS tools, and other programming languages such as Python and Tableau. This innovative program also emphasizes business development skills and communications. An exciting fact is that 90% of program graduates in 2017 received job offers within three months of completion, with an average annual salary above $80,000, with additional signing bonuses. Also, 100% of students received summer paid internships in the summer of 2017.

Degree Name: M.S. in Business Analytics
Program Duration: 37 credits

35

George Mason University



George Mason University, also known as Mason is the largest public research institution in the nation. Established as a University of Virginia Branch in 1942, it became an independent university in 1972. They have four Virginia Based campuses and one located in Songdo, South Korea within the bounds of the Incheon Free Economic Zone. The university is the home to 500 student groups and 41 sororities and fraternities. The Nobel Memorial Prize in Economics was won by two faculty members. Between 2009-2013, they received a 21% increase in the number of submitted applications, and has enrolled 4% more degree seeking students.

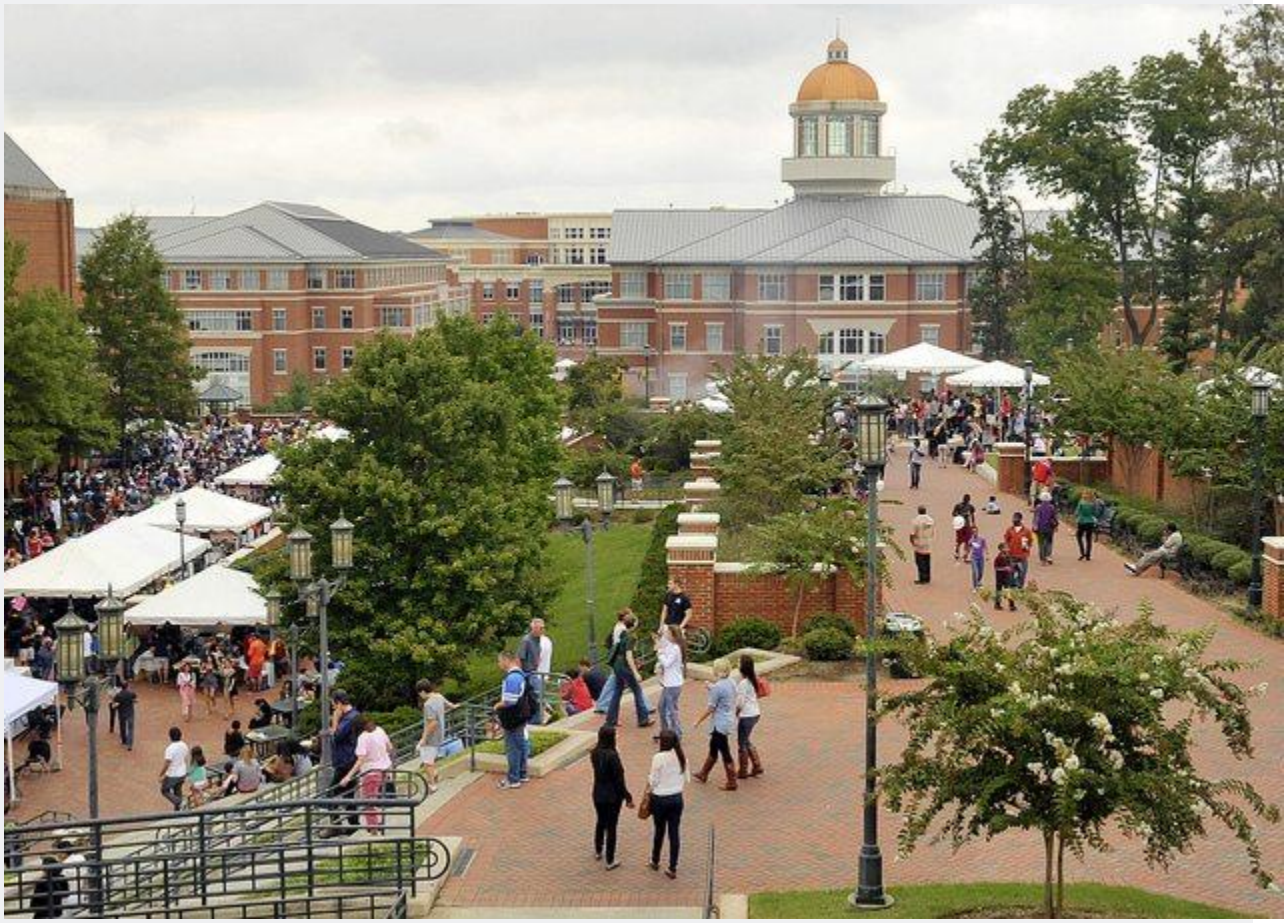
George Mason University's Volgenau's School of Engineering offers a residential graduate program in Data Analytics Engineering. This Master's in Science degree granting program gives students the conceptual and technical skills for

data driven decision making. Topics that are engaged in the core course work are: data mining, statistical models, optimization, data visualization, and risk analysis. This course of study is perfect for students who are inclined to enter the field of data science in a leadership role. Students will also be able to delve into topics that support their individual professional goals like digital forensics, business analytics, and financial engineering.

Degree Name: M.S. in Data Analytics Engineering
Program Duration: 30 semester credits

36

University of North Carolina at Charlotte



The University of North Carolina at Charlotte, is also known as UNC Charlotte, and is a public research university offering 140 undergraduate degrees programs, 64 Master's degree programs, and 23 Doctoral. UNC Charlotte has three distinct campuses: Charlotte Research Institute Campus, Center City Campus, and the main campus located in University City, and sits on 1,000 wooded acres, just 8 minutes from Uptown Charlotte. UNC Charlotte is the largest postsecondary institution in the Charlotte region, which is the third largest banking center in the nation. UNC Charlotte has grown 33.3% in enrollment in the past ten years, which translates to the fastest growing institution in the UNC System.

UNC Charlotte offers a residential graduate program in Data Science and Business Analytics. This Master's of Science program is an interdisciplinary course of study that synthesizes operations research, information sciences, data science, and statistics. Students are given the innovative data driven skillset that is more and more in demand in today's world, and they are therefore prepared to work in the data-intensive fields that include financial services, retail/ supply chain, and healthcare. Classes are taught in the evenings at the Center City Uptown campus, and the program can be completed with full time or part time status. A practicum is required for graduation, giving students the opportunities to work first-hand with leaders in the field.

Degree Name: M.S. in Data Science and Business Analytics
Program Duration: 33 semester credits

37

University of Cincinnati

The University of Cincinnati, colloquially referred to as Cincinnati or UC, is a public research institution, and part of the University of Ohio system. Established in 1819 as Cincinnati

College, it is the oldest post secondary institution in the nation. It is the second largest university in Ohio, and one of the 50 largest institutions in the United States, with 44,000 enrolled students. They were ranked by Times Higher Education in the category of top 100 universities in North America, and top 250 in the world. The US News and World Report followed suit tanking UC within the top 200 universities in the world.

The University of Cincinnati's Carl H. Lindner College of Business offers a residential graduate program in Business Analytics. This Master's in Science degree program gives students a competitive foundation in predictive, descriptive, and prescriptive analytics, which most-likely why this program has been consistently ranked within the top 10 Business Analytics Programs in the United States. This ranking is illustrated by the fact that so many graduates are in leadership positions at companies such as google, Disney, Proctor and Gamble, and Facebook. One of the program highlights is that corporate partners present seminars and recruit students on an almost weekly basis.

Degree Name: M.S. in Business Analytics
Program Duration: 33 credits

38

University of Texas at Austin

The University of Texas at Austin is also known as UT, UT Austin, or simply Texas. It is a public research institution, and the flagship of the University of Texas System. Established in 1881, this large campus is located less than one mile from the Texas state capitol building. As Texas would have it, they are the eighth largest student body in the country, with over 50,000 graduate and undergraduate students enrolled, and 24,000 staff and faculty. They are known as a major hotbed for academic research with expenditures with funds over $550 million for the 2014-2015 school year.

One of University of Texas' exemplary programs is the residential Master's in Business Analytics. This graduate offering is STEM certified, and embedded with a tremendous amount of support and industry leader recruitment from companies including Walmart and Deloitte Consulting. Networking is also a critical component of the program, as many bridges are built with industry leaders, and the career development team helps support students with this trajectory. This multi-faceted program is designed to be completed within a 10 month period. This course of study is open to students who have majors in different disciplines, but it is highly recommended that applicants complete coursework in Statistics, Probability and Calculus prior to applying.

Degree Name: M.S. in Business Analytics
Program Duration: 36 semester credits

39

University of New Hampshire at Durham

The University of New Hampshire at Durham (UNH) is a public research university in the University of New Hampshire System. Situated in the Seacoast region of the state, Durham's campus is the parent to six colleges. Established in 1866 as a land grant college in Hanover, in connection to Dartmouth, In 1893 UNH moved to Durham. UNH is the largest university in the state, and is one of only nine in the country that are land, sea, and space grant institutions. There are over 15,000 students enrolled between the Durham, Manchester, and Concord Campuses. UNH is a member of the New England Board of Higher Education.

The University of New Hampshire at Durham offers a residential Master's of Science program in Analytics. This is a full-time graduate program that spans three semesters of coursework with an interdisciplinary approach to topics such as: quantitative analysis, critical thinking, reasoning, applications, and analytics. Students are given multiple opportunities to engage with industry partners, and are given the responsibility to use their skills as applied to real-world analytic inquiries. Students are able to declare specializations such as health, business, or self-declared foci. Graduates will be prepared to sit for several SAS qualifications exams. Applications are accepted on a rolling basis, and the programs begin each May.

Degree Name: M.S. in Analytics
Program Duration: 36 credits

40

University of Rochester

The University of Rochester, which is also known as U of R, or simply Rochester is a private, nonsectarian research university, located in Rochester, New York. This institution offers six schools with a wide variety of interdisciplinary programs, all of which grant undergraduate, graduate, doctoral, and professional degrees. U of R is known for many key features, including it's Eastman School of Music, and the resident Institute of Optics, which happens to be the first educational institution dedicated to the discipline of optics, and was established in 1929. Rochester also has a laboratory for Laser Energetics, and is the host to the world's most energetic laser. Within this well-known university is the Simon Business School and the Goergen Institute.

The Masters of Science in Business Analytics program at Rochester combines foundational business knowledge with analytical expertise. The latter of these, coupled with an

economics-based curriculum, gives students the theories and practical frameworks to be a sought-after analyst. At the Simon School, faculty teach advanced programs like Python, R, and Tableau to help students leverage the Big Data insights that companies need. With an entire chart of co-curricular, career building, and social and networking opportunities built right in, students enter the Masters of Science in Business Analytics program with their sights on their future of success, the support to meet those goals, and the tools to expand their vision even further.

The Simon school is located on the University's River Campus in Rochester, New York. It was renamed after William E. Simon (1927–2000), the 63rd United States Secretary of the Treasury, in 1986. What started as a small business school has grown in the last 60 years to offer a Masters of Business Administration and Executive Masters of Business Administration, a Masters of Science degree with seven different programs students can choose from and an extensive Ph.D. program with a variety of concentrations.

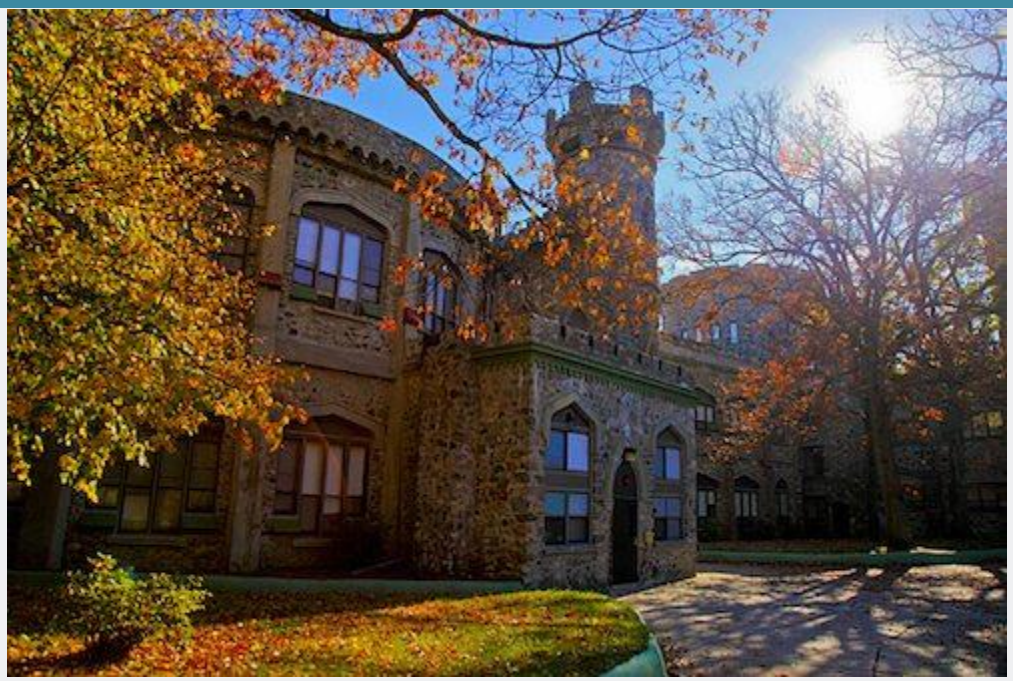
The Goergen Institute for Data Science at the University of Rochester offers a residential graduate program in Data Science, with a Master's of Science accreditation. Students with full-time status can complete the program in as little as two or three semesters. The two semester option is directed towards participants who already have a strong foundation in complex mathematics and computer science, and are equipped to take on an extra demanding course load, which includes four classes per semester, to make way for an expedient start to graduation duration. The three semester model gives students the opportunity to take three courses per semester. Students who are accepted into the program without a strong computer

science background, will be required to take a Summer Bridging Course, and will be notified of such in their acceptance letter.

Degree Name: M.S. in Data Science//M.S. in Business Analytics
Program Duration: 32 semester credits//40 credit hours

41

Brandeis University



Brandeis University is a world renowned private research university located in Waltham, Massachusetts, a mere 9 miles from Boston. Established in 1948 as a non-sectarian co-educational school, with sponsorship from the Jewish Community, they Brandeis was built on the site of the former Middlesex University. The school was named for Louis Brandeis, the first Jewish Justice of the Supreme Court. The 235 acre suburban campus had a total enrollment of 5,532 students in 2015. They have offerings that include 43 majors

and 46 minors. Well over half of the courses have fewer than 20 students in attendance, and all classes enjoy the intimacy of a low student to faculty ratio.

Brandeis University's International Business School offers a residential graduate program in Business Analytics. The close proximity to Boston gives students the advantage of being in the hub of financial and business innovation. This Master's of Science degree granting program is STEM designated, and students have many opportunities to work on faculty mentored Data Analytics Consulting Projects. A benefit of Brandeis is that all admitted students will be considered for generous scholarship opportunities. Applicants are not expected to have prior work experience in the field. The program spans the course of 11 months, and the faculty are well-respected scholars and leaders in the field.

Degree Name: M.S. in Business Analytics (MSBA)
Program Duration: 40 semester hours

42

Rutgers University

Also known as Rutgers, or RU, Rutgers University is an American public research university, which happens to be the largest postsecondary institution in the state of New Jersey. Established in November 10, 1766, chartered as Queen's College prior to the American Revolution. In 1825, it was renamed Rutgers University, after Colonel Henry Rutgers, NYC Landowner, philanthropist, and former military officer. RU is the eighth oldest college in the United States. Rutgers is the flagship state university for New Jersey. They have 9,000 members of the faculty and 175 academic departments, to support the 45,000 undergraduates, and 20,000 graduates that are currently enrolled.

Rutgers University's residential Master of Business and Science degree in Analytics program is a degree for the times in which we live; one that brings data driven decision making to the table as they develop in prospective leaders in the field the skills to master computer science, statistics, data mining, and machine learning. Students are prepared to immediately enter many industries including: business intelligence, predictive modeling, analytics, and data driven fields such as marketing, finance, biotechnology and healthcare. Prospective students without a background in STEM are encouraged to apply, though it is mandatory that they have at least one year of calculus. There are many opportunities for students to learn more about the program through online or in person information sessions, aimed to determine if one is a good fit.

Degree Name: Master of Business and Science with concentration in Analytics
Program Duration: 43 credits

43

University of Maryland, College Park

The University of College Park. also known as UM, UMD, UMCP, or simply referred to as College Park, is a public research institution. It is located just four miles from the Northeast border of Washington D.C, in the city College Park in Prince George's County, Maryland. Established in 1856, this institution is the flagship of the University of Maryland system. The enrollment is 2016 included a total of 29,000 students. They offer over 100 undergraduate majors and 120 graduate programs. UMCP is the largest in the state, including the Washington Metropolitan Area. They are a member of the Association of American Institutions.

The University of Maryland's residential graduate Master's in Science degree program in Business Analytics is a professional track, providing students with the critical tools to be leaders in the fields of data management. In this era of "big data," it is critical that the new skillset of data analysis, statistics, computational comprehension, and qualitative/ quantitative decision making. The program is designed to span a 16 month period on a full-time basis. Classes take place on campus in the afternoons and evenings. All program participants are required

to take part in a culminating capstone project in teams of 4-5 students, with faculty supervision.

Degree Name: M.S. in Business Analytics
Program Duration: 30 credits

44

University of Tennessee at Knoxville

The University of Tennessee at Knoxville, also known as UTK or UT, is a public sun and land grant university situated in Knoxville, Tennessee. Established in 1794, just 2 years before Tennessee became the 16th state in the union. UTK is the flagship campus of the University of Tennessee system. Their enrollment is 28,000 students and represents all 50 states and 100 countries worldwide. The US News and World Report ranked UT 103 among all national universities, and 46th among all public universities, in 2017. In their archives, they host

collections of papers of all three US Presidents from the state of Tennessee.

The Haslam College of Business at the University of Tennessee Knoxville, offers a residential Master's of Science Program in Business Analytics which is designed to develop students into experts in the fields related to business analytics. Students will gain hands-on experience, including a required internship that takes place in the summer, to work directly in real world scenarios to solve complex business data related problems. They are seeking applicants who have some skills in decoding quantitative knowledge, and a desire to contribute to businesses data-driven decision making. There are many financial aid opportunities, for at least one semester. It is also an almost 100% statistic that graduates had a job directly out of the gate.

Degree Name: M.S. in Business Analytics
Program Duration: 38 credits

45

Fairfield University

Fairfield University is a private co-educational undergraduate and graduate level university, nestled within New England's town of Fairfield, Connecticut. Established by the Society of Jesus in 1942, it is still one of 28 additional member institutions of the Association of Jesuit Colleges and Universities. Fairfield has distinct institutional objectives, which include ethical and religious values, and a sense of social responsibility coupled with the development of the creative and intellectual potential of their students. Their total enrollment is close to 3,500 undergraduates and 1,200 graduate students within all five of Fairfield's schools. They have been nationally recognized for their accounting, nursing, liberal arts and science programs, which have produced 62 Fulbright scholars, a MacArthur fellow, and a Guggenheim fellow.

Fairfield University offers an exciting residential Master's in Science program in Business Analytics. This graduate program provides students with the highest quality education to bring them into leadership roles in the ever-expanding world of tech. Students will learn key competencies enabling them to gain the necessary skills to analyze big sets of data and apply the findings to contribute to the success of businesses, corporations and organizations. This program is known for it's flexibility and convenience, as well as its relationships and proximity to Fortune 500 Companies in NYC, and has systematically been nationally ranked as an excellent and relevant program.

Degree Name: M.S. in Business Analytics
Program Duration: 30 credit hours

46

Syracuse University

Syracuse University, also referred to as Cuse, or CU, is a private research institution in Syracuse New York, with a history rooting back to the Genesee Wesleyan Seminary, which then became the Genesee College. Established in 1831, it was a project of the Methodist Episcopal Church, situated in the town of Lima, New York. In 1870, the college decided to undergo transformation, and this independent university in Syracuse was established. In 1920, they became explicitly non-sectarian. Currently, they are located in the University Hill neighborhood of Syracuse, just southeast of downtown. This large campus features many architectural styles, including Romanesque Revival.

The School of Information Studies at Syracuse University, also known as the iSchool, offers a residential degree program in Information Management. This program is known for their interdisciplinary approach to the innovative coursework that sets this Master's in Science degree granting program apart from the others. This course of study brings together many aspects of STEM disciplines with in demand IT Communication skills, design and implementation, along with data analysis. Students will work with faculty who are the top in their field, and bring mentorship based on personal experience. Graduates of the

program have found success in many fields including as consultants and analysts, working with some of the largest named companies in tech, and many have begun their own enterprise.

Degree Name: M.S. in Applied Data Science
Program Duration: 42 semester credits

47

Indiana University Bloomington

Indiana University Bloomington, also known as IU, and often referred to as simply Indiana, is a public research school situated in Bloomington, Indiana. They are both the flagship to the Indiana University System, as well as the largest, with a total enrollment of over 48,000 students. They are considered one of the Public Ivy schools in the country, meaning one of the top public universities in the country, with an education equivalent to that of an Ivy League School. Students attend IU from every state in the country, though 55% are from the state of Indiana. The student body also represents 165 additional

countries, as well as Puerto Rico. They are also known for an active student life, with over 750 student run organizations.

Indiana University Bloomington has a School of Informatics, Computing and Engineering, which offers the innovative residential graduate program in Data Science. This well-respected program has a 30 credit hour requirement, and aimed at imparting the critical and timely skillset for students to be successful in the evolving fields of data science. Prospective students are not required to have previous experience academically or in the workforce in the fields of STEM, as they make a note of valuing every student's orientation and experience in this interdisciplinary program. There are options to achieve this program not only as a Master's in Science degree program, but the Certificate program could be a good path for some students. There are also options to studying in a hybrid format with some in person classroom experience and some online learning.

Degree Name: M.S. in Data Science
Program Duration: 30 semester credits

48

Kennesaw State University

Kennesaw State University is also known as KSU, and is a public research institution with its main campus in Kennesaw Georgia, just 20 minutes northwest of Atlanta. Their secondary campus is located in Marietta, Georgia. They also host classes at a myriad of additional locations including Cobb Galleria Centre, Dalton State College, and Paulding County, which is in Dallas. Their current enrollment is just over 35,000 students, bringing them to the third largest post-secondary institution in

the state of Georgia, and within the top 50 largest universities in the United States. Both campuses are in suburban settings, with a total acreage of 581.

The Ph.D. in Analytics and Data Science at Kennesaw State University utilizes principles of interdisciplinary approach, in this doctoral program that bridges the divide between application and theory. Both academia and industry are heavily present in the program concepts, as faculty integrate Computer Science, Mathematics and Statistics with applications in Business, Healthcare, Transportation and Public Policy, preparing the next generation of Data Scientists. Some of the projects that doctoral students are working on in the Kennesaw State residential Doctoral Big Data program are aimed to improve processes in sales and marketing of online programs, bringing new insights in the field of healthcare such as identifying better strategies in vaccination distribution both in the U.S. and in developing countries, analyzing patterns of crime and weather to see how the two are related, finding ways to attract and sustain engagement of underrepresented groups in the STEM fields and many, many others. With faculty who hold a U.S. patent in an algorithm developed for one of these very projects, PhD students at Kennesaw State are changing the world of Big Data.

Degree Name: Applied Statistics, MS // Ph.D. in Analytics and Data Science
Program Duration: 36 credit hours // 48 credit hours + dissertation

49

University of Alabama in Huntsville

The University of Alabama in Huntsville was just a simple extension center for the University of Alabama until 1961; beginning in 1950, classes were held in the local high school. The change came with NASA – the Marshall Space Flight Center turned Huntsville into “Rocket City,” making a small Alabama city into the center of the space program with the largest NASA site in the nation. NASA called for a public research university to tend to the educational needs of their engineers and managers, and the humble UA Huntsville Center, almost overnight, became the University of Alabama in Huntsville, a major STEM research institution. Even now, UAH is ranked by U.S. News as a top-tier national university, and the Carnegie Foundation recognizes UAH as one of the most productive research universities in the nation.

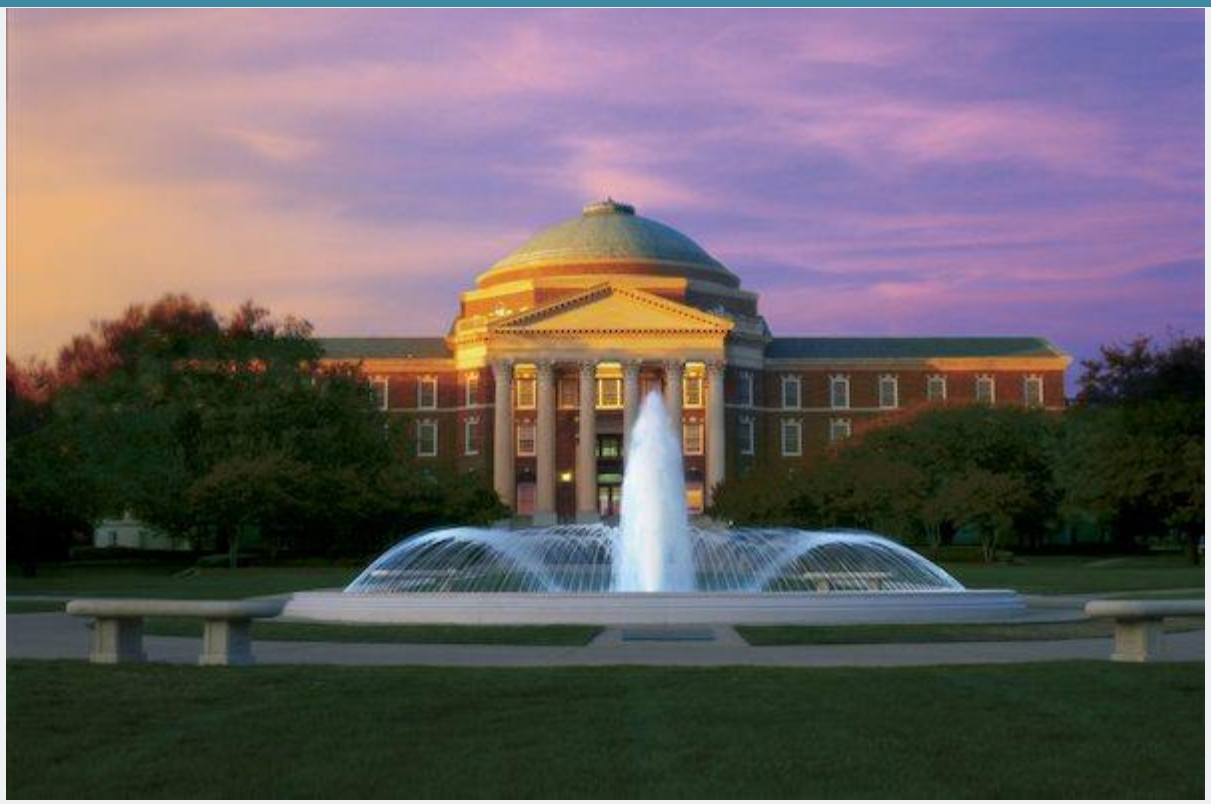
UAH, then, is an ideal place for a world-class Big Data degree program, and the UAH residential MS in Management Science-Business Analytics fits the bill. The UAH College of Business has been recognized by both the National Security Agency and

the National Science Foundation for their fusion of technology and business, and the MSMS-BA focuses on the application of data analytics to business. This program is primarily designed for working professionals who want to learn to make the best use of the tons of data available today, and practice using that data to meet customer needs. With a 30 credit program, full-time students can complete their MS in as little as one year, and with UAH's best value ranking from U.S. News and the Princeton Review, students can be sure they're getting a deal.

Degree Name: MS in Management Science-Business Analytics
Program Duration: 30 credit hours

50

Southern Methodist University



A private research university in Dallas, TX, Southern Methodist University came about almost by chance. The city of Dallas, which was growing into a major metropolitan area, had attempted to attract another college to the city, hoping to establish an institution in Dallas similar to the Southern Ivies (like Houston's Rice University). That attempt was unsuccessful, but the Methodist Church wanted to open a university in the area. It was a fortunate choice; as Dallas's population and economy boomed, SMU's access to Fortune 500 companies, finance, technology, and other industries helped make it one of the leading institutions in the nation for business and technology, ranked near the top 50 national universities by U.S. News & World Report.

With top-ranked STEM and business programs – including the Cox School of Business, one of the most selective and prestigious in the US – SMU is an ideal location for Big Data. In fact, Southern Methodist offers two data-related master's degree programs. The Department of Statistical Science at SMU offers the MS in Applied Statistics and Data Analytics, while the Cox School offers SMU's MS in Business Analytics. These two programs come at the Big Data field from two different angles. The MASDA approaches data from the technical standpoint, preparing students for the hard mathematics and computational demands of data analysis. The MS in Business Analytics, on the other hand, focuses on the organization, processing, and interpretation of data for business purposes. Both carry the SMU reputation in Dallas and around the world, making either a best value investment in a Big Data degree.

Degree Name: M.S. in Applied Statistics and Data Analytics / / M.S. in Business Analytics
Program Duration: 36 credit hours / / 33 credit hours

Master of Science in Data Science

Ready or not, we are living in a data-driven world. Every sector in our global economy is using data to make better management and policy decisions. In fact, *Forbes* and Glassdoor ranked Data Scientist as the **number one job in America** for the last four years with a median base salary of **$120,931**. By 2020, it is predicted that there will be a 28% increase in data science jobs.

Data scientists have leadership roles in business, e-commerce, energy, entertainment, manufacturing, media, science, security, telecommunications, and politics. In government, data scientists inform policy decisions. In public health, data scientists detect disease outbreaks. In finance, big data is analyzed to determine risk and make economic projections. Environmental scientists use data science to understand the impact of human activity on the planet. And in business, data scientists help improve efficiency and increase sales.

Across public, private, and nonprofit sectors, **big data is making a big difference**. Those with the right skills will lead this change. American University's MS in Data Science offers students these skills through a diverse set of courses in a flexible, innovative learning space that takes full advantage of the Washington, DC, location.

Program tracks for this degree include:

- Applied Public Affairs
- Business Analytics

- Computer Science
- Cybersecurity
- Environmental Science
- Finance
- International Economic Relations
- Investigative Journalism
- Microeconomic Analysis

This degree program is jointly administered by the AU School of Public Affairs and College of Arts and Sciences. During coursework, you'll learn statistical methods and programming, regression, statistical machine learning, and political analysis. If you're interested in this program, please email spainfo@american.edu.

American University's MS in Data Science prepares students to acquire, process, analyze, and present complex data. Students will master both the theoretical knowledge and practical skills used by data scientists in academia, industry, and government. Core courses such as **Statistical Machine Learning**, **Data Science**, and **Statistical Programming in R** will train students to clean process, visualize, and archive modern datasets, including text, imagery, and biometric data, apply machine learning algorithms to real data, and use the mathematical and statistical language of data scientists. Each student will **choose an applied field** in which they will analyze data and explain the associated issues in that field, culminating in the final semester capstone experience, DATA-793 Practicum, working directly with faculty and other researchers' ongoing projects.

45 Best Master's (MS) in Analytics in USA | Top US Universities for MS Data/Business Analytic

Analytics goes beyond traditional data analysis and business intelligence. It aims to turn big data into actionable intelligence. The Analytics market in the US is estimated to be worth USD $29.53 billion by 2019. So, undoubtedly, the US is the market leader when it comes to education and training in advanced analytics. In this post, we will look at the best **MS Analytics** (*Data Analytics & Business Analytics*) programs in the US.

45 Best Master's (MS) in Analytics in USA

As an industry, big data is growing at a rapid pace, with nearly 2.3 trillion gigabytes of data being generated every day. The data galaxy is doubling up every couple of years. So, the ability to make sense of all this data and use it to change the way we live and interact with one another is set to inevitably become a vital part of the business.

While Business Intelligence (BI) is concerned with measuring the past performance, Business Analytics (BA) programs tend to focus more on prescriptive and predictive techniques (e.g. "what should we do?" and "what happens next?"). Data Science is closely related and overlapping with data/business analytics.

MS Data Science programs are primarily focussed on data modeling & programming. In contrast, MS Analytics degrees usually concentrate on mining data in order to achieve specific business goals (e.g. more profits).

Those who are interested in pursuing MS Data Science, please refer to the article on Top Universities for MS in Data Science in USA.

If you are not sure about the right program, read MBA vs MS Business Analytics vs MS Data Science and MS Machine Learning & AI vs MS Data Science vs MS Analytics.

It's a bit hard (and sometimes not fair) to rank programs as each curriculum has its own strengths and uniqueness. Hence, I have listed the best MS Analytics programs in the US in tiers. Priority has been given on the full-time programs, curriculum, student feedback, and reviews on popular & credible portals like *Analytics Vidhya*, *Data Science Central*, *Masters in Data Science*, *KDnuggets*, *MBA Crystal Ball*, *Predictive Analytics Today*, *Masters Portal* etc.

Top US Universities for MS Data/Business Analytics – Tier 1

Massachusetts Institute of Technology (Sloan)

The one-year Master of Business Analytics is at MIT is one of the best programs in analytics in the US. The program prepares students for careers that apply and manage modern data science to solve critical business challenges.

The program is tailored for current students or recent college graduates who plan to pursue a career in the data science industry, as well as those seeking career advancement or change, especially engineers, mathematicians, physicists, computer programmers, and other high-tech professionals.

MIT MANAGEMENT SLOAN SCHOOL

MBAn Class of 2018 Profile

Class Size: 30 Students

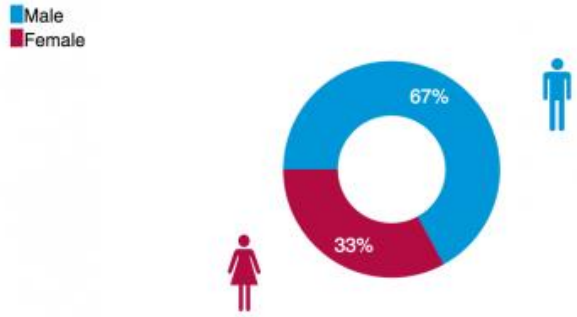
Average Work Experience*: 0-2 years

Median Undergraduate GPA: 3.9/4.0

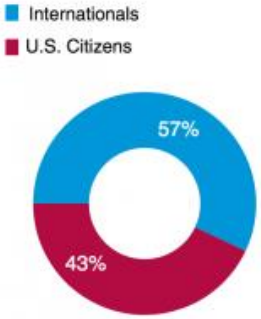
Median GRE Quant: 168/170

*including internships

Gender



Citizenship



Countries Represented

- USA
- Canada
- China
- Hong Kong
- France
- Germany
- India
- Morocco
- Netherlands
- Mexico
- Singapore
- Taiwan

Academic Backgrounds

Undergrad Universities

- Berkeley
- Brown
- Cambridge
- Duke
- Georgia Tech
- Harvard
- IIT
- ITAM
- MIT
- Northwestern
- Stanford
- UCLA
- US Air Force
- Waterloo
- Wellesley

Undergrad Degrees

- Computer Science
- Economics
- Electrical Engineering
- Finance
- Industrial Engineering
- Math
- Mechanical Engineering
- Physics

n=30 (as of June 2017)

The curriculum includes various scientific and analytics tools that build up the career of the student. The curriculum consists of five required fundamental subjects, one required Analytics Capstone project, one required project class, six three-hour modules, one pro-seminar, and a minimum of three focused electives. Partner companies include brands like BCG, McKinsey, IBM, BMW and many more.

The capstone project is a key component of the Master's program. The project gives students an opportunity to work on real-life data science problems with industry practitioners. The capstone project allows students to work in teams of two with sponsor organizations around the US and abroad for seven months.

University of Texas at Austin (McCombs)

The MS Business Analytics at UT Austin focuses on statistics, programming, and predictive analysis in a business context. It's a great program for the folks from math, computer science, and engineering backgrounds. The 10-month program prepares students for how to harness vast amounts of data and use it to build a better business.

Industry partners include names like Walmart and Deloitte. It's a great program and provides offers excellent opportunities and support to the students related to the field of big data and data analytics.

University of Southern California (Marshall)

The Masters in Business Analytics program at USC Marshall is one of the most competitive programs in the US. In 2014, the MS Business Analytics (MSBA) program at the *USC Marshall School of Business* had received 300 applications in the early round. Only 18 were admitted in that cohort.

It's an 18-month program for full-time students. Students will gain in-depth training in the analysis of big data, including statistical modeling, data management, visualization, information security, optimization, and decision-making under uncertainty.

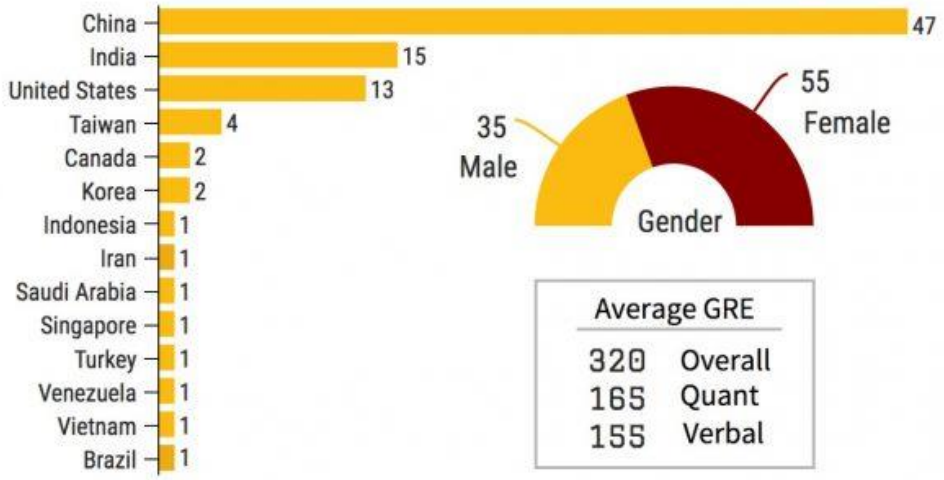
University of Southern California

MS. Business Analytics

Class Profiles 2016-2017

2017

Cohort Size: 90

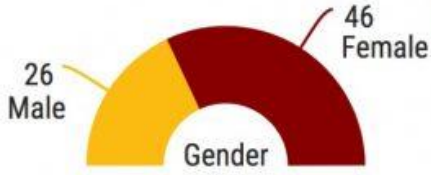


Average GRE	
320	Overall
165	Quant
155	Verbal

	Average	Min	Max
Age	25	21	66
Work Experience (months)	18	0	252

	Average
GPA	3.6
GMAT	729

2016



Cohort Size: 72

	Average
Age	25
GPA	3.6
GMAT	722
Work Experience (months)	15

Average GRE	
320	Overall
166	Quant
154	Verbal

It’s a perfect program for the managers who want to develop or sharpen their analytical skills, and recent college graduates with a strong analytical background, who are interested in pursuing a career in the exciting and fast-growing field of business analytics and data sciences.

North Carolina State University (Institute of Advanced Analytics)

North Carolina State University is considered to be one of the premium institutions for data science. It also offers a Masters in Analytics program. In fact, it is the first and preeminent analytics degree in the US. In terms of placements, acceptance rate, curriculum, and value for money, the MSA program at NCSU is placed along with MIT, Carnegie Mellon, UC Berkeley and Harvard for data science talent.

It is a 10 months master’s program that teaches all the tools and skills regarding data science. The curriculum in the NCSU changes regularly on the basis of the business and market conditions. The curriculum includes Mathematics; statistical techniques; analytics tools, methods, applications; analytics practicum. Various techniques and tools related to the data mining data analysis and data engineering are taught to the students with real-time projects. Read MS Analytics at NC State: Interview with Admissions Head.

University of Rochester (Simon)

The MS Business Analytics program at Rochester provides excellent training in Python, R, and Tableau to help you leverage the big data insights. It's a great program for the folks who are from Economics, Marketing, and Non-Engineering (but quantitative) backgrounds.

Students can finish the MS program either in 10 months (without internship) or 17 months (with internship). Please be advised that the institute does not guarantee an internship; though the careers service of the Simon Business School has got a good reputation among the employers.

University of Minnesota (Carlson)

The Master's in Business Analytics program at Minnesota teaches students how to extract insights with creative data analysis, and then apply the results in real business settings. It's a great program and highly ranked in all the popular tables like QS, Poets & Quants, etc.

As per the program website, 100% of 2015 and 2016 MS in Business Analytics graduates received offers for employment across the country

within 3 months of finishing the course. Graduates now work at McKinsey & Company, Amazon, Facebook, Ford, Expedia, Ameriprise, Google, and other standout companies.

Have a look at the student feedback:

> I firmly believe that the MSBA program at the Carlson School of Management has the one of the best MSBA curriculum's and faculty in the U.S. It has everything it takes to start one's career as a "rounded" data science professional in a top notch analytics firm. This 45- credit, 11-month highly intense program starts off with foundational courses in Statistics, Python, and Marketing Management. Marketing Analytics is a very good exit option post graduation and the Marketing Management course teaches students how data analytics is being used for decision making in the marketing industry. Besides the above three courses, the summer term also has a dedicated course to motivate and discuss various analytics case studies licensed from Harvard, Stanford, and Kellogg Business schools.
>
> Come Fall semester, the program gets very intense with courses in Exploratory Data Analysis & Visualization, Predictive Modeling, Harvesting Big Data, Databases & Data Warehousing, and Analytics Project Management. These courses are taught by the best professors in Information Systems and Decision Sciences Department – Professor Ravi Bapna, Professor Gedas Adomavicius, and Professor De Liu. This was a period of immense learning and I remember not having slept for more than seven hours on any day in this semester. Few of the concepts learnt in this semester are:

Southern Methodist University (Cox)

The Master of Science in Business Analytics program at SMU Cox provides students with the expertise and specialized Analytic skills.

Career opportunities are available in IT, marketing, operations, consulting, and other domains.

The broad and intensive curriculum across a range of management disciplines and functions explores the latest methods of translating data into relevant information for decision-makers.

The program can be finished in two semesters. The program attracts candidates from a wide range of backgrounds that include economics, math, engineering, marketing, and science.

Students have the option of **two program specialization** or tracks:

- Business Process Analytics- DB Concepts & Applications
- Customer Analytics – Marketing Research & Customer-Centric Analytics

Michigan State University (Eli Broad)

The Master's program at MSU is taught by the Faculties from the Schools of Business, computer science, and statistics. It's a one-year program and very practical-oriented in nature. The program offers students the

opportunity to work on data sets arranged with corporate and faculty support. It is also a full-time, fixed curriculum, cohort-based program to help ensure a solid team focused program.

The curriculum includes courses in business strategy, data mining, applied statistics, project management, marketing technologies, and communications taught by leading MSU faculty.

The annual base salary for the class of 2016 was $85,200. For the class of 2017, the summer internship average pay is approximately$26/hr. 32 of the 33 students landed a summer internship.

Carnegie Mellon University (Heinz College)

Within the MISM program, the Business Intelligence & Data Analytics (BIDA) track is a blended program for folks who are interested in the intersection of data analytics, management, strategy and IT.

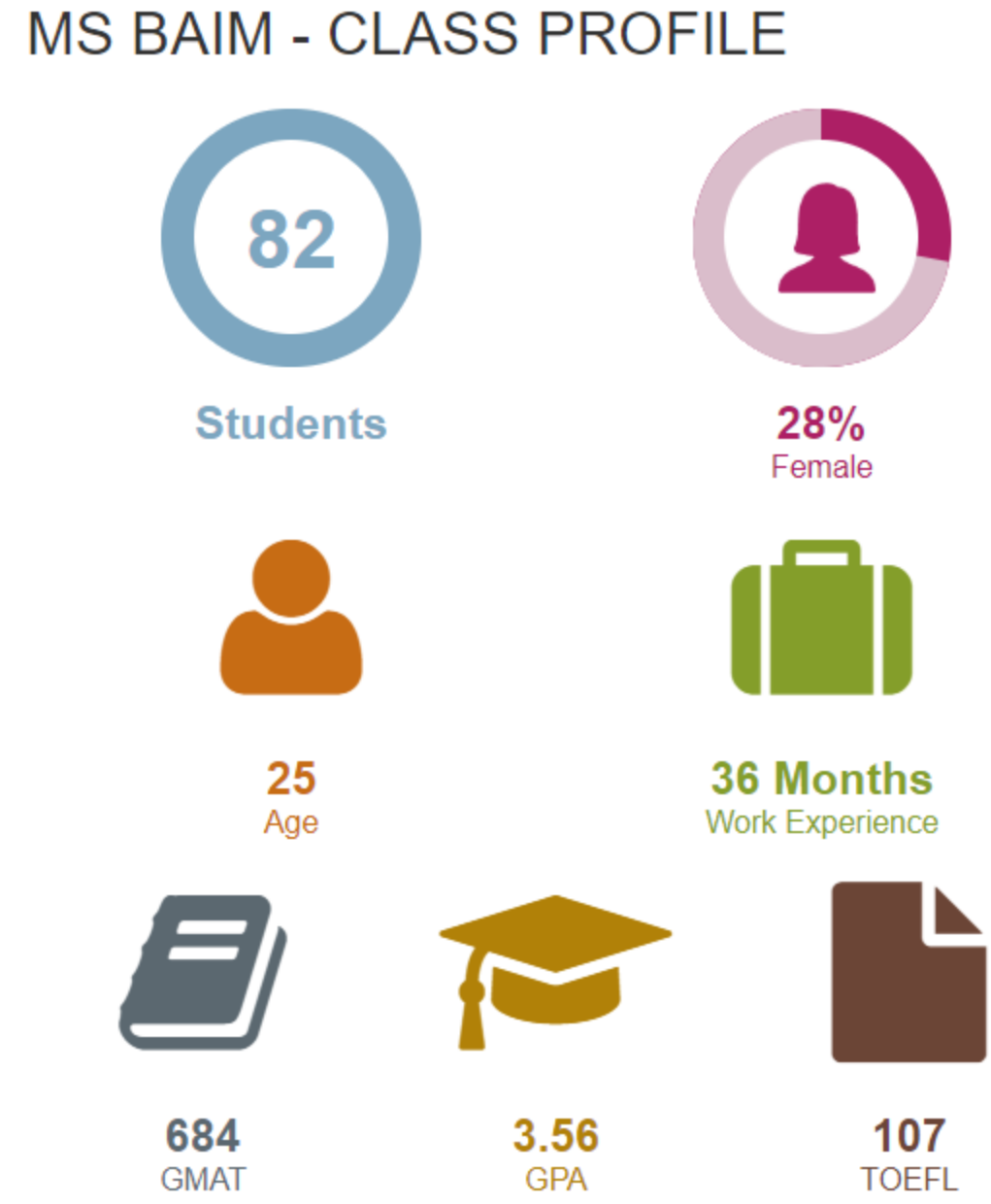
The curriculum includes advanced analytics coursework in machine learning, structured and unstructured data analytics and predictive modeling. Additionally, the range of electives is also very impressive at Carnegie Mellon.

The 16-month program includes a 10-week summer internship. The program can also be finished in 12 months if you have at least 3 years of work experience.

Purdue University (Krannert)

The MS Business Analytics and Information Management at Purdue's Krannert School of Management is a full-time, 11-month designed to provide knowledge regarding various information technologies and

analytical techniques. A rigorous curriculum will be followed by the students to understand various software tools used in the market. The program is ideal for students that will pursue analytical-type careers (e.g. Data/OR Analyst, Data Scientist, Decision Scientist, Business Consultant, etc.).



Krannert School of Management at Purdue University is specialized in developing the students from STEM backdrop to growing in prominent roles in the organization. Students will have an opportunity to work with real-time projects through experimental learning at Krannert School of Management. The school has been

using an analytical business approach to develop industry leaders for a long time.

Duke University (Fuqua)

Duke Fuqua's 10-month Master of Quantitative Management (MQM) Business Analytics focuses on analytics strategies and tools within a selected industry, such as finance, marketing, forensics, or strategy.

The program will help the student to solve actual business problems by leveraging interdisciplinary perspectives. It will also help students to understand how data is used in various aspects of the business.

The student gets training in communications and analytics in the context of all the specializations like finance, marketing, forensics, or strategy. Duke University provides a special Capstone project which is a 6-week intensive program specially designed to work with specific tools of the Major or Core subjects preferred by the students. In this program, the individual gets an opportunity to work with the real-time projects of the students.

Georgia Institute of Technology

The one-year MS analytics program at Georgia Tech is designed for students from a wide range of backgrounds – business, computer science, economics, linguistics, engineering, theater, statistics, and history.

Georgia Tech's interdisciplinary approach to analytics gives students the opportunity to learn and work in cross-disciplinary teams. Students gain hands-on experience with cutting-edge analytics techniques in statistics and operations research, big data, and high-performance computing.

The core curriculum lays the foundation; 15 hours of electives allow students to specialize in one of three analytics tracks.

Additionally, students receive funding to attend an analytics conference; have access to free cloud computing resources, software, and training; and take part in Georgia Tech's Business Analytics and Big Data Industry Forum.

Last but not least, opportunities for networking and career placement are excellent at Georgia Tech.

University of Chicago (Graham)

The Master of Science in Analytics at Chicago is a 12-month program. The curriculum emphasizes core tools and methodologies, including a large range of applications (e.g. risk analytics). Students work for three quarters on a capstone project and have the option to take workshops and short courses in R, Hadoop, Python, etc.

This program is aimed at the candidates who have got a love for analytics and intend to be the future leaders. Cohorts include recent BS graduates (science or social science) with two years of working experience; early- or mid-career professionals who want to advance up the job ladder; and mid-career leaders who need better analytical skills to improve their decision-making.

Graduates have ended up as data scientists, senior analysts and strategy directors.

Northwestern University (McCormick)

The 15-month MS Analytics program at Northwestern takes a three-pronged approach to its MSIA, with rigorous training in business skills, critical IT tools and every area of analytics — predictive (e.g. forecasting), descriptive (e.g. data mining) and prescriptive (e.g. optimization and simulation).

Courses are taught by tenured faculty from the schools of engineering, management & journalism, and industry leaders.
Additionally, the opportunities for real-world experience is awesome – 8 months for an industry project, 3 months for the summer internship and 10 weeks for the capstone.

Indiana University Bloomington (Kelley)

The MS in Business Analytics from the Kelley School of Business at Indiana is one of the most sought-after programs in the US. The 15-month Online Master of Science in Business Analytics attracts the majority of students with at least two years of work experience.

This program adds professional smarts to a bachelor's degree in IT, stats or business—one that's designed to prepare folks for technical management and consultancy positions.

Graduates went on to become a brand analytics manager, e-commerce project manager, web metrics analyst, or a consultant to senior decision-makers.

DePaul University (Driehaus & Kellstadt)

The MS Business Analytics program at DePaul is designed to meet the industry-wide need for financial professionals with exceptional data analysis skills. It's a great opportunity to study in the center of Chicago's financial district, steps away from firms like Aon, Boeing, JP Morgan Chase, and United Airlines.

DePaul University is a center for Predictive Analytics and Data mining where the students work with real-time analytics cases through sponsored projects and internship. Students will have the opportunity to specialize in growth industries and occupations including Economics, Finance, Healthcare, Hospitality, Human Resources, Marketing, and Supply Chain Management.

Read: Best Online Courses on Data Science, Machine Learning, AI, Big Data & Analytics

Best MS Data/Business Analytics Program in USA – Tier 2

New York University (Stern)

This is a great program for analytics. The only reason that I have not listed NYU in Tier 1 because it's not a full-time program. This program is explicitly aimed at senior-level professionals interested in business strategy and data science.

Its unique format blends online study with five campus sessions—called "modules"—at NYU Stern and rotating global locations (e.g. New York, London, Dubai, Shanghai, etc.). Though it's part-time, the degree can be completed in 12 months.

Arizona State University (W.P. Carey)

The 9-month MS Business Analytics program at ASU is designed for recent college graduates with a strong quantitative background. Courses are available in Enterprise Analytics, Applied Analytics, Data Mining, Business Analytics Strategy, and Marketing Analytics.

University of Notre Dame

The Notre Dame Masters in Business Analytics program is a special program which enables the student to understand both the business and analytics. Various analytical principles and techniques are taught to understand the data which ultimately results in decision making. Master's in Business Analytics is a 12 months analytics program which has more importance towards the integration of business and data analytics so that the student will have the power of transforming the data, build up the ideas and provide insights to make a better decision in the organization.

George Washington University

The George Washington School of Business offers M.S. in Business Analytics. College's faculty are committed to providing advanced scientific principles as a curriculum to make the students understand the deep foundations of various analytical methodologies and technologies which help them to take decisions in a real-time project. It is an intensive 10 months program where the students work more with live projects. The college also offers the part-time format of the course which has a course duration of 24 months.

Drexel University

Drexel LeBow's Master of Science in Business Analytics program is a program designed specially in Decision Sciences and MIS department. The students who are interested in being a part of the Drexel University will learn various quantitative methods, strategy, and operations with fact-based insights and an in-depth understanding of business performance using statistical explanatory and predictive modeling and quantitative analysis of data.

Read: MS Business Analytics in USA: Admissions, Academics and Job Prospects for International Students

Bentley University

Since the last 20 years, Bentley University has been in the industry of data analytics training. Bentley University helps in developing the distinctive skills and the knowledge to translate the data by combining both the analytics and business together which in turns adds up the value in decisions taken in the organization. It offers 3 Major programs in Data Analytics like Audit Analytics, Business Analytics & Marketing Analytics which allows the students to elect a core subject as a career elective.

University of Texas at Dallas

The Master of Science in Business Analytics (MS BUAN) provides students with an in-depth foundation in business analytics and data science. The program educates the students with various aspects of data science, big data, and business analytics. Finance, Healthcare, IT, Marketing and operations are the core courses at the University of Dallas which provide the base foundation of all the tools and techniques used in the core subjects. The association with the core subjects in business applications is considered to be the specialty of the university.

Read: UT Dallas and UConn Admits for Indian Applicant with NGO Experience

University of Cincinnati

The Master of Science in Business Analytics program is a 12 months program at Lindner College of Business, the University of Cincinnati is ranked among top 20 Universities in the world by QS Ranking and top 10 business analytics program in the US. Master of Science in Business Analytics is a program at the University of Cincinnati which provides expert knowledge regarding descriptive, predictive and prescriptive analytics. Various skills which are necessary to understand the data like data visualization, statistical modeling, data mining, optimization, and simulation are taught to the students to analyze and understand various datasets and insights. Real-time projects are assigned to the students so that the student can also be a part of corporate learning.

University of Connecticut

MS in Business Analytics and Project Management at the University of Connecticut is specially designed to meet the growing demand for professionals who can handle both business analytics and project management skills in the organization. MS in Business Analytics and

Project Management is accredited by SAS Institute and the Project Management Institute (PMI) which indicates the value of the curriculum at the University of Connecticut. The university ranked top 6th college for predictive analysis. Many tools, techniques, and technologies are taught to the students so that they can understand, analyze and predict the data with appropriate scientific explanations. The university offers both part-time and full-time courses to the students along with real-time projects and internships.

Other Top Universities for MS Analytics in USA (Tier 3)

- Louisiana State University – MS Analytics
- Emory University – MS Business Analytics
- Harrisburg University of Science and Technology – MS Analytics
- Rensselaer Polytechnic Institute (Lally) – MS Business Analytics
- Washington (Olin) – Masters in Business Analytics
- Boston University – MS Computer Science with Data Analytics Concentration
- Rutgers – MBS Analytics: Discovery Informatics & Data Sciences
- Texas A&M – MS Analytics
- University of Maryland – MS Business Analytics

- Columbia University – MS Information and Knowledge Strategy
- Cornell University – MEng Data Analytics Track
- University of Iowa (Tippie) – MS Business Analytics
- Virginia (McIntire) – MS Commerce: Business Analytics Track
- UC San Diego (Rady) – Master of Science in Business Analytics
- UC Davis – MS Business Analytics
- San Francisco State Univesity – MS Business Analytics
- University of Florida – MS ISOM: Business Intelligence & Analytics Specialization
- Fordham University – MS Business Analytics
- University of Miami – MS Business Analytics
- University of Colorado Denver – MS Business Analytics

The 20 best schools to study big data analytics

Big data analytics is one of the fastest growing fields in enterprise technology. Here are the best places to study in the field.

Big data, data analytics, data science--whichever direction you want to take it, there's no denying that the technology fields around data and data services are hot. Data is the currency of the enterprise, and more organizations than ever are trying to get some value out of it.

Many folks in tech careers who want to pursue a job in big data often head back to school to get the proper credentials. Value Colleges, an online publication that compares universities, recently released its list of the Top 50 Best Value Big Data Graduate Programs of 2016, where it details the best schools to learn about data.
Here are the top 20.

1. Carnegie Mellon University
Carnegie Mellon University's Heinz College houses the School of Information Systems and Management, where students can earn a master's degree in information systems management with a focus in data analytics. The school also offers a program for aspiring CISOs.

2. Stanford University

One of the building blocks of Silicon Valley, it's no surprise that Stanford is on this list. Stanford's master of science degree in statistics with a concentration in big data, combined with their ties to the entrepreneurial community, make it a great fit for someone who wants to work in startups.

3. Santa Clara University

Santa Clara University is one of the oldest schools in the state of California. Its Leavey School of Business offers students a Master of Science degree in business analytics, which is one of the best in the country.

4. University of Michigan - Dearborn
Michigan isn't just the home of the biggest automakers; it's also a good place to learn about big data. The university's

Dearborn campus offers a master of science in business analytics, and is fairly well priced compared to the private universities on the list.

5. University of Texas at Dallas
They say everything's bigger in Texas, and that includes the data. The University of Texas at Dallas Jindal School of Management is a great option to pursue a graduate degree in analytics, as the surrounding Dallas economy is home to many major businesses.

6. University of Virginia
The University of Virginia is known, all-round, as a quality school, and its Data Science Institute is a great place to study in the field. Its Master of Science in Data Science (MSDS) is only 10 months long, offering a quick turnaround for professionals.

7. University of Florida
Coming in at no. 7 is the University of Florida, where students can pursue a master of science in information systems and operations management. The university is also routinely ranked as a top business school, making it a great option for those who wish to also pursue management down the line.

8. Purdue University
Considered by some to be the Ivy League school of the Midwest, Purdue is a great place to study data analytics.

9. University of Maryland

One the whole, the University of Maryland is known as a great public university, but it also has a great graduate program for aspiring data scientists. Also, being close to Washington, DC means that government jobs are a great place to start your career.

10. Georgia Institute of Technology
The Georgia Institute of Technology Master of Science in Analytics is a great interdisciplinary degree, and nearby Atlanta is home to a booming tech job market.

11. University of California - San Diego
Most of the schools in the University of California fold are known or being tech-centric, but UCSD's Rady School of Management where students can get an in-depth graduate degree business analytics.

12. Bentley University
Bentley University is known for its history in accounting, but it also is a data powerhouse. The school was also ranked by Value Colleges as a top school for UX design, for students looking to broaden their skills.

13. Oklahoma State University
At Oklahoma State University, students can take two paths to data dominance: a master's degree in management information systems or in business analytics. The low cost of tuition is an added perk.

14. Northwestern University
Northwestern University's McCormick School is a great option for an advanced degree in analytics. The school is

also known for business and journalism, which could provide additional skills in management or data reporting.

15. North Carolina State University
The birthplace of SAS, North Carolina State University has been working with data for a long time. Also, the state's Research Triangle is home to many tech companies that are likely looking for data analysts.

16. Indiana University - Bloomington
Indiana University's main campus in Bloomington hosts the School of Informatics and Computing, where students can get a master's degree in data science in only 30 credit hours. The school is also known for its focus on business, making it a great option for students who are also looking to go into management.

17. Worcester Polytechnic Institute
Though a smaller school, Worcester Polytechnic Institute in Worcester, Massachusetts offers one of the best master's degree in data science and analytics.

18. University of Massachusetts - Amherst
Regarded as one of the best research universities in the nation, Univerity of Massachusetts Amherst is a great place to study data science as well.

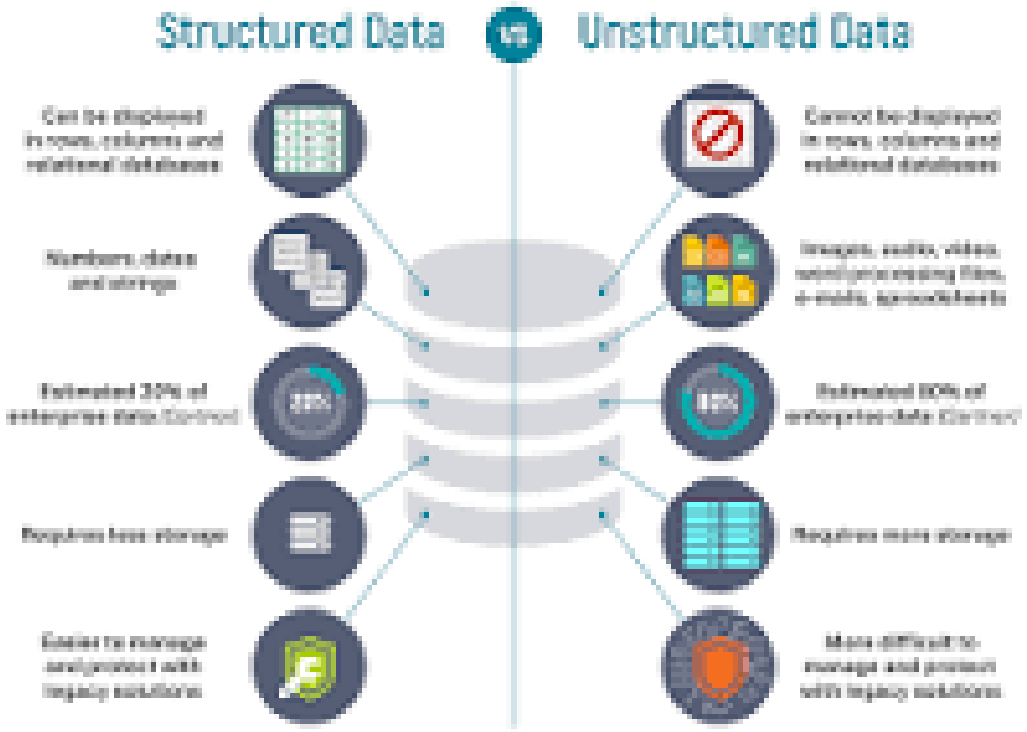
19. Seattle University
Like many schools on this list, Seattle University offers a master's degree in business analytics. However, the close proximity to tech giants like Microsoft and Amazon make it a strategic geographic location to start your big data journey.

20. George Mason University

Rounding out the top 20 is George Mason University. Known for its work in computing, business, and economics, George Mason is a great option for students who want to use data analytics in business environment.

Chapter 10:

UNSTRUCTURED DATA



Unstructured data

From Wikipedia, the free encyclopedia

Unstructured data (or **unstructured information**) is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.

In 1998, Merrill Lynch cited a rule of thumb that somewhere around 80-90% of all potentially usable business information may originate in unstructured form.[1] This rule of thumb is not based on primary or any quantitative research, but nonetheless is accepted by some.[2] Other sources have reported similar or higher percentages of unstructured data.[3][4][5]

As of 2012, IDC and Dell EMC project that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010.[6] More recently, IDC and Seagate predict that the global data sphere will grow to 163 zettabytes by 2025 [7] and majority of that will be unstructured. The Computer World magazine states that unstructured information might account for more than 70%–80% of all data in organizations.[1]

Background

The earliest research into business intelligence focused in on unstructured textual data, rather than numerical data.[8] As early as 1958, computer science researchers like H.P. Luhn were particularly concerned with the extraction and classification of unstructured text.[8] However, only since the turn of the century has the technology caught up with the research interest. In 2004, the SAS Institute developed the SAS Text Miner, which uses Singular Value Decomposition (SVD) to reduce a hyper-dimensional textual space into smaller dimensions for significantly more efficient machine-analysis.[9] The mathematical and technological advances

sparked by machine textual analysis prompted a number of businesses to research applications, leading to the development of fields like sentiment analysis, voice of the customer mining, and call center optimization.[10] The emergence of Big Data in the late 2000s led to a heightened interest in the applications of unstructured data analytics in contemporary fields such as predictive analytics and root cause analysis.[11]

Issues with terminology

The term is imprecise for several reasons:

1. Structure, while not formally defined, can still be implied.
2. Data with some form of structure may still be characterized as unstructured if its structure is not helpful for the processing task at hand.
3. Unstructured information might have some structure (semi-structured) or even be highly structured but in ways that are unanticipated or unannounced.

Dealing with unstructured data

Techniques such as data mining, natural language processing (NLP), and text analytics provide different methods to find patterns in, or otherwise interpret, this information. Common techniques for structuring text usually involve manual tagging with metadata or part-of-speech tagging for further text mining-based structuring.

The Unstructured Information Management Architecture (UIMA) standard provided a common framework for processing this information to extract meaning and create structured data about the information.[12]

Software that creates machine-process able structure can utilize the linguistic, auditory, and visual structures that exist in all forms of human communication.[13] Algorithms can infer this inherent structure from text, for instance, by examining word morphology, sentence syntax, and other small- and large-scale patterns. Unstructured information can then be enriched and tagged to address ambiguities and relevancy-based techniques then used to facilitate search and discovery. Examples of "unstructured data" may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects (e.g. in files or documents …) that they have structure and are thus a mix of structured and unstructured data, but collectively this is still referred to as "unstructured data".[14] For example, an HTML web page is tagged, but HTML mark-up typically serves solely for rendering. It does not capture the meaning or function of tagged

elements in ways that support automated processing of the information content of the page. XHTML tagging does allow machine processing of elements, although it typically does not capture or convey the semantic meaning of tagged terms.

Since unstructured data commonly occurs in electronic documents, the use of a content or document management system which can categorize entire documents is often preferred over data transfer and manipulation from within the documents. Document management thus provides the means to convey structure onto document collections.

Search engines have become popular tools for indexing and searching through such data, especially text.

Approaches in natural language processing

Specific computational workflows have been developed to impose structure upon the unstructured data contained within text documents. These workflows are generally designed to handle sets of thousands or even millions of documents, or far more than manual approaches to annotation may permit. Several of these approaches are based upon the concept of online analytical processing, or OLAP, and may be supported by data models such as text cubes.[15] Once document metadata is available through a data model, generating summaries of subsets of

documents (i.e., cells within a text cube) may be performed with phrase-based approaches.[16]

Approaches in medicine and biomedical research

Biomedical research generates one major source of unstructured data as researchers often publish their findings in scholarly journals. Though the language in these documents is challenging to derive structural elements from (e.g., due to the complicated technical vocabulary contained within and the domain knowledge required to fully contextualize observations), the results of these activities may yield links between technical and medical studies[17] and clues regarding new disease therapies.[18] Recent efforts to enforce structure upon biomedical documents include self-organizing map approaches for identifying topics among documents,[19] general-purpose unsupervised algorithms,[20] and an application of the CaseOLAP workflow[16] to determine associations between protein names and cardiovascular disease topics in the literature.[21] CaseOLAP defines phrase-category relationships in an accurate (identifies relationships), consistent (highly reproducible), and efficient manner. This platform offers enhanced accessibility and empowers the biomedical community with phrase-mining tools for widespread biomedical research applications.[21]

References

^ ***Shilakes, Christopher C.; Tylman, Julie (16 Nov 1998). "Enterprise Information Portals" (PDF). Merrill Lynch.***

---

1. ^ *Grimes, Seth (1 August 2008). "Unstructured Data and the 80 Percent Rule". Breakthrough Analysis - Bridgepoints. Clarabridge.*
2. ^ *Gandomi, Amir; Haider, Murtaza (April 2015). "Beyond the hype: Big data concepts, methods, and analytics". International Journal of Information Management.* ***35*** *(2): 137–144. doi:10.1016/j.ijinfomgt.2014.10.007. ISSN 0268-4012.*
3. ^ *"The biggest data challenges that you might not even know you have - Watson". Watson. 2016-05-25. Retrieved 2018-10-02.*
4. ^ *"Structured vs. Unstructured Data". www.datamation.com. Retrieved 2018-10-02.*
5. ^ *"EMC News Press Release: New Digital Universe Study Reveals Big Data Gap: Less Than 1% of World's Data is Analyzed; Less Than 20% is Protected". www.emc.com. EMC Corporation. December 2012.*

6. ^ *"Trends | Seagate US"*. *Seagate.com. Retrieved 2018-10-01.*
7. ^ Jump up to:[a] [b] *Grimes, Seth. "A Brief History of Text Analytics". B Eye Network. Retrieved June 24, 2016.*
8. ^ *Albright, Russ. "Taming Text with the SVD" (PDF). SAS. Retrieved June 24, 2016.*
9. ^ *Desai, Manish (2009-08-09). "Applications of Text Analytics". My Business Analytics @ Blogspot. Retrieved June 24, 2016.*
10. ^ *Chakraborty, Goutam. "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining" (PDF). SAS. Retrieved June 24, 2016.*
11. ^ *Holzinger, Andreas; Stocker, Christof; Ofner, Bernhard; Prohaska, Gottfried; Brabenetz, Alberto; Hofmann-Wellenhof, Rainer (2013). "Combining HCI, Natural Language Processing, and Knowledge Discovery – Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field". In Holzinger, Andreas; Pasi, Gabriella (eds.). Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science. Springer. pp. 13–24. doi:10.1007/978-3-642-39146-0_2. ISBN 978-3-642-39146-0.*

12. ^ *"Structure, Models and Meaning: Is "unstructured" data merely unmodeled?"*. *InformationWeek. March 1, 2005.*
13. ^ *Malone, Robert (April 5, 2007). "Structuring Unstructured Data"*. *Forbes.*
14. ^ *Lin, Cindy Xide; Ding, Bolin; Han, Jiawei; Zhu, Feida; Zhao, Bo (December 2008). Text Cube: Computing IR Measures for Multidimensional Text Database Analysis*. *2008 Eighth IEEE International Conference on Data Mining.* *IEEE. CiteSeerX 10.1.1.215.3177. doi:10.1109/icdm.2008.135. ISBN 9780769535029.*
15. ^ Jump up to:[a] [b] *Tao, Fangbo; Zhuang, Honglei; Yu, Chi Wang; Wang, Qi; Cassidy, Taylor; Kaplan, Lance; Voss, Clare; Han, Jiawei (2016). "Multi-Dimensional, Phrase-Based Summarization in Text Cubes" (PDF).*
16. ^ *Collier, Nigel; Nazarenko, Adeline; Baud, Robert; Ruch, Patrick (June 2006). "Recent advances in natural language processing for biomedical applications". International Journal of Medical Informatics.* ***75*** *(6): 413–417. doi:10.1016/j.ijmedinf.2005.06.008. ISSN 1386-5056. PMID 16139564.*
17. ^ *Gonzalez, Graciela H.; Tahsin, Tasnia; Goodale, Britton C.; Greene, Anna C.; Greene, Casey S. (January 2016). "Recent Advances and*

Emerging Applications in Text and Data Mining for Biomedical Discovery". *Briefings in Bioinformatics*. ***17*** *(1): 33–42*. *doi:10.1093/bib/bbv087*. *ISSN 1477-4054*. *PMC 4719073*. *PMID 26420781*.

18. ^ *Skupin, André; Biberstine, Joseph R.; Börner, Katy (2013). "Visualizing the topical structure of the medical sciences: a self-organizing map approach"*. *PLOS One*. ***8*** *(3): e58779. doi:10.1371/journal.pone.0058779*. *ISSN 1932-6203*. *PMC 3595294*. *PMID 23554924*.
19. ^ *Kiela, Douwe; Guo, Yufan; Stenius, Ulla; Korhonen, Anna (2015-04-01). "Unsupervised discovery of information structure in biomedical documents"*. *Bioinformatics*. ***31*** *(7): 1084–1092. doi:10.1093/bioinformatics/btu758*. *ISSN 1367-4811*. *PMID 25411329*.
20. ^ Jump up to: ***a*** ***b*** *Liem, David A.; Murali, Sanjana; Sigdel, Dibakar; Shi, Yu; Wang, Xuan; Shen, Jiaming; Choi, Howard; Caufield, John H.; Wang, Wei; Ping, Peipei; Han, Jiawei (Oct 1, 2018). "Phrase mining of textual data to analyze extracellular matrix protein patterns across cardiovascular disease"*. *American Journal of Physiology. Heart and Circulatory Physiology*. ***315*** *(4): H910–H924. doi:10.1152/ajpheart.00175.2018*. *ISSN 1522-1539*.

UNSTRUCTURED DATA DEFENTIONS

Unstructured data is information, in many different forms, that doesn't hew to conventional data models and thus typically isn't a good fit for a mainstream relational database. Thanks to the emergence of alternative platforms for storing and managing such data, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications.

Traditional structured data, such as the transaction data in financial systems and other business applications, conforms to a rigid format to ensure consistency in processing and analyzing it. Sets of unstructured data, on the other hand, can be maintained in formats that aren't uniform, freeing analytics teams to work with all of the available data without necessarily having to consolidate and standardize it first. That enables more comprehensive analyses than would otherwise be possible.

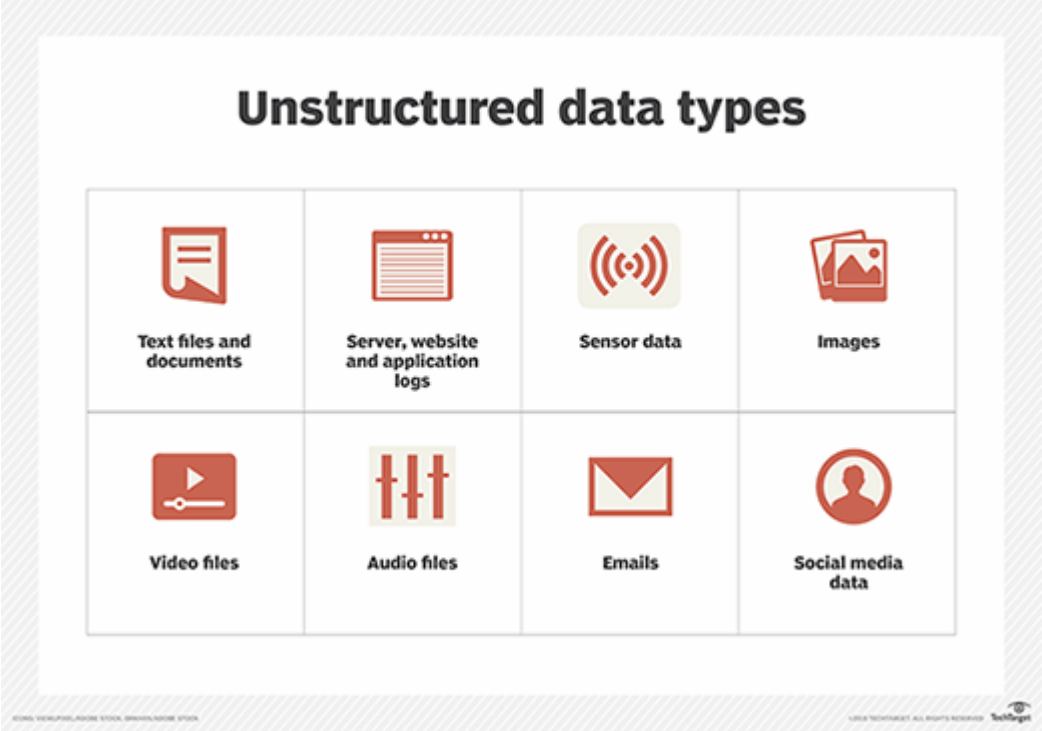
Types of unstructured data

One of the most common types of unstructured data is text. Unstructured text is generated and collected in a wide range

of forms, including Word documents, email messages, PowerPoint presentations, survey responses, transcripts of call center interactions, and posts from blogs and social media sites.

Other types of unstructured data include images, audio and video files. Machine data is another category, one that's growing quickly in many organizations. For example, log files from websites, servers, networks and applications -- particularly mobile ones -- yield a trove of activity and performance data. In addition, companies increasingly capture and analyze data from sensors on manufacturing equipment and other internet of things (IoT) connected devices.

In some cases, such data may be considered to be semi-structured -- for example, if metadata tags are added to provide information and context about the content of the data. The line between unstructured and semi-structured data isn't absolute, though; some data management consultants contend that all data, even the unstructured kind, has some level of structure.



Unstructured data analytics

Because of its nature, unstructured data isn't suited to transaction processing applications, which are the province of structured data. Instead, it's primarily used for BI and analytics. One popular application is customer analytics. Retailers, manufacturers and other companies analyze unstructured data to improve customer relationship management processes and enable more-targeted marketing; they also do sentiment analysis to identify both positive and negative views of products, customer service and corporate entities, as expressed by customers on social networks and in other forums.

Predictive maintenance is an emerging analytics use case for unstructured data. For example, manufacturers can analyze sensor data to try to detect equipment failures before they occur in plant-floor systems or finished products

in the field. Energy pipelines can also be monitored and checked for potential problems using unstructured data collected from IoT sensors.

Analyzing log data from IT systems highlights usage trends, identifies capacity limitations and pinpoints the cause of application errors, system crashes, performance bottlenecks and other issues. Unstructured data analytics also aids regulatory compliance efforts, particularly in helping organizations understand what corporate documents and records contain.

Unstructured data techniques and platforms

Analyst firms report that the vast majority of new data being generated is unstructured. In the past, that type of information often was locked away in soloed document management systems, individual manufacturing devices and the like -- making it what's known as dark data, unavailable for analysis.

But things changed with the development of big data platforms, primarily Hadoop clusters, NoSQL databases and the Amazon Simple Storage Service (S3). They provide the required infrastructure for processing, storing and managing large volumes of unstructured data without the imposition of a common data model and a single database schema, as in relational databases and data warehouses.

A variety of analytics techniques and tools are used to analyze unstructured data in big data environments. Text

analytics tools look for patterns, keywords and sentiment in textual data; at a more advanced level, natural language processing technology is a form of artificial intelligence that seeks to understand meaning and context in text and human speech, increasingly with the aid of deep learning algorithms that use neural networks to analyze data. Other techniques that play roles in unstructured data analytics include data mining, machine learning and predictive analytics.

The phrase *unstructured data* usually refers to information that doesn't reside in a traditional row-column database. As you might expect, it's the opposite of structured data — the data stored in fields in a database.

Examples of Unstructured Data

Unstructured data files often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, web pages and many other kinds of business documents. Note that while these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database.

Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly — often many times faster than structured databases are growing.

Mining Unstructured Data

Many organizations believe that their unstructured data stores include information that could help them make better business decisions. Unfortunately, it's often very difficult to analyze unstructured data. To help with the problem, organizations have turned to a number of different software solutions designed to search unstructured data and extract important information. The primary benefit of these tools is the ability to glean actionable information that can help a business succeeds in a competitive environment.

Because the volume of unstructured data is growing so rapidly, many enterprises also turn to technological solutions to help them better manage and store their unstructured data. These can include hardware or software solutions that enable them to make the most efficient use of their available storage space.

Unstructured Data and Big Data

As mentioned above, unstructured data is the opposite of structured data. Structured data generally resides in a relational database, and as a result, it is sometimes called *relational data*. This type of data can be easily mapped into pre-designed fields. For example, a database designer may set up fields for phone numbers, zip codes and credit card numbers that accept a certain number of

digits. Structured data has been or can be placed in fields like these. By contrast, unstructured data is not relational and doesn't fit into these sorts of pre-defined data models.

Semi-Structured Data

In addition to structured and unstructured data, there's also a third category: semi-structured data. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. Examples of semi-structured data might include XML documents and NoSQL databases.

The term big data is closely associated with unstructured data. Big data refers to extremely large datasets that are difficult to analyze with traditional tools. Big data can include both structured and unstructured data, but IDC estimates that 90 percent of big data is unstructured data. Many of the tools designed to analyze big data can handle unstructured data.

Unstructured Data Management

Organizations use of variety of different software tools to help them organize and manage unstructured data. These can include the following:

Big data tools

Software like Hadoop can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly.

Business intelligence software

Also known as BI, business intelligence is a broad category of analytics, data mining, dashboards and reporting tools that help companies make sense of their structured and unstructured data for the purpose of making better business decisions.

Data integration tools

These tools combine data from disparate sources so that they can be viewed or analyzed from a single application. They sometimes include the capability to unify structured and unstructured data.

Document management systems

Also called enterprise content management systems, a DMS can track, store and share unstructured data that is saved in the form of document files.

Information management solutions

This type of software tracks structured and unstructured enterprise data throughout its lifecycle.

Search and indexing tools

These tools retrieve information from unstructured data files such as documents, Web pages and photos.

Unstructured Data Technology

A group called the Organization for the Advancement of Structured Information Standards (OASIS) has published

the Unstructured Information Management Architecture (UIMA) standard. The UIMA "defines platform-independent data representations and interfaces for software components or services called analytics, which analyze unstructured information and assign semantics to regions of that unstructured information."

Many industry watchers say that Hadoop has become the de facto industry standard for managing Big Data. This open source project is managed by the Apache Software Foundation.

Definition - What does *Unstructured Data* mean?

Unstructured data represents any data that does not have a recognizable structure. It is unorganized and raw and can be non-textual or textual. For example, email is a fine illustration of unstructured textual data. It includes time, date, recipient and sender details and subject, etc., but an email body remains unstructured. Unstructured data also may be identified as loosely structured data, wherein the data sources include a structure, but not all data in a data set follow the same structure.

In customer-centered businesses, the data found in an unstructured form may be examined to enhance relationship marketing and customer relationship management (CRM). As social media apps, such as Facebook and Twitter, go mainstream, unstructured data development is likely to outrun the progress of structured data.

Techopedia explains *Unstructured Data*

Unstructured data refers to data that follows a form that is less ordered than items like spreadsheet pages, database tables or other linear or ordered data sets. In fact, the term "data set" is helpful because it is associated with data that is in neat, accessible arrays, without any extra content, and that is linked or tagged in a specific structure.

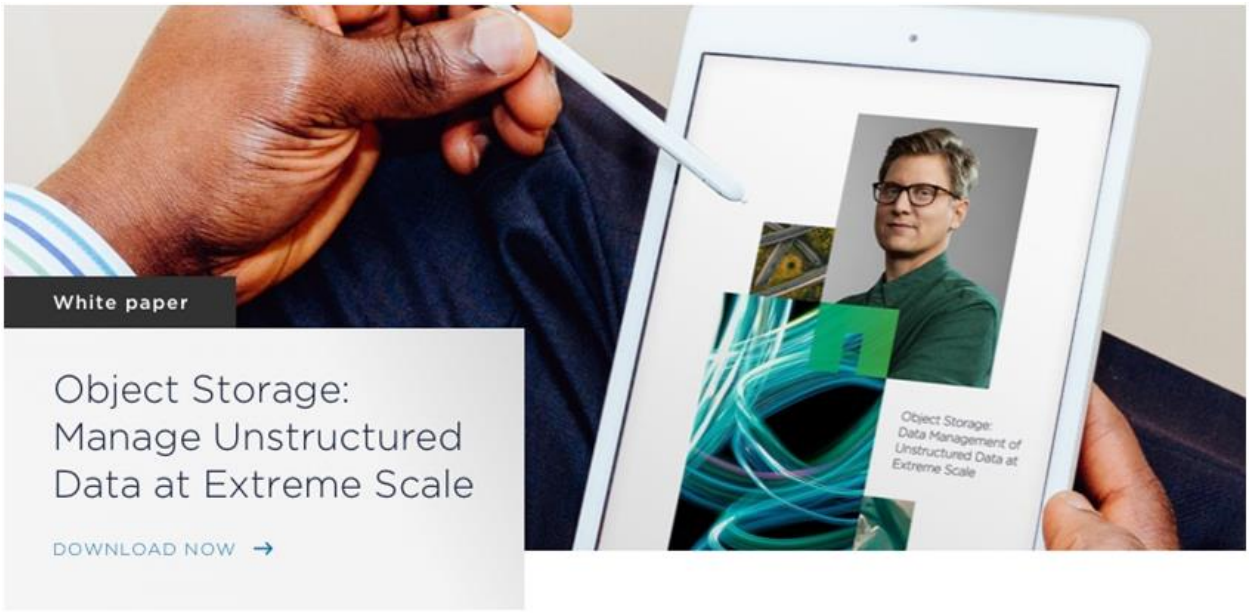
Other instances of unstructured textual data include Word documents, PowerPoint presentations, instant messages, collaboration software, documents, books, social media posts and medical records. Non-textual unstructured data is generally created in media, such as MP3 audio files, JPEG images and Flash video files, etc.

Unstructured data usually does not include a predefined data model, and it may not match well with relational tables. Unstructured data is usually text heavy. However, it may include numbers and dates, as well as facts. This leads to ambiguities that are difficult to identify using conventional software programs.

The storage of huge volumes of unstructured data generated within an enterprise, if poorly managed, may lead to higher expenses. Data in hard copy documents or in an electronic format must be scanned in order for a search application to parse out ideas, depending on words used in certain contexts. This is known as enterprise or semantic search.

What Is Unstructured Data?

In the modern world of big data, unstructured data is the most abundant. It's so prolific because unstructured data could be anything: media, imaging, audio, sensor data, text data, and much more. *Unstructured* simply means that it is datasets (typical large collections of files) that aren't stored in a structured database format. Unstructured data has an internal structure, but it's not predefined through data models. It might be human generated, or machine generated in a textual or a non-textual format.



Unstructured Data vs. Structured Data

Unstructured data can be thought of as data that's not actively managed in a transactional system; for example, data that doesn't live in a relational database management system (RDBMS). Structured data can be thought of as records (or transactions) in a database environment; for example, rows in a table of a SQL database.

There is no preference as to whether data is structured or unstructured. Both have tools that allow users to access information. Unstructured data just happens to be in greater abundance than structured data is.

Examples of unstructured data are:

- **Rich media.** Media and entertainment data, surveillance data, geo-spatial data, audio, weather data
- **Document collections.** Invoices, records, emails, productivity applications
- **Internet of Things (IoT).** Sensor data, ticker data
- **Analytics.** Machine learning, artificial intelligence (AI)

Until the advent of object-based storage, most, if not all, of this unstructured data was stored in file-based systems.

What Challenges Does Working with Unstructured Data Present?

The way to think about how to deal with the challenges of unstructured data is to ask: What do enterprises face with traditional approaches to managing unstructured data?

Scale

It's common in many enterprises to encounter unstructured datasets at the scale of tens or hundreds of billions of items. These items, objects, or files can be anything from a few bytes (for example, a temperature reading from a production-line instrument) to terabytes in size (for example, a full-length 8K resolution motion picture). Managing this scale with traditional file approaches rapidly moves from difficult to impossible as more and more resources are required just to maintain a "balance" of servers, file systems, arrays, and so on.

Collaboration

Increasingly, these massive unstructured datasets deliver value as they are shared (for example, researchers at multiple hospitals who share a common massive bank of genomic sequences). With traditional approaches, the ability to share massive sets of unstructured data across geographies, corporate entities, and so on, has required extremely expensive replication and governance.

Overcoming These Challenges by Using Object Storage

Today's object storage solutions meet the challenges of scale and collaboration by delivering a geo-distributed active namespace. This namespace enables a user at any location to retrieve an object or a file from any location with a simple GET command (without having to specify a data center, server, file system, or director). Similarly, PUT commands enable the ingest of data so that all locations can easily have access.

The simplicity and scalability of a single global namespace combined with a simple stateless data management protocol (for example, Amazon S3 and Swift) help organizations deliver a scalable and collaborative environment across geography, organization, and application boundaries.

NetApp and Object Storage

You can store and manage unstructured data at scale by using NetApp® StorageGRID® technology for secure, durable object storage for private and public clouds. With StorageGRID, you can build a massive (multilocation) single namespace, and you can also integrate a unique information lifecycle policy into that data. With the StorageGRID integrated policy engine, you can be confident that your data is available:

- In the right geographic location
- At the right level of performance
- At the right level of durability and protection
- At the right time and changing over time automatically as business needs evolve

Unstructured data: A cheat sheet

Unstructured data is approximately 80% of the data that organizations process daily. This primer covers what unstructured data is, why it enriches business data, and how it speeds up decision making.

By 2025, IDG projects that there will be 163 zettabytes of data in the world, and estimates indicate that 80% of this data is unstructured.

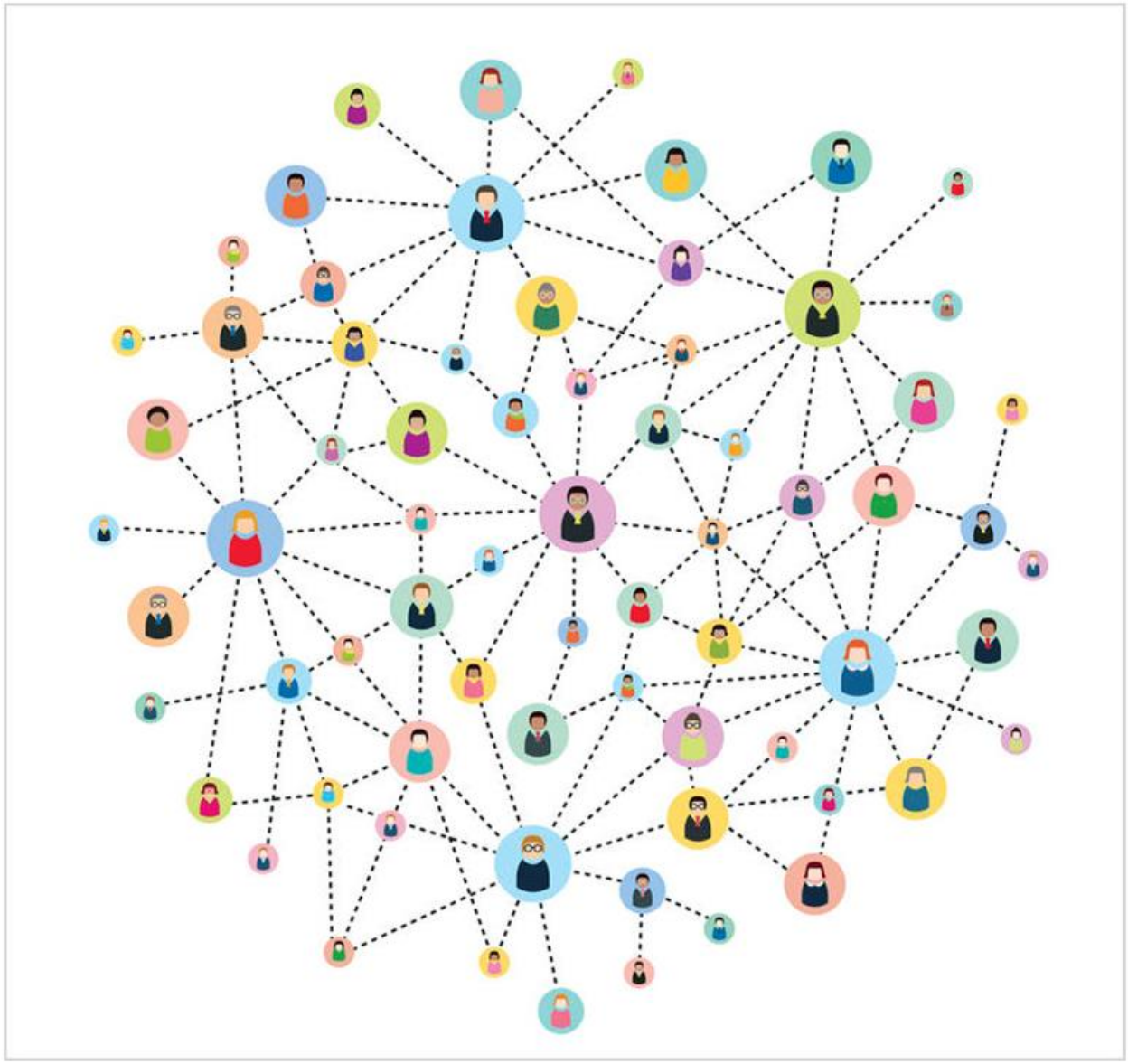
With structured data, data fields are aligned side-by-side in fixed record lengths, with specific data fields appearing at static locations within each record. Unstructured data does not contain a set record format--it can come in any shape or form. Unstructured data comes from documents, social media feeds, digital pictures and videos, audio transmissions, sensors used to gather climate information, and unstructured content from the web.

Learn more about unstructured data by reading this smart person's guide. We will update this resource periodically with the latest information and tips about unstructured data.

Executive summary

- **What is unstructured data?** Unstructured data is data that aren't stored in a fixed record length format. Examples include documents, social media feeds, and digital pictures and videos.
- **Why does unstructured data matter?** Since 80% of data that organizations see and process daily is unstructured, businesses must adapt to handle the increasing stores of unstructured data.
- **Who does unstructured data affect?** Internally, almost every corporate department uses unstructured data in some form; externally, unstructured data is used to monitor and report on movements of shipments and/or assets with sensors and more.
- **When will businesses use unstructured data?** Unstructured data is used in every company and organization.
- **How can companies use unstructured data?** The preparation and processing of unstructured data, and the ability to append it to systems of record or store it for future use, is available in an on-premise system and in the cloud.

What is unstructured data?



Unstructured data is any data that aren't stored in a fixed record length format, which is known as transactional data. Examples of unstructured data include:

- A paper-based vendor invoice that comes into your accounting department;
- A product photo that gets associated with an order;

- A bar code that is assigned to an item in your warehouse;
- A doctor treating a patient needs more than a patient's written records--she might also require x-rays, MRIs, and other types of imagery;
- A police officer wants to be able to call up a photo of a possible suspect, as well as the suspect's driving record; and
- A surveyor mapping a section of land needs more than his own measurements; he combines this data with satellite and topographical imagery that enhances his records in your accounting, inventory management, or order systems are not unstructured data because those records are all structured in the same, uniform way. Every record consists of a series of contiguous data fields, and one of these fields is the access point, or key, into the records (e.g., the key to an order record might be the order number).

To process unstructured data, your systems and databases need be able to read this data by looking at a key or a reference point; for example, the key to a stored photo of Jim Smith would likely be Jim Smith. After that, the system must access the entire data object (i.e., the photo of Jim Smith).

The difference between an unstructured object of Jim Smith and a structured record of him is that the unstructured data object would be Jim's photo, which would require quite a bit of storage; the structured record of Jim Smith would be information about him, such as his address or his phone

number. There is much more data in a full photo of Jim than in a small, fixed string of data elements like an address, phone number, etc. that you would find in a structured data record; for this reason, unstructured data, with the large objects that are assigned to its keys, requires more processing and more storage.

Why does unstructured data matter?

More about Big Data

Because 80% of the data in companies is unstructured, organizations need to understand the types of unstructured data they are accumulating and the best ways to process and store this data for business advantages. Without data management strategies and guidance in these areas, companies run the risks of not capitalizing on unstructured data, failing to keep up with competitors, or storing more unstructured data than they really need, thereby running up data center costs.

In a majority of cases, unstructured data is ultimately related back to the company's structured data records. As an example, every x-ray or MRI image for a patient is related back to the patient's record in the hospital's record system. The patient record in the record system is enriched with unstructured data that is linked to it, and the doctor gets a more complete picture of the patient.

This is the value of unstructured data: It enriches corporate data and enables leaders to work smarter.

Who does unstructured data affect?

Unstructured data can affect everyone at the company, from the entry-level staffer to the CEO.

Internally, almost every corporate department uses unstructured data in some form--from engineering with its rastor drawings to marketing with its social media engagements and photo imagery, to financial and office operations with scanned documents.

Externally, unstructured data is used to monitor and report on movements of shipments and/or assets with sensors, to monitor school campuses with security cameras, and to exchange videos, photos, images, audio transmissions, etc. with suppliers and other business partners.

All of these unstructured data users need access, policies, and guidelines for using the data.

The organizations that do the best job of using unstructured data assess their business and determine where they need unstructured data most. Then they find ways to blend this data with systems of record transactional data so that employees have more complete information at their fingertips.

This unstructured data could be strategic, such as a map that integrates sales and demographics information with warehouse locations so a company can plan its next facility moves.

On the operational side, unstructured data can be appended to manufacturing rework orders to show how and where parts failed, so problems can be anticipated and eliminated in the future.

In both cases, leaders have access to more information than ever before, which speeds up decision making.

When will businesses use unstructured data?

It was reported in 2016 that the Economist Intelligence Unit interviewed 476 executives around the world, and according to Forbes contributor Bernard Marr, the report concluded that: "Big Data analysis, or the mining of extremely large data sets to identify trends and patterns, is fast becoming standard business practice. Global technology infrastructure, too, has matured to an extent that reliability, speed, and security are all typically robust enough to support the seamless flow of massive volumes of data, and consequently encourage adoption."
The same report stated that 58% of companies in the US, 56% of companies in Europe, and 60% of companies in Asia were deriving high business value from their use of data.

How can companies use unstructured data?

Every business has unstructured data--the key of knowing how to use and process it to the smartest business advantage, and which unstructured data to store and what to discard. It is very expensive to store and maintain unstructured data because of its size, the processing power it requires, and the fact that not all of it is useful.

Organizations use many types of unstructured data at face value, such as photographs, documents, audio and video recordings, and web content. The next stage is figuring out how to get at the data so that all of the information it contains is fully used. In the case of a photograph, the photo can be linked with spatial events through GPS technology that links it to a specific location; or, the photo can be enriched by being linked to contextual and associative data, such as who is in the photo, the year the photo was taken, and so on.

The preparation and processing of unstructured data, and the ability to append it to systems of record, or to store it for future use, is available in an on-premise system and in the cloud. In both cases, data analysts must groom the unstructured data so it can work hand-in-hand with other types of unstructured and structured data. Adequate storage and processing must also be provisioned.

The ways that companies use unstructured data varies widely. Larger enterprises often have their IT staffs and experts prepare and process unstructured data in house. Small, midsize, and even some large businesses use cloud-based unstructured data preparation and processing. Cloud-based options such as Amazon AWS, Microsoft Azure, and IBM Cloud are services that help make big data affordable and attainable to companies of any size.

Some Approaches to Contextualization

The notion that finding context in Non-repetitive unstructured data is a challenge is not a new idea. Indeed, people have been attempting to contextualize text for a long time. The earliest attempt to trying to contextualize text is a technology called "NLP." NLP stands for natural language processing (or sometimes "natural language programming.")

NLP has been around a long time and has met with modest success. There are several inherent limitations to NLP. The first limitation is that NLP makes the assumption that context of text can be derived from text itself. The problem is that only a small amount of context comes from text itself. In the case of the two gentlemen standing around and saying—"She's hot"—the vast majority of the context comes from external sources, not textual sources. Is the lady young and attractive? Is it Houston, Texas, in the summertime? Is the conversation taking place in a hospital? All of these circumstances that provide

context are external to the words that are being spoken.

The second limitation of NLP is that NLP does not account for emphasis. Suppose the words are spoken—“I love you.” How are these words to be interpreted?

If you say “I love you” where the emphasis is on “I,” the meaning is that it is me and not someone else who loves you. If the emphasis is on the word “love,” the meaning is that the emotion I feel is strong, one of love. I don’t like you—I actually love you. If the emphasis is on the word “you,” the meaning is that it is you and not someone else that I love.

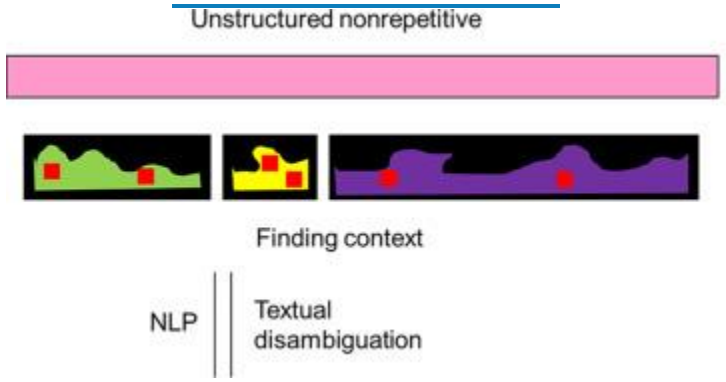
So, the same words can have very different meaning based on the way the words are said.

But there is a very different reason why NLP has had a hard time showing concrete results. That reason is that NLP—in order to be implemented effectively—must understand the logic behind words. The problem is that the English language has evolved over many years and many circumstances, and at the end of the day, the logic behind the English language is very complex. Trying to map out the logic of the English language is very difficult to do. It is tortuous.

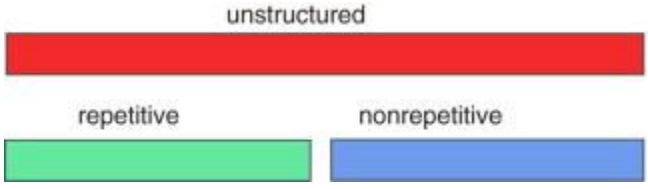
For these reasons (and probably more), NLP processing has met with modest success.

A much more practical approach is that of textual disambiguation

Down here shows the two approaches toward contextualization of text.



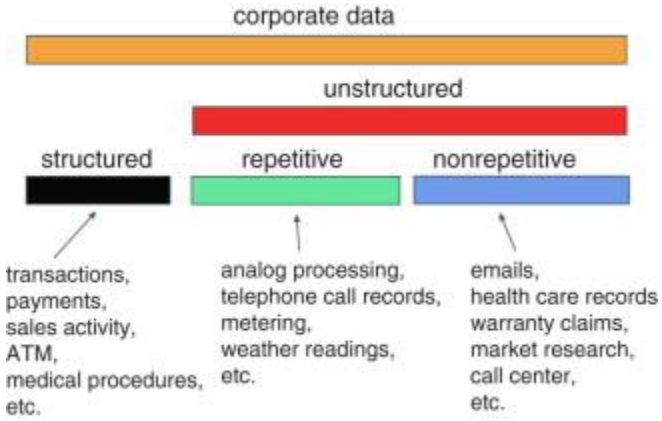
Unstructured data can further be divided into two basic forms of data – repetitive unstructured data and no repetitive unstructured data. As is the case with the division of corporate data, there are many ways to subdivide unstructured data. The method shown here is but one of many ways to subdivide unstructured data. This simple subdivision of unstructured data is shown in down here.



Repetitive unstructured data is data that occurs many times, often in the same structure and even in the exact same embodiment. Typically, repetitive data occurs many, many times. The structure of repetitive data looks exactly the same or substantially the same as the previous record. There is no massive and elaborate infrastructure managing the content of repetitive unstructured data.

Non-repetitive unstructured data is data where the records are substantially different from each other. In general each Non-repetitive record is markedly different from each other record.

The division of data types in the corporation has many different embodiments. Consider the data as shown down here.



Structured data is typically found as a by-product of transactions. Every time a sale is made, every time a bank account encounters a withdrawal, every time someone transacts an ATM activity, and every time a bill is sent a record of the transaction is made. The record of the transaction ends up as a structured record.

Unstructured repetitive data is quite different. Unstructured repetitive records are typically records of machine interactions, such as the analog verification of product coming off a manufacturing process or the metering of energy usage by a consumer. Consider metering. There is great repetition of records in both form and substance that are created when looking at metered readings.

Unstructured Non-repetitive information is fundamentally different than unstructured repetitive records. With unstructured Non-repetitive records there is little or no repetition of either form or content from one record to the next. Some examples of unstructured Non-repetitive information include email, call center conversations, and market research. When you look at one email, the odds are very good that the next email in the database will be different than the previous email. The same is true for call center

information, warranty claims, market research, and so forth.

Data Stream Disposing for Real-Time Data Operation.

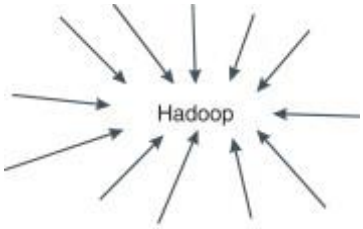
Unstructured data such as video data can not be stored into a structured database system for analysis purpose. And data mining on data stream form different data sources with non-persisted association is a new but important issue. There are several different directions to process data stream with some dynamic methods, for example, to retrieve features from continuous data stream so as to build data association, or to process the whole body of a fragment of data stream by function transformation. In [81], WoT infrastructure should focus on real-time interaction in the future research. Therefore, a WoT micro-benchmark is designed to combine cloud computing, service decomposition and multi-threading programming. And the benchmark is evaluated over a real WoT system.

The Streaming Linked Data (SLD) framework [82] provides a pluggable system for analysis of RDF streams. By means of a set of visualization widgets, data streams could be collected

and analysed based on semantic techniques. And the Streaming Linked Data Format can be used in distributed environments with flexibility.
Data Stream mining involves uncertain reasoning based on partition data and utilizes intermediate result for high efficiency. When unstructured and semi-structured data are also involved in the processing process, there are lots of researches and technical problems left to do.

Repetitive Unstructured Data

The repetitive unstructured data is said to be “Hadoop centric.” Being Hadoop centric means that processing of repetitive unstructured data revolves around processing and managing the Hadoop environment. The centricity of the repetitive unstructured data is seen down here.



Hadoop centric

The center of the Hadoop environment naturally enough is Hadoop. Hadoop is the technology by which very large amounts of data can be managed. Hadoop is at the center of what is known as Big Data. Hadoop is the primary storage mechanism for Big Data. The essential characteristics of Hadoop are:

–

It is capable of managing very large volumes of data.

–

It manages data on less expensive storage.

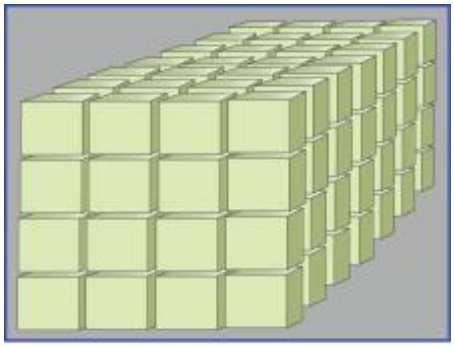
–

It manages data by the “Roman census” method.

–

It stores data in an unstructured manner.

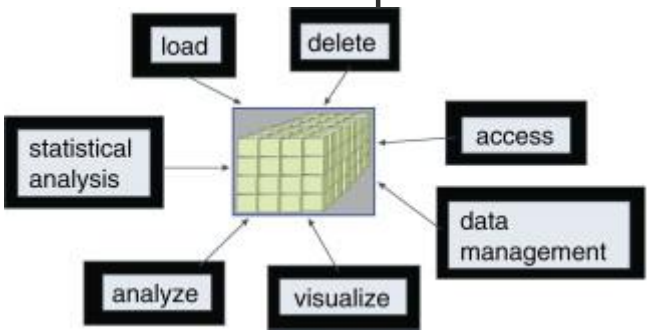
Because of these operating characteristics of Hadoop, very large volumes of data can be managed. Hadoop is capable of managing volumes of data significantly larger than standard relational database management systems. The Big Data technology of Hadoop is depicted down here.



Hadoop

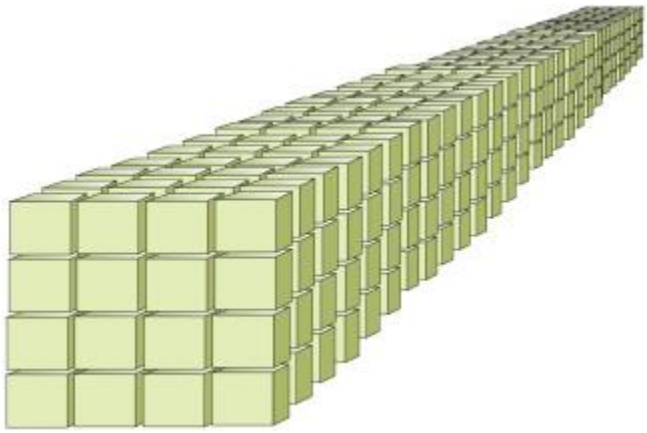
But Hadoop is a raw technology. In order to be useful, Hadoop requires its own unique infrastructure.

The technologies that surround Hadoop serve to manage the data and to access and analyze the data found in Hadoop. The infrastructure services that surround Hadoop are seen down here.



The services that surround Hadoop are familiar to anyone that has ever used a standard DBMS. The difference is that in a standard DBMS the services are found in the DBMS itself, while in Hadoop many of the services have to be done externally. A second major difference is that throughout the Hadoop environment there is the need to service huge volumes of data. The developer in the Hadoop environment must be prepared to manage and handle

extremely large volumes of data. This means that many infrastructure tasks can be handled only in the Hadoop environment itself.
Indeed, the Hadoop environment is permeated by the need to be able to handle extraordinarily large amounts of data. The need to handle nearly unlimited amounts of data is seen down.



There is then an emphasis on doing the normal tasks of data management in the Hadoop environment where the process must be able to handle very large amounts of data.

Some Approaches to Contextualization

The notion that finding context in Non-repetitive unstructured data is a challenge is not a new idea. Indeed people have been attempting to contextualize text for a long time. The earliest attempt

to trying to contextualize text is a technology called "natural language processing" (sometimes called "natural language programming"; NLP)
NLP has been around a long time and has met with modest success. However, there are several inherent limitations to NLP. The first limitation is that NLP makes the assumption that context of text can be derived from text itself. The problem is that only a small amount of context comes from text itself. In the case of the preceding example with the two gentlemen, the vast majority of the context comes from external sources, not textual sources. Is the woman young and attractive? Is it Houston, Texas, in the summertime? Is the conversation taking place in a hospital? All of these circumstances that provide context are external to the words that are being spoken.

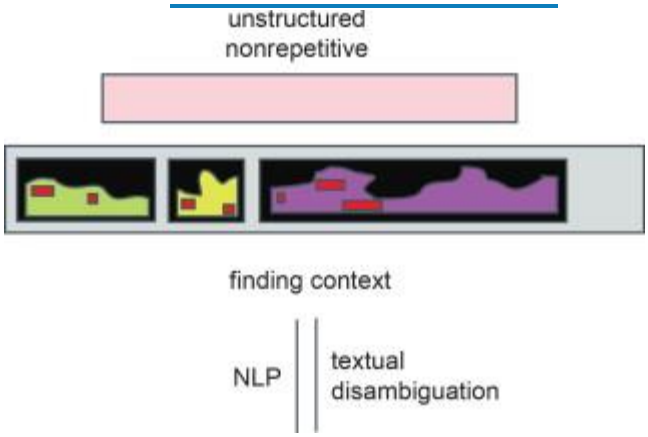
The second limitation of NLP is that NLP does not account for emphasis. Suppose the words, "I love you" are spoken. How are these words to be interpreted?

If you say "I love you" where the emphasis is on "I," the meaning is that it is me and not someone else who loves you. If the emphasis is on the word "love," the meaning is that the emotion I feel is strong, one of love. I don't merely like you, I actually love you. If the

emphasis is on the word “you,” the meaning is that it is you, and not someone else that I love.
So the same words can have very different meaning based on the way the words are said.

But there is a very different reason why NLP has had a hard time showing concrete results. That reason is that NLP, in order to be implemented effectively, must understand the logic behind words. The problem is that the English language has evolved over many years and many circumstances, and at the end of the day, the logic behind the English language is very complex. Trying to map out the logic of the English language is not only difficult, but tortuous.

For these reasons (and probably more) NLP processing has met with modest success. A much more practical approach is that of textual disambiguation. Here it shows the two approaches toward contextualization of text.



Unstructured

Unstructured data are any data that are not associated with a discrete data field. In other words, unstructured data are data that do not reside in a spreadsheet file or in a database format. Examples of unstructured data include files that reside on a file share—often text or binary files like Word, PowerPoint, audio files, video files, image files, and more.

The problem with unstructured data is that they are hard to sort, manage, and organize. If you think of the files stored on servers or network storage, you know that there may be thousands of duplicate copies (or almost duplicate copies) of data in many different formats. It's not uncommon for medium-to-large organizations to have terabytes of unstructured data that they need to manage, backup, and potentially recover. It's both a problem for the IT department and a significant, if not always visible, cost to the organization in terms of resources (storage space, backup space, backup time, staff resources) to manage so much unstructured data.

Conclusion

Unstructured data, of which text data is a major portion, appears to be doubling in volume every three years (Mayer-Schonberger, 2013). The ability to automatically process and mine information from such digital data will become an important skill in the future. These techniques can be used to classify and predict just as the other techniques throughout the book, except one is now working on text documents and even voice recordings that have been transcribed to text.

Chapter 11:

STRUCTURED DATA

Structured Data Definitions:

Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.

A data structure is a kind of repository that organizes information for that purpose. In a database, for example, each field is discrete and its information can be retrieved either separately or along with data from other fields, in a variety of combinations. The power of the database is its ability to make data comprehensive, so that it yields useful information. A database query language, such as SQL (standard query language), allows a database administrator to interact with the database.

Structured data contrasts with unstructured and semi-structured data. The three can be considered to exist on a continuum, with unstructured data being the least formatted and structured data being the most formatted. Data is increasingly amenable to processing as it is increasingly structured.

STRUCTURED DAT

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.

Characteristics of Structured Data

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).
Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. Anything that couldn't fit into a tightly organized structure would have to be stored on paper in a filing cabinet.

Managing Structured Data

Structured data is often managed using Structured Query Language (SQL) – a programming language created for managing and querying data in relational database management systems. Originally developed by IBM in the early 1970s and later developed commercially by Relational Software, Inc. (now Oracle Corporation). Structured data was a huge improvement over strictly paper-based unstructured systems, but life doesn't always fit into neat little boxes. As a result, the structured data always had to be supplemented by paper or microfilm storage. As technology performance has continued to improve, and prices have dropped, it was possible to bring into computing systems unstructured and semi-structured data.

Unstructured and Semi-Structured Data

Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and

graphic images, videos, streaming instrument data, web pages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text. Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments. Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics. XML and other markup languages are often used to manage semi-structured data.

Structured Data Technology Standards

SQL has been a standard of the American National Standards Institute since 1986. It is managed by InterNational Committee for Information Technology Standards (INCITS) Technical Committee DM 32 – Data Management and Interchange. The committee has two task groups, one for databases and the other for metadata. HP, CA, IBM, Microsoft, Oracle, Sybase (SAP) and Teradata all participate, as well as several federal government agencies. Both of the committee project documents have links to further information on each project. SQL became an International Organization for Standards (ISO) standard in 1987. The published standards are available for purchase from the ANSI eStandards Store, under the INCITS/ISO/IEC 9075 classification.

Understand how structured data works

Google Search works hard to understand the content of a page. You can help us by providing explicit clues about the meaning of a page to Google by including structured data on the page. Structured data is a standardized format for providing information about a page and classifying the page content; for example, on a recipe page, what are the ingredients, the cooking time and temperature, the calories, and so on.

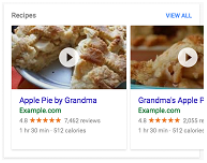
Google uses structured data that it finds on the web to understand the content of the page, as well as to gather information about the web and the world in general. For example, here is a JSON-LD structured data snippet that might appear on the contact page of the Unlimited Ball Bearings Corporation, describing their contact information:

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Organization",
  "url": "http://www.example.com",
  "name": "Unlimited Ball Bearings Corp.",
  "contactPoint": {
    "@type": "ContactPoint",
    "telephone": "+1-401-555-1212",
    "contactType": "Customer service"
```

```
   }
}
</script>
```

Google Search also uses structured data to enable special search result features and enhancements. For example, a recipe page with valid structured data is eligible to appear in a graphical search result, as shown here:

Search result feature	Equivalent structured data (trimmed for brevity)

	<html>
	<head>
	<title>Grandma's Holiday Apple Pie</title>
	<script type="application/ld+json">
	{
	"@context": "https://schema.org/",
	"@type": "Recipe",
	"name": "Grandma's Holiday Apple Pie",
	"author": "Elaine Smith",
	"image": "http://images.edge-generalmills.com/56459281-6fe6-4d9d-984f-385c9488d824.jpg",
	"description": "A classic apple pie.",
	"aggregateRating": {
	"@type": "AggregateRating",
	"ratingValue": "4",
	"reviewCount": "276",
	"bestRating": "5",
	"worstRating": "1"
	},
	"prepTime": "PT30M",
	"totalTime": "PT1H",
	"recipeYield": "8",
	"nutrition": {
	"@type": "NutritionInformation",
	"servingSize": "1 medium slice",
	"calories": "230 calories",

Because the structured data labels each individual element of the recipe, users can search for your recipe by ingredient, calorie count, cook time, and so on.

Structured data is coded using in-page markup on the page that the information applies to. The structured data on the page should describe the content of that page. You should not create blank or empty pages just to hold structured data; nor should you add structured data about information

that is not visible to the user, even if the information is accurate. For more technical and quality guidelines, see the Structured data general guidelines.

Structured data format

This documentation describes which properties are required, recommended, or optional for structured data with special meaning to Google Search. Most Search structured data uses schema.org vocabulary, but you should rely on the documentation on developers.google.com as definitive for Google Search behavior, rather than the schema.org documentation. Attributes or objects not described here are not required by Google Search, even if marked as required by schema.org.

Be sure to test your structured data using the Structured Data Testing tool during development, and the Rich result status reports after deployment, to monitor the health of your pages, which might break after deployment due to tinplating or serving issues.

You must include all the required properties for an object to be eligible for appearance in Google Search with enhanced display. In general, defining more recommended features can make it more likely that your information can appear in Search results with enhanced display. **However**, it is more important to supply fewer but complete and accurate recommended properties rather than trying to provide every possible recommended property with less complete, badly-formed, or inaccurate data.

In addition to the properties and objects documented here, Google can make general use of the sameAs property and other schema.org structured data. Some of these elements may be used to enable future Search features, if they are deemed useful.

Google Search supports structured data in the following formats, unless documented otherwise:

Format	Description and Placement
JSON-LD* (Recommended)	JavaScript notation embedded in a <script> tag in the page head or body. The markup is not interleaved with the user-visible text, which makes nested data items easier to express, such as the Country of a PostalAddress of a MusicVenue of an Event. Also, Google can read JSON-LD data when it is dynamically injected into the page's contents, such as by JavaScript code or embedded widgets in your content management system.
Microdata	An open-community HTML specification used to nest structured data within HTML content. Like RDFa, it uses HTML tag attributes to name the properties you want to expose as structured data. It is typically used in the page body, but can be used in the head.
RDFa	An HTML5 extension that supports linked data by introducing HTML tag attributes that correspond to the user-visible content that you want to describe for search engines. RDFa is commonly used in both the head and body sections of the HTML page.

Google recommends using JSON-LD for structured data whenever possible.

Structured data guidelines

Be sure to follow the general structured data guidelines, as well as any guidelines specific to your structured data type; otherwise your structured data might be ineligible for rich result display in Google Search.

See our recommended process for creating new structured data pages.

Building, testing, and releasing your structured data

Google recommends becoming a verified site owner of your site in Search Console. Search Console provides many useful debugging and monitoring reports and tools to test rendering, availability, and structured data indexing.

Create new pages

Try our new Structured Data Codelab. This codelab walks you through adding several types of structured data to a simple HTML site, including where to place structured data on a site and how to validate it.

Here is our recommended process for creating new structured data pages:

1. Follow the structured data guidelines for your page and feature type. For example, if your page is a recipe, be sure to follow the guidelines for Recipes.
2. Validate your code using the Structured Data Testing Tool. Paste your code into the tool and click the validate button.
   Tip: For some result types, the Structured Data Testing Tool lets you generate a sample view of a result from your data. For those types, after you click "validate" on the tool, if the code is valid you will see a "Preview" button, which will generate a sample search result display based on your data.
3. Deploy a few pages using your structured data and use the URL Inspection tool to test how Google sees the page. Be sure that your page is accessible to Google and not blocked by robots, noindex, or login requirements. If the page looks okay, use that tool to request an index of the page. Read here to find out how to test a page hosted on a local server or behind a firewall.

4. After Google has indexed the page, look for errors and properly processed data using the applicable Rich result status reports. Ideally you should see an increase of rich results or structured data types and no increase in error rates.

- If you see your structured data being crawled without errors, update your sitemap to have Google crawl those pages regularly.
- If you find problems in your structured data, fix it, retest on URL Inspection, and request another index until everything looks good in the reports. Then deploy the rest of your pages.

5. Monitor your error rates periodically using the applicable Rich result status reports, especially after releasing new templates or updating your code.

Fix existing pages

1. Sign up as a verified site owner of your site in Search Console. If Search Console notices any dramatic changes in error rates, you might be notified by email about the change.
2. Periodically monitor your pages using the applicable Rich result status reports. Here are some common reasons for errors:

- If you see an **increase in errors**, perhaps you rolled out a new template that doesn't work, or your site interacts with the existing template in a new and bad way.
- If you see a total **decrease in structured data**, check whether the sudden decrease in indexed structured data is matched by an increase in errors. If not, perhaps you are no longer embedding structured data in your pages, or the pages are inaccessible to Google (perhaps pages are blocked by robots.txt or noindex). Test page availability using the URL Inspection tool.

  You can retest a specific page using the Structured Data Testing Tool.

3. Fix your code and test it using the Structured Data Testing Tool. Read here to find out how to test a page hosted on a local server or behind a firewall.
4. Roll out your fix and wait for Google to recrawl your page. Remember that this can take a few days.

Allow time for re-crawling and re-indexing! Remember that it takes several days after publishing a page for Google to find and crawl it.

Fix pages with manual actions

If you have a structured data manual action against your page, the structured data on the page will be ignored (although the page can still appear in Google Search results).

Verified site owners in Search Console receive an email message when their site is affected by a manual action. You can open the Manual Actions page in Search Console to see any manual actions affecting your site.

If you are not a verified site owner in Search Console, you will not be able to check if your site has manual actions.

To diagnose and fix your code:

1. **Open** the Manual Actions page for your site.
2. **Expand** the **Site-wide matches** and **Partial matches** sections, if they are collapsed, and look for any structured data issues.
3. **Determine which pages and which rich result types are affected.**

- For partial matches (actions that affect only part of your site), you should see the URLs of some example pages that are non-compliant.
- For issues that affect the entire site, you won't see URLs for example pages, so you will have to do some detective work to determine the cause

of the failure. Look into recent changes in your templates as a clue to what changed recently and might be triggering the action.

4. **Read all the guidelines** (*very carefully!*) for each rich result type affected by the action to find which guideline you are not compliant with:

- General technical guidelines for all result types
- General quality guidelines for all result types
- Guidelines for your specific rich result type (for example, Event, Review snippet, or Job Posting)

  Take a look at *Common structured data errors* below. The problem can be caused by either spammy content or spammy markup usage. However, the issue may not be a syntax issue, and so a markup validator won't be able to identify these issues.

5. Fix the issue and republish your pages.
6. Submit a reconsideration request in the Manual Actions report for each specific manual action. *Do not submit a reconsideration request unless you think the issue is fixed*; filing multiple reconsideration requests on unfixed content can negatively affect the page review time.
7. **Wait for the results** of the request, which can take **several days** (more than 3). You will get either a confirmation that you fixed the issue, or a notice that the problem was not fixed.

Common structured data errors

Here are some common causes of structured data manual actions.

Affected types	Description
All items	The information in the structured data is not shown in the user-visible text. For example, review stars are in the structured data but not on the page. This would cause the stars to appear only in Search results, but not on the page. This often occurs with Products or Reviews. Relevant guidelines: The marked-up content is hidden from the user. Markup should not be used to hide content not visible to users in any form, since it might create a misleading or deceptive search experience. For example, if the JSON-LD markup describes a performer, the HTML body should describe that same performer. Make sure the reviews and ratings you mark up are readily available to users from the marked-up page.
Event	A page is using event markup but there is no visible event content on the page. The visible text or description in markup is more concerned with promoting/selling the event than describing what it is. Relevant guidelines: Avoid marking non-events as events: Don't promote non-event products or services such as "Trip package: San Diego/LA, 7 nights" as events. Don't add short-term discounts or purchase opportunities, such as: "Concert — buy your tickets now," or "Concert - 50% off until Saturday".
Job	Job markup used, but no job related content on the page No way to apply for the job

Everything You Want to Know About Structured Data for SEO

What is the definition of structured data? You've heard it mentioned when reading about search engine optimization (SEO) and have undoubtedly seen it while browsing search results, but what is it and how does it work on a site like yours?

This beginner's guide to structured data (or schema markup) answers those questions and more.

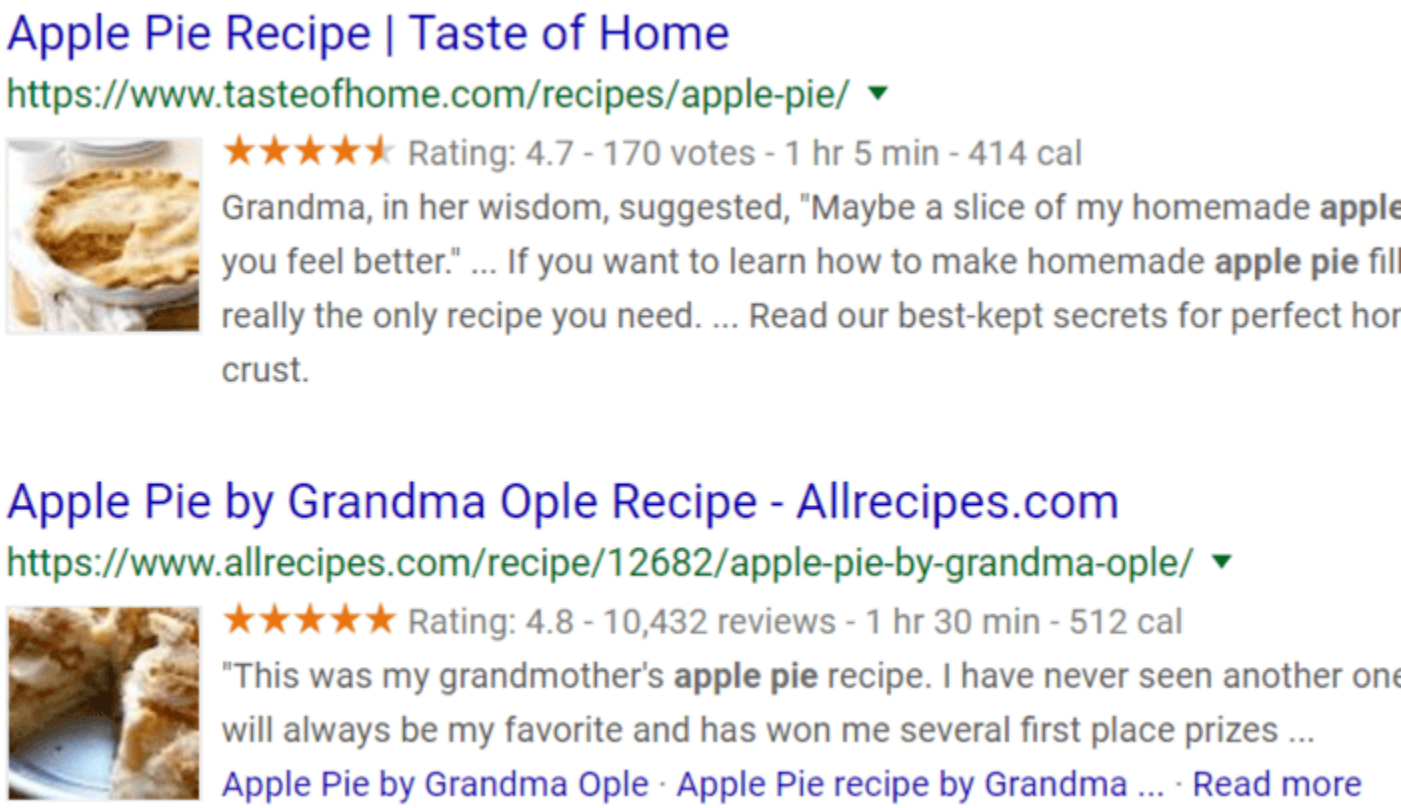
What is structured data?

Structured data describes any data, from numbers to letters, located in a fixed field of a record or file.

```
{
  "@context": "https://schema.org/",
  "@type": "Recipe",
  "name": "Grandma's Holiday Apple Pie",
  "author": "Elaine Smith",
  "image": "http://images.edge-generalmills.com/56459
  "description": "A classic apple pie.",
  "aggregateRating": {
    "@type": "AggregateRating",
    "ratingValue": "4",
    "reviewCount": "276",
    "bestRating": "5",
    "worstRating": "1"
  },
```


From a search engine perspective, structured data provides additional information or explicit clues about a page's content, like if it features a recipe or a how-to series of steps.

A search engine like Google, for example, uses structured data snippets and data labels that appear in your page's HTML code to deliver this information.



These additional details about your page allow Google to tweak how search results appear in the search engine result pages (SERPs). With structured data, you can offer users more value when they search on Google.

How does structured data work?

Understanding how structured data works requires a refresher on how sites appear in search results.

For your website to show up in search results, a crawler, spider, or bot from a search engine (like Google) needs to crawl and index your site. When bots crawl your website, they compile the information on it, from its text to its images to its structured data, and indexes it all.

The search engine then interprets that information using its algorithm.

Deciphering a page isn't always easy, though, which is why search engines continue to update their algorithms. Adopting structured data can help make the work of crawlers and search engines easier too. With structured data, you provide immediate, easy-to-follow information about your page.

Instead of guessing what your page is about, a search engine can determine if your page includes a:

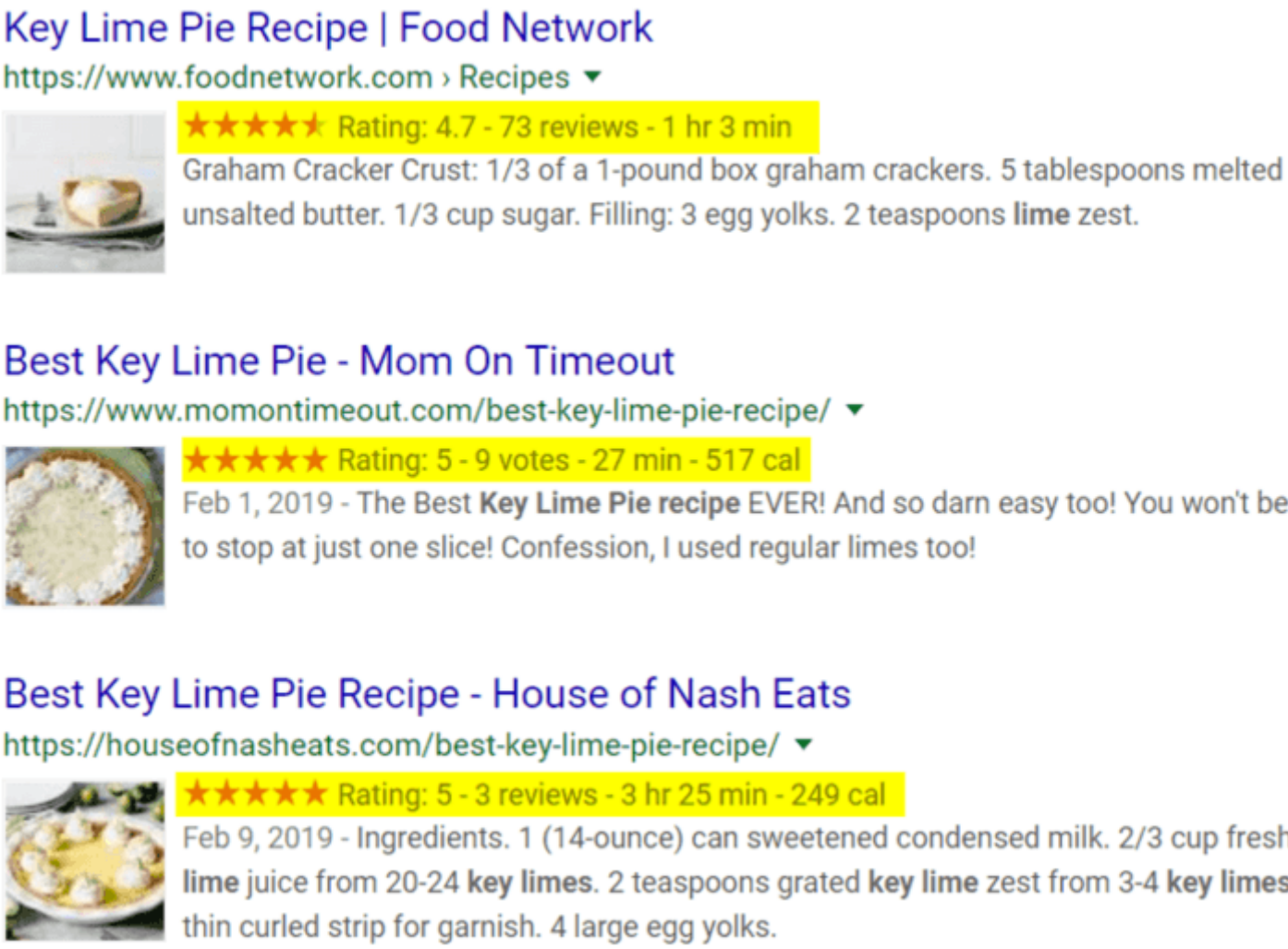
- Recipe
- How-to guide
- Ratings
- Pricing
- Phone number
- FAQ
- Event

- And more

  The search engine can then use that information to streamline a user's search.

  If someone's looking for a key lime pie recipe, for example, they probably have a few preferences that they didn't include in their initial search. For instance, they may need a recipe that takes less than an hour. They most likely want one that delivers when it comes to tastiness, too!

  Structured data can provide all this information for a simple search like "key lime pie recipe."



In the structured data example above, for example, you can see that Google includes the following:

- Each recipe's user rating
- Each recipe's cooking time
- Each recipe's calorie count

Without structured data, Google couldn't provide this information reliably. It would have to guess whether a page

details the cooking time or a user comment and rating section describes the recipe's quality.

When talking about how structured data works, it's worth mentioning that Google doesn't always adapt its search results. Even if you include structured data for a page, like product pricing, Google may skip on adding it for a range of reasons, like search intent.

Structured data analysis (statistics)

From Wikipedia, the free encyclopedia

.

Structured data analysis is the statistical data analysis of structured data. This can arise either in the form of an *a priori* structure such as multiple-choice questionnaires or in situations with the need to search for structure that fits the given data, either exactly or approximately. This structure can then be used for making comparisons, predictions, manipulations etc.[1][2]

Types of structured data analysis

- Algebraic data analysis
- Bayesian analysis
- Cluster analysis
- Combinatorial data analysis
- Formal concept analysis
- Functional data analysis
- Geometric data analysis
- Regression analysis
- Shape analysis
- Topological data analysis
- Tree structured data analysis

References

1. ^ *Brigitte Le Roux; Henry Rouanet (2004). Geometric Data Analysis: from Correspondence Analysis to*

Structured Data Analysis. Springer. ISBN 978-1402022357.

2. ^ *Lawrence J. Hubert, Phipps Arabie, Jacqueline Meulman (2001). Combinatorial Data Analysis: Optimization by Dynamic Programming. SIAM. ISBN 978-0898714784.*

Further reading

- Carlsson, G. (2009) "Topology and Data", *Bulletin (New Series) of the American Mathematical Society*, 46 (2), 255–308
- *James O. Ramsay; B. W. Silverman (2005). Functional data analysis. Springer. ISBN 9780387400808.*
- Leland Wilkinson, (1992) *Tree Structured Data Analysis: AID, CHAID and CART*

Structured data analysis (systems analysis)

From Wikipedia, the free encyclopedia

Structured data analysis (SDA) is a method for analyzing the flow of information within an organization using data flow diagrams. It was originally developed by IBM for systems analysis in electronic data processing, although it has now been adapted for use to describe the flow of information in any kind of project or organization, particularly in the construction industry where the nodes could be departments, contractors, customers, managers, workers etc.

See also

- Information management
- Data modeling
- Structured Systems Analysis and Design Method

Further reading

- CONSTRUCTION SITE INFORMATION NEEDS
- *Norman Fisher; Li Yin Shen; Shen L. Yin (1992). Information Management in a Contractor: A Model of the Flow of Project Data. Thomas Telford. ISBN 978-0727716668.*
- *Institution of Civil Engineers (1991). Applications of Information Technology in Construction. Thomas Telford. ISBN 978-0727716538.*

Structure mining

From Wikipedia, the free encyclopedia

Structure mining or **structured data mining** is the process of finding and extracting useful information from semi-structured

data sets. Graph mining, sequential pattern mining and molecule mining are special cases of structured data mining [citation needed].

Description

The growth of the use of semi-structured data has created new opportunities for data mining, which has traditionally been concerned with tabular data sets, reflecting the strong association between data mining and relational databases. Much of the world's interesting and mineable data does not easily fold into relational databases, though a generation of software engineers have been trained to believe this was the only way to handle data, and data mining algorithms have generally been developed only to cope with tabular data.

XML, being the most frequent way of representing semi-structured data, is able to represent both tabular data and arbitrary trees. Any particular representation of data to be exchanged between two applications in XML is normally described by a schema often written in XSD. Practical examples of such schemata, for instance NewsML, are normally very sophisticated, containing multiple optional sub trees, used for representing special case data. Frequently around 90% of a schema is concerned with the definition of these optional data items and sub-trees.

Messages and data, therefore, that are transmitted or encoded using XML and that conform to the same schema are liable to contain very different data depending on what is being transmitted.

Such data presents large problems for conventional data mining. Two messages that conform to the same schema may have little data in common. Building a training set from such data means that if one were to try to format it as tabular data for conventional data mining, large sections of the tables would or could be empty.

There is a tacit assumption made in the design of most data mining algorithms that the data presented will be complete. The other necessity is that the actual mining algorithms employed, whether supervised or

unsupervised, must be able to handle sparse data. Namely, machine learning algorithms perform badly with incomplete data sets where only part of the information is supplied. For instance methods based on neural networks.[*citation needed*] or Ross Quinlan's ID3 algorithm.[*citation needed*] are highly accurate with good and representative samples of the problem, but perform badly with biased data. Most of times better model presentation with more careful and unbiased representation of input and output is enough. A particularly relevant area where finding the appropriate structure and model is the key issue is text mining.

XPath is the standard mechanism used to refer to nodes and data items within XML. It has similarities to standard techniques for navigating directory hierarchies used in operating systems user interfaces. To data and structure mine XML data of any form, at least two extensions are required to conventional data mining. These are the ability to associate an XPath statement with any data pattern and sub statements with each data node in the data pattern, and the ability to mine the presence and count of any node or set of nodes within the document.

As an example, if one were to represent a family tree in XML, using these extensions one could create a data set containing all the individuals in the tree, data items such as name and age at death, and counts of related nodes, such as number of children. More sophisticated searches could extract data such as grandparents' lifespan etc.

The addition of these data types related to the structure of a document or message facilitates structure mining.

References

- Andrew N Edmonds, *On data mining tree structured data in XML', Data mining UK conference, University of Nottingham, Aug 2003*
- Gusfield, D., *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press, **1997**. ISBN 0-521-58519-8
- R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, John Wiley & Sons, **2001**. ISBN 0-471-05669-3

- F. Hadzic, H. Tan, T.S. Dillon, Mining of Data with Complex Structures, Springer, **2010**. ISBN 978-3-642-17556-5

Chapter 12:

Which American Main Organization & Institutions Using Big Data Centers?

Google Big Data Centers

Google data centers

From Wikipedia, the free encyclopedia

Former Google data center in Eemshaven, Netherlands

Google Data Centers are the large data center facilities Google uses to provide their services, which combine large amounts of digital storage (mainly hard drives and solid-state drives),

compute nodes organized in aisles of racks, internal and external networking, environmental controls (mainly cooling and dehumidification), and operations software (especially as concerns load balancing and fault tolerance).

There is no official data on how many servers there are in Google data centers, but Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers. This number is always changing as the company expands capacity and refreshes its hardware.[1]

Locations

Google data center in The Dalles, Oregon

The locations of Google's various data centers by continent are as follows:[2]

North America:

1. Berkeley County, South Carolina (33°03′50.8″N 80°02′36.1″W) — since 2007, expanded in 2013, 150 employees
2. Council Bluffs, Iowa (41°13′17.7″N 95°51′49.92″W) — announced 2007, first phase completed 2009, expanded 2013 and 2014, 130 employees
3. Douglas County, Georgia (33°44′59.04″N 84°35′5.33″W) — since 2003, 350 employees
4. Bridgeport, Jackson County, Alabama (34°54′48.4″N 85°44′53.1″W)[3][4]
5. Lenoir, North Carolina (35°53′54.78″N 81°32′50.58″W) — announced 2007, completed 2009, over 110 employees
6. Montgomery County, Tennessee (36°37′37.7″N 87°15′27.7″W) — announced 2015
7. Mayes County, Oklahoma at MidAmerica Industrial Park (36°14′28.1″N 95°19′48.22″W) — announced 2007, expanded 2012, over 400 employees[5]
8. The Dalles, Oregon (45°37′57.04″N 121°12′8.16″W) — since 2006, 80 full-time employees

9. Henderson, Nevada — announced in 2018 : 1,210 acres of land bought in 2017 in the Tahoe Reno Industrial Center;[6] project approved by the state of Nevada in November 2018[7]

South America:

1. Quilicura, Chile (33°21′30.5″S 70°41′50.4″W) — announced 2012, online since 2015, up to 20 employees expected. A $140 million investment plan to increase capacity at Quilicura was announced in 2018.[8]
2. Cerrillos, Chile - announced for 2020 [9]

Europe:

1. Saint-Ghislain, Belgium (50°28′09.6″N 3°51′55.7″E) — announced 2007, completed 2010, 12 employees
2. Hamina, Finland (60°32′11.68″N 27°7′1.21″E) — announced 2009, first phase completed 2011, expanded 2012, 90 employees
3. Dublin, Ireland (53°19′12.39″N 6°26′31.43″W) — announced 2011, completed 2012, 150 employees[10]

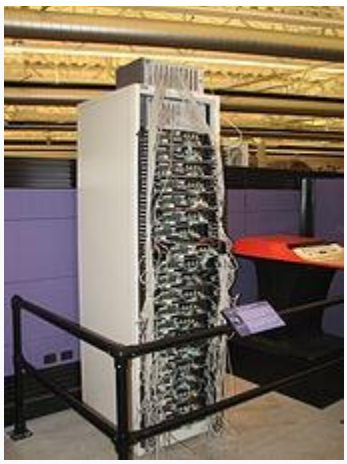
4. Eemshaven, Netherlands (53.425659°N 6.8593522°E) — announced 2014, completed 2016, 200 employees
5. Fredericia, Denmark — announced 2018, planned completion in 2021[11]

Asia:

1. Jurong West, Singapore (1°21′04.8″N 103°42′35.2″E) — announced 2011, completed 2013
2. Changhua County, Taiwan (24°08′18.6″N 120°25′32.6″E) — announced 2011, completed 2013, 60 employees
3. Mumbai, India — announced 2017, completed 2019[12]

Hardware

Original Hardware



Google's first production server rack, circa 1998

The original hardware (circa 1998) that was used by Google when it was located at Stanford University included:[13]

- Sun Microsystems Ultra II with dual 200 MHz processors, and 256 MB of RAM. This was the main machine for the original Backrub system.
- 2 × 300 MHz dual Pentium II servers donated by Intel, they included 512 MB of RAM and 10 × 9 GB hard drives between the two. It was on these that the main search ran.
- F50 IBM RS/6000 donated by IBM, included 4 processors, 512 MB of memory and 8 × 9 GB hard disk drives.
- Two additional boxes included 3 × 9 GB hard drives and 6 x 4 GB hard disk drives respectively (the original storage for Backrub). These were attached to the Sun Ultra II.
- SDD disk expansion box with another 8 × 9 GB hard disk drives donated by IBM.
- Homemade disk box which contained 10 × 9 GB SCSI hard disk drives.

Production Hardware

As of 2014, Google used a heavily customized version of Debian (GNU/Linux). They migrated from a Red Hat-based system incrementally in 2013.[14]

The customization goal is to purchase CPU generations that offer the best performance per dollar, not absolute performance. How this is measured is unclear, but it is likely to incorporate running costs of the entire server, and CPU power consumption could be a significant factor.[15] Servers as of 2009–2010 consisted of custom-made open-top systems containing two processors (each with several cores[16]), a considerable amount of RAM spread over 8 DIMM slots housing double-height DIMMs, and at least two SATA hard disk drives connected through a non-standard ATX-sized power supply unit.[17] The servers were open top so more servers could fit into a rack. According to CNET and a book by John Hennessy, each server had a novel 12-volt battery to reduce costs and improve power efficiency.[16][18]

According to Google their global data center operation electrical power ranges between 500 and

681 megawatts.[19][20] The combined processing power of these servers might have reached from 20 to 100 petaflops in 2008.[21]

Network topology

Details of the Google worldwide private networks are not publicly available, but Google publications[22][23] make references to the "Atlas Top 10" report that ranks Google as the third largest ISP behind Level 3.[24]

In order to run such a large network, with direct connections to as many ISPs as possible at the lowest possible cost, Google has a very open peering policy.[25]

From this site, we can see that the Google network can be accessed from 67 public exchange points and 69 different locations across the world. As of May 2012, Google had 882 Gbit/s of public connectivity (not counting private peering agreements that Google has with the largest ISPs). This public network is used to distribute content to Google users as well as to crawl the Internet to build its search indexes.

The private side of the network is a secret, but a recent disclosure from Google[26] indicate that they

use custom built high-radix switch-routers (with a capacity of 128 × 10 Gigabit Ethernet port) for the wide area network. Running no less than two routers per datacenter (for redundancy) we can conclude that the Google network scales in the terabit per second range (with two fully loaded routers the bi-sectional bandwidth amount to 1,280 Gbit/s).

These custom switch-routers are connected to DWDM devices to interconnect data centers and point of presences (PoP) via dark fibre.

From a datacenter view, the network starts at the rack level, where 19-inch racks are custom-made and contain 40 to 80 servers (20 to 40 1U servers on either side, while new servers are 2U rackmount systems.[27] Each rack has an Ethernet switch). Servers are connected via a 1 Gbit/s Ethernet link to the top of rack switch (TOR). TOR switches are then connected to a gigabit cluster switch using multiple gigabit or ten gigabit uplinks.[28] The cluster switches themselves are interconnected and form the datacenter interconnect fabric (most likely using a dragonfly design rather than a classic butterfly or flattened butterfly layout[29]).

From an operation standpoint, when a client computer attempts to connect to Google, several DNS servers resolve www.google.com into multiple IP addresses via Round Robin policy. Furthermore, this acts as the first level of load balancing and directs the client to different Google clusters. A Google cluster has thousands of servers, and once the client has connected to the server additional load balancing is done to send the queries to the least loaded web server. This makes Google one of the largest and most complex content delivery networks.[30]

Google has numerous data centers scattered around the world. At least 12 significant Google data center installations are located in the United States. The largest known centers are located in The Dalles, Oregon; Atlanta, Georgia; Reston, Virginia; Lenoir, North Carolina; and Moncks Corner, South Carolina.[31] In Europe, the largest known centers are in Eemshaven and Groningen in the Netherlands and Mons, Belgium.[31] Google's Oceania Data Center is claimed to be located in Sydney, Australia.[32]

Project 02

One of the largest Google data centers is located in the town of The Dalles, Oregon, on the Columbia River, approximately 80 miles (129 km) from Portland. Codenamed "Project 02", the $600 million[33] complex was built in 2006 and is approximately the size of two American football fields, with cooling towers four stories high.[34] The site was chosen to take advantage of inexpensive hydroelectric power, and to tap into the region's large surplus of fiber optic cable, a remnant of the dot-com boom. A blueprint of the site appeared in 2008.[35]

Summa Peppermill

In February 2009, Stora Enso announced that they had sold the Summa paper mill in Hamina, Finland to Google for 40 million Euros.[36][37] Google invested 200 million euros on the site to build a data center and announced additional 150 million euro investment in 2012.[38][39] Google chose this location due to the availability and proximity of renewable energy sources.[40]

Modular container data centers[edit]

In 2005,[41] Google was researching a containerized modular data center. Google filed a patent application for this technology in 2003.[42]

Floating Data Centers

In 2013, the press revealed the existence of Google's floating data centers along the coasts of the states of California (Treasure Island's Building 3) and Maine. The development project was maintained under tight secrecy. The data centers are 250 feet long, 72 feet wide, 16 feet deep. The patent for an in-ocean data center cooling technology was bought by Google in 2009[43][44] (along with a wave-powered ship-based data center patent in 2008[45][46]). Shortly thereafter, Google declared that the two massive and secretly-built infrastructures were merely "interactive learning centers, [...] a space where people can learn about new technology."[47]

Google halted work on the barges in late 2013 and began selling off the barges in 2014.[48][49]

Software

Most of the software stack that Google uses on their servers was developed in-house.[50] According to a

well-known Google employee, C++, Java, Python and (more recently) Go are favored over other programming languages.[51] For example, the back end of Gmail is written in Java and the back end of Google Search is written in C++.[52] Google has acknowledged that Python has played an important role from the beginning, and that it continues to do so as the system grows and evolves.[53]

The software that runs the Google infrastructure includes:[54]

- Google Web Server (GWS) – custom Linux-based Web server that Google uses for its online services.
- Storage systems:
  - Google File System and its successor, Colossus[55][56]
  - Bigtable – structured storage built upon GFS/Colossus[55]
  - Spanner – planet-scale database, supporting externally-consistent distributed transactions[55][57]

    - Google F1 – a distributed, quasi-SQL DBMS based on Spanner, substituting a custom version of MySQL.[58]
- Chubby lock service
- MapReduce and Sawzall programming language
- Indexing/search systems:
    - TeraGoogle – Google's large search index (launched in early 2006), designed by Anna Patterson of Cuil fame.[59]
    - Caffeine (Percolator) – continuous indexing system (launched in 2010).[60]
    - Hummingbird – major search index update, including complex search and voice search.[61]
- Borg declarative process scheduling software

Google has developed several abstractions which it uses for storing most of its data:[62]

- Protocol Buffers – "Google's lingua franca for data",[63] a binary serialization format which is widely used within the company.
- SSTable (Sorted Strings Table) – a persistent, ordered, immutable map from keys to values, where both keys and values are arbitrary byte strings. It is also used as one of the building blocks of Bigtable.[64]

- RecordIO – a sequence of variable sized records.[62][65][66]

Software Development Practices

Most operations are read-only. When an update is required, queries are redirected to other servers, so as to simplify consistency issues. Queries are divided into sub-queries, where those sub-queries may be sent to different ducts in parallel, thus reducing the latency time.[27]

To lessen the effects of unavoidable hardware failure, software is designed to be fault tolerant. Thus, when a system goes down, data is still available on other servers, which increases reliability.

Search infrastructure

Index

Like most search engines, Google indexes documents by building a data structure known as inverted index. Such an index obtains a list of documents by a query word. The index is very large due to the number of documents stored in the servers.[30]

The index is partitioned by document IDs into many pieces called shards. Each shard is replicated onto multiple servers. Initially, the index was being served from hard disk drives, as is done in traditional information retrieval (IR) systems. Google dealt with the increasing query volume by increasing number of replicas of each shard and thus increasing number of servers. Soon they found that they had enough servers to keep a copy of the whole index in main memory (although with low replication or no replication at all), and in early 2001 Google switched to an *in-memory index* system. This switch "radically changed many design parameters" of their search system, and allowed for a significant increase in throughput and a large decrease in latency of queries.[67]

In June 2010, Google rolled out a next-generation indexing and serving system called "Caffeine" which can continuously crawl and update the search index. Previously, Google updated its search index in batches using a series of MapReduce jobs. The index was separated into several layers, some of which were updated faster than the others, and the main layer wouldn't be updated for as long as two weeks. With Caffeine the entire index is updated

incrementally on a continuous basis. Later Google revealed a distributed data processing system called "Percolator"[68] which is said to be the basis of Caffeine indexing system.[60][69]

Server types[edit]

Google's server infrastructure is divided into several types, each assigned to a different purpose:[27][30][70][71][72]

- Web servers coordinate the execution of queries sent by users, then format the result into an HTML page. The execution consists of sending queries to index servers, merging the results, computing their rank, retrieving a summary for each hit (using the document server), asking for suggestions from the spelling servers, and finally getting a list of advertisements from the ad server.
- Data-gathering servers are permanently dedicated to spidering the Web. Google's web crawler is known as GoogleBot. They update the index and document databases and apply Google's algorithms to assign ranks to pages.
- Each index server contains a set of index shards. They return a list of document IDs ("docid"), such that documents corresponding to a certain docid

contain the query word. These servers need less disk space, but suffer the greatest CPU workload.
- Document servers store documents. Each document is stored on dozens of document servers. When performing a search, a document server returns a summary for the document based on query words. They can also fetch the complete document when asked. These servers need more disk space.
- Ad servers manage advertisements offered by services like AdWords and AdSense.
- Spelling servers make suggestions about the spelling of queries.

Security

In October 2013, *The Washington Post* reported that the U.S. National Security Agency intercepted communications between Google's data centers, as part of a program named MUSCULAR.[73][74] This wiretapping was made possible because Google did not encrypt data passed inside its own network.[75] Google began encrypting data sent between data centers in 2013.[76]

Environmental Impact

Google's most efficient data center runs at 35 °C (95 °F) using only fresh air cooling, requiring no electrically powered air conditioning; the servers run so hot that humans cannot go near them for extended periods.[77]

In December 2016, Google announced that—starting in 2017—it will power all of its data centers, as well as all of its offices, from 100% renewable energy. The commitment will make Google "the world's largest corporate buyer of renewable power, with commitments reaching 2.6 gigawatts (2,600 megawatts) of wind and solar energy".[78][79][80]

Reference

1. ^ *"How Many Servers Does Google Have?"*. *Data Center Knowledge. Retrieved 20 September 2018.*
2. ^ *marcharla/locations/index.html "Google data centers, locations" Check |url= value (help). Google. Retrieved 21 July 2014.*
3. ^ https://www.google.com/about/datacenters/inside/locations/jackson-county/

4. ^ *Us, Contact; Directory, Staff; Notification, Local Project. "Google kicks off construction on $600M Alabama data center". Made in Alabama. Retrieved 2019-08-19.*
5. ^ *Dawn-Hiscox, Tanwen (February 20, 2018). "Google to spend $600m on Pryor data center expansion". Data Centre Dynamics. Archived from the original on April 23, 2019. Retrieved April 23, 2019.*
6. ^ *Tanwen Dawn-Hiscox (18 April 2017). "Google is planning a massive data center in Nevada". Datacenterdynamics.com. Retrieved 8 December 2018.*
7. ^ *Jason Hidalgo (16 November 2018). "Nevada approves Google's $600M data center near Las Vegas, $25.2M in tax incentives". Rgj.com. Retrieved 8 December 2018.*
8. ^ *"Google ha decido de invertir $140 millones de dólares en su centro de datos en Chile". Newtechmag.net (in Spanish). 28 September 2018. Retrieved 8 December 2018.*
9. ^ *http://www.diarioeldia.cl/tendencias/google-instalara-nuevo-data-center-en-chile*. Missing or empty |title= (help)

10. ^ *"Dublin, Ireland – Data Centers – Google". www.google.com. Retrieved 2019-04-02.*
11. ^ *Sverdlik, Yevgeniy (November 20, 2018). "Google to Build €600M Data Center in Denmark". Data Center Knowledge. Archived from the original on April 16, 2019. Retrieved April 23, 2019.*
12. ^ *Stiver, Dave (1 November 2017). "GCP arrives in India with launch of Mumbai region". Google Cloud Blog. Retrieved 2019-07-30.*
13. ^ *""Google Stanford Hardware"". Archived from the original on February 9, 1999. Retrieved 2017-03-23.. Stanford University (provided by Internet Archive).* Retrieved on July 10, 2006.
14. ^ *Merlin, Marc (2013). "Case Study: Live upgrading many thousand of servers from an ancient Red Hat distribution to a 10 year newer Debian based one" (PDF). Linux Foundation. Retrieved 2017-06-09.*
15. ^ *Tawfik Jelassi; Albrecht Enders (2004). "Case study 16 — Google". Strategies for E-*

business. Pearson Education. p. 424. ISBN 978-0-273-68840-2.

16. ^ Jump up to:[a] [b] Computer Architecture, Fifth Edition: A Quantitative Approach, ISBN 978-0123838728; Chapter Six; 6.7 "A Google Warehouse-Scale Computer" page 471 "Designing motherboards that only need a single 12-volt supply so that the UPS function could be supplied by standard batteries associated with each server"
17. ^ Google's secret power supplies on YouTube
18. ^ Google on-server 12V UPS, 1 April 2009.
19. ^ Google Green infographics
20. ^ *"Analytics Press Growth in data center electricity use 2005 to 2010". Archived from the original on 2012-01-11. Retrieved 2012-05-22.*
21. ^ Google Surpasses Supercomputer Community, Unnoticed?, May 20, 2008.
22. ^ *"Fiber Optic Communication Technologies: What's Needed for Datacenter Network Operations", Research, Google*
23. ^ *"FTTH look ahead — technologies & architectures", Research, Google*

24. ^ *James Pearn. How many servers does Google have?. plus.google.com.*
25. ^ *"kumara ASN15169", Peering DB*
26. ^ *"Urs Holzle", Speakers, Open Network Summit, archived from the original on 2012-05-10, retrieved 2012-05-22*
27. ^ Jump up to: a b c Web Search for a Planet: The Google Cluster Architecture (Luiz André Barroso, Jeffrey Dean, Urs Hölzle)
28. ^ Warehouse size computers
29. ^ Denis Abt High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities
30. ^ Jump up to: a b c *Fiach Reid (2004). "Case Study: The Google search engine". Network Programming in .NET. Digital Press. pp. 251–253. ISBN 978-1-55558-315-6.*
31. ^ Jump up to: a b *Rich Miller (March 27, 2008). "Google Data Center FAQ". Data Center Knowledge. Archived from the original on March 13, 2009. Retrieved March 15, 2009.*
32. ^ *Brett Winterford (March 5, 2010). "Found: Google Australia's secret data network". ITNews. Retrieved 2010-03-20.*

33. ^ Google "The Dalles, Oregon Data Center" Retrieved on January 3, 2011.
34. ^ Markoff, John; Hansell, Saul. "Hiding in Plain Sight, Google Seeks More Power." *New York Times*. June 14, 2006. Retrieved on October 15, 2008.
35. ^ Strand, Ginger. "Google Data Center" *Harper's Magazine*. March 2008. Retrieved on October 15, 2008. Archived August 30, 2012, at the Wayback Machine
36. ^ *"Stora Enso divests Summa Mill premises in Finland for EUR 40 million"*. *Stora Enso*. *2009-02-12. Archived from the original on 2009-04-13. Retrieved 2009-12-02.*
37. ^ [*dead link*] *"Stooora yllätys: Google ostaa Summan tehtaan"*. *Kauppalehti(in Finnish). Helsinki. 2009-02-12. Archived from the original on 2009-02-14. Retrieved 2009-02-12.*
38. ^ *"Google investoi 200 miljoonaa euroa Haminaan"*. *Taloussanomat (in Finnish). Helsinki. 2009-02-04. Retrieved 2009-03-15.*
39. ^ *"Hamina, Finland"*. *Google. Retrieved 2018-04-23.*

40. ^ Finland – First Choice for Siting Your Cloud Computing Data Center.Archived 2013-07-06 at the Wayback Machine Accessed 4 August 2010.
41. ^ https://www.theregister.co.uk/2009/04/10/google_data_center_video
42. ^ *"United States Patent: 7278273"*. *Patft.uspto.gov. Retrieved 2012-02-17.*
43. ^ *Rory Carroll (30 October 2013). "Google's worst-kept secret: floating data centers off US coasts". Theguardian.com. Retrieved 8 December 2018.*
44. ^ *Rich Miller (29 April 2009). "Google Gets Patent for Data Center Barges". Datacenterknowledge.com. Retrieved 8 December 2018.*
45. ^ *Martin Lamonica (8 September 2008). "Google files patent for wave-powered floating data center". Cnet.com. Retrieved 8 December 2018.*
46. ^ *"Google's ship based datacenter patent application surfaces". Datacenterdynamics.com. 7 September 2008. Retrieved 8 December 2018.*
47. ^ *"Google barge mystery solved: they're for 'interactive learning*

centers'". Theguardian.com. 6 November 2013. Retrieved 8 December 2018.

48. ^ *Brandon Bailey (2014-08-01). "Google confirms selling a mystery barge". San Jose Mercury News. Retrieved 2015-04-07.*
49. ^ *Chris Morran (2014-11-07). "What Happened To Those Google Barges?". Consumerist. Retrieved 2017-01-15.*
50. ^ *Mark Levene (2005). An Introduction to Search Engines and Web Navigation. Pearson Education. p. 73. ISBN 978-0-321-30677-7.*
51. ^ *"Python Status Update". Artima. 2006-01-10. Retrieved 2012-02-17.*
52. ^ *"Warning". Panela. Blog-city. Archived from the original on December 28, 2011. Retrieved 2012-02-17.*
53. ^ *"Quotes about Python". Python. Retrieved 2012-02-17.*
54. ^ *"Google Architecture". High Scalability. 2008-11-22. Retrieved 2012-02-17.*
55. ^ Jump up to: *a b c Fikes, Andrew (July 29, 2010), "Storage Architecture and Challenges", TechTalk (PDF), Google*[*permanent dead link*]

56. ^ *"Colossus: Successor to the Google File System (GFS)". SysTutorials. 2012-11-29. Retrieved 2016-05-10.*
57. ^ *Dean, Jeffrey 'Jeff' (2009), "Design, Lessons and Advice from Building Large Distributed Systems", Ladis (keynote talk presentation), Cornell*
58. ^ *Shute, Jeffrey 'Jeff'; Oancea, Mircea; Ellner, Stephan; Handy, Benjamin 'Ben'; Rollins, Eric; Samwel, Bart; Vingralek, Radek; Whipkey, Chad; Chen, Xin; Jegerlehner, Beat; Littlefield, Kyle; Tong, Phoenix (2012), "F1 — the Fault-Tolerant Distributed RDBMS Supporting Google's Ad Business", Research (presentation), Sigmod: Google*
59. ^ *"Anna Patterson – CrunchBase Profile". Crunchbase.com. Retrieved 2012-02-17.*
60. ^ Jump up to:**a** **b** The Register. Google Caffeine jolts worldwide search machine
61. ^ *"Google official release note". Google.com. Retrieved 2013-09-28.*
62. ^ Jump up to:**a** **b** *"Google Developing Caffeine Storage System | TechWeekEurope UK". Eweekeurope.co.uk. 2009-08-18. Archived*

from the original on 2011-11-15. Retrieved 2012-02-17.

63. ^ *"Developer Guide – Protocol Buffers – Google Code". Code.google.com. Retrieved 2012-02-17.*
64. ^ [1]
65. ^ *windley on June 24, 2008 1:10 PM (2008-06-24). "Phil Windley's Technometria | Velocity 08: Storage at Scale". Windley.com. Retrieved 2012-02-17.*
66. ^ *"Message limit – Protocol Buffers | Google Groups". Groups.google.com. Retrieved 2012-02-17.*
67. ^ *"Jeff Dean's keynote at WSDM 2009" (PDF). Retrieved 2012-02-17.*
68. ^ Daniel Peng, Frank Dabek. (2010). Large-scale Incremental Processing Using Distributed Transactions and Notifications. Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation.
69. ^ The Register. Google Percolator – global search jolt sans MapReduce comedown
70. ^ *Chandler Evans (2008). "Google Platform". Future of Google Earth. Madison*

Publishing Company. p. 299. ISBN 978-1-4196-8903-1.

71. ^ *Chris Sherman (2005). "How Google Works". Google Power. McGraw-Hill Professional. pp. 10–11. ISBN 978-0-07-225787-8.*

72. ^ *Michael Miller (2007). "How Google Works". Googlepedia. Pearson Technology Group. pp. 17–18. ISBN 978-0-7897-3639-0.*

73. ^ *Gellman, Barton; Soltani, Ashkan (October 30, 2013). "NSA infiltrates links to Yahoo, Google data centers worldwide, Snowden documents say". The Washington Post. Retrieved November 1, 2013.*

74. ^ *Savage, Charlie; Miller, Claire Cain; Perlroth, Nicole (October 30, 2013). "N.S.A. Said to Tap Google and Yahoo Abroad". The New York Times. Retrieved March 9, 2017.*

75. ^ *Gallagher, Sean (October 31, 2013). "How the NSA's MUSCULAR tapped Google's and Yahoo's private networks". Ars Technica. Condé Nast. Retrieved March 9, 2017.*

76. ^ *Miller, Claire Cain (October 31, 2013). "Angry Over U.S. Surveillance, Tech*

Giants Bolster Defenses". *The New York Times*. *Retrieved March 9, 2017.*

77. ^ *Humphries, Matthew (March 27, 2012). "Google's most efficient data center runs at 95 degrees". geek.com. Archived from the original on June 13, 2016. Retrieved June 13, 2016.*

78. ^ *Hölzle, Urs (December 6, 2016). "We're set to reach 100% renewable energy — and it's just the beginning". The Keyword Google Blog. Retrieved December 8, 2016.*

79. ^ *Statt, Nick (December 6, 2016). "Google just notched a big victory in the fight against climate change". The Verge. Vox Media. Retrieved December 8, 2016.*

80. ^ *Etherington, Darrell (December 7, 2016). "Google says it will hit 100% renewable energy by 2017". TechCrunch. AOL. Retrieved December 8,2016.*

Further Reading

- *L.A. Barroso; J. Dean; U. Hölzle (March–April 2002). "Web search for a planet: The Google cluster architecture" (PDF). IEEE Micro. **23** (2): 22–28. doi:10.1109/MM.2003.1196112.*

- Shankland, Stephen, CNET news "Google uncloaks once-secret server." April 1, 2009.

External links[edit]

- Google Research Publications
- Web Search for a Planet: The Google Cluster Architecture (Luiz André Barroso, Jeffrey Dean, Urs Hölzle)
- Underneath the Covers at Google: Current Systems and Future Directions (Talk given by Jeff Dean at *Google I/O* conference in May 2008)
- Search Engine Optimization

Google Cloud infrastructure

Our planet-scale infrastructure delivers the highest level of performance and availability in a secure, sustainable way.

Trusted global presence

Google Cloud’s footprint spans 61 zones and over 130 points of presence across more than 35 countries worldwide. From redundant cloud regions to high-bandwidth connectivity via subsea cables, every aspect of our infrastructure is designed to deliver your services to your users, no matter where they are around the world.

What’s new

Investing in Google infrastructure, investing in Nevada

Today we're announcing new infrastructure investments in the state of Nevada: a new Google data center and Google Cloud region. These investments will expand our footprint in the southwestern U.S., creating more jobs in the area, improving connectivity and speed for users of Google services and Google Cloud customers, and ensuring that Nevada will become one of the world's onramps to the internet.

Google infrastructure

Infrastructure is a key area of investment for us because it underpins all of the work that we do and supports all of our products. Data centers are the engines of the internet, and as demand for online

content and cloud services continues to increase, our data centers are growing too. They support all of our products including Search, Ads, Maps, YouTube, and Google Cloud. In total, we've invested $47 billion in capex between 2016 and 2018, which includes investments in our infrastructure.
We announced earlier this year that we'll invest another $13 billion in the U.S. alone, including this investment in Nevada.

Globally, Google operates data centers in sixteen locations, and Google Cloud customers are served by 20 cloud regions and 61 availability zones around the world. Once complete, our new site in Nevada will be part of this worldwide network of data centers.

Economic growth and technology in the Silver State

The Las Vegas metro area is home to over two million people and a booming entertainment and gaming industry. Whether you're a Gmail user, a global retailer, or one of the world's largest entertainment corporations, fast access to online content and cloud services are critical to keeping your day running smoothly.

Caesars Entertainment Corporation is among the world's largest hotel operators, with 40,000 rooms around the globe. The data analytics team at Caesars leverages Google Cloud's BigQuery serverless data warehouse and TensorFlow machine learning framework to aggregate data and derive meaningful insights from it. With these valuable insights, Caesars Entertainment has improved the results of their marketing and hospitality initiatives. "Caesars Entertainment selected Google Cloud because we

depend on highly reliable performance as well as scalability for our data analytics initiatives," said Gene Lee, SVP Chief Analytics Officer for Caesars Entertainment. "The addition of a Google Cloud region in Las Vegas, combined with the sophisticated capabilities of BigQuery and TensorFlow, should enable Caesars to further differentiate the gaming, hospitality and entertainment experiences we are able to offer to individual guests."

A new Google data center

At our groundbreaking event in Henderson today, we marked the start of construction on a new Google data center. Today's celebration was attended by Senator Catherine Cortez Masto, Senator Jacky Rosen, Representative Susie Lee, and Governor Steve Sisolak who spoke about how Google continues to invest in the state by bringing tech jobs to the area, giving local non-profits access to over $1 million in funding, and providing additional support for small and large businesses in the state. When it comes online in 2020, the new data center will enhance our ability to provide the fastest and most reliable services for all our users and customers. We're creating more jobs, servicing more customers in the area, and creating economic opportunity by supporting local nonprofits.

A cloud for Nevada

When it launches, the new Google Cloud region in Las Vegas will give organizations in the Western U.S. and those doing business in Nevada faster access to Google Cloud Platform products and tools that will help supercharge their businesses. The region will have three availability zones and will support our portfolio of key GCP products, delivering simple, reliable, and secure infrastructure and lightning fast data analytics and ML/AI capabilities.

And we aren't stopping there—we'll launch our Salt Lake City cloud region in early 2020, for a total of seven Google Cloud regions in the continental United States. These new regions will enable Google Cloud customers to distribute their workloads across up to four regions in the west—Los Angeles, Oregon, Salt Lake City, and Las Vegas—providing even greater connectivity than ever before. Contact sales to learn more about cloud region availability and to get started on GCP today.

A new Home in the Southwest

We believe it's important to invest in the communities that we call home. In this spirit,

today we announced the Google.org Impact Challenge Nevada, a $1,000,000 commitment to Silver State nonprofits with bold and innovative ideas to create economic opportunity in their communities. Beginning today, local nonprofits can submit their proposals to a panel of local judges who will select five winners to receive $175,000 grants and training from Google.org to jumpstart their ideas. Additionally, Nevadans will have a chance to vote for their favorite idea from the five winners, and the “People’s Choice Winner” will receive an additional $125,000 in funding.

Google is proud to call Nevada its newest home, and we’ll continue to invest in communities throughout the state. Thank you for welcoming Google into your communities—we look forward to building out our infrastructure in Nevada and welcoming Google Cloud customers to our Las Vegas region soon.

Secure, efficient data centers

Google Cloud's global network of data centers—among the world's most secure and energy-efficient facilities—run your services 24x7 with the highest possible speed and reliability. Our data centers employ layered security, built-in redundancy and fault tolerance, and strictly limit employee access.

Fast, reliable global network

With a highly provisioned, low-latency network—the same network that powers products like Gmail, Google Search, and YouTube—your traffic stays on Google's private backbone for most of its journey, ensuring exceptional user experience and high performance.

Multi-layered security

We protect your data through progressive infrastructure layers that deliver defense-in-depth. Google Cloud's adherence to data privacy and security standards has earned the trust of third-party auditors who attest that our infrastructure and operations keep user data more secure and compliant.

Designed for high availability

Our data centers and network architecture are designed for maximum reliability and uptime. Your workloads are securely distributed across multiple regions, availability zones, points of presence, and network cables to provide strong built-in redundancy and application availability.

Sustainability built in

Google matches 100% of the energy consumed by our global operations with purchases of renewable energy, so every Google Cloud product you use has zero net carbon emissions. And our hyper-efficient data centers use 50% less energy than most systems.

PayPal processes billions of transactions across the globe. With Google Cloud, we have access to the world's largest network, which helps us reach our infrastructure goals and best serve millions of users.

Sri Shivananda, Senior Vice President and Chief Technology Officer, PayPal

Data and Security

At our data centers, we take security very seriously. We keep your data safe and secure by using dozens of critical security featur

We've designed our data centers with security in mind.

Security is part of our data centers' DNA. We custom-build servers exclusively for our data centers, never selling or distributing them externally. And our industry-leading security team works 24/7 around the globe making our facilities one of the safest places for your data to live.

We also have robust disaster recovery measures in place. For example, in the event of a fire or any other disruption, we shift data access automatically and seamlessly to another data center so that our users can keep working, uninterrupted. Our emergency backup generators continue to power our data centers even in the event of a power failure.

We safeguard your data.

Rather than storing each user's data on a single machine or set of machines, we distribute all data — including our own — across many computers in different locations. We then chunk and replicate the data over multiple systems to avoid a single point of failure. We name these data chunks randomly, as an extra measure of security, making them unreadable to the human eye.

While you work, our servers automatically back up your critical data. So when accidents happen — if your computer crashes or gets stolen — you can be up and running again in seconds.

Lastly, we rigorously track the location and status of each hard drive in our data centers. We destroy hard drives that have reached the end of their lives in a thorough, multi-step process to prevent access to the data.

Our security team is on duty 24/7.

Our data centers are protected with several layers of security to prevent any unauthorized access to your data. We use secure perimeter defense systems, comprehensive camera coverage, biometric authentication, and a 24/7 guard staff. In addition, we enforce a strict access and security policy at our data centers and ensure all staff is trained to be security minded.

We also have local and regional security operations centers covering the entire data center fleet. These

SOCs monitor and respond to alarms at all of our facilities, and are constantly tracking local and global events that could impact operations at our data centers. The security teams also run year-round testing to ensure that we're always prepared to respond to any situation. And along with routine testing, the teams run a robust enterprise risk management program to proactively assess and mitigate any risks to the data centers.

Google’s DATA CENTERS Locations

North America

Berkeley County, South Carolina

Council Bluffs, Iowa

The Dalles, Oregon

Douglas County, Georgia

Henderson, Nevada

Jackson County, Alabama

Lenoir, North Carolina

Loudoun County, Virginia

Mayes County, Oklahoma

Midlothian, Texas

Montgomery County, Tennessee

South America

Quilicura, Chile

Europe

Dublin, Ireland

Eemshaven, Netherlands

Fredericia, Denmark

Hamina, Finland

St. Ghislain, Belgium

Asia

Changhua County, Taiwan

Singapore

Efficiency

When you use Google products, the servers in our data centers do the work for you — around the clock and around the world. Our servers support many products at a time. That's "the cloud." By keeping our servers busy, we can do more with less — more searches, more Gmail, and more YouTube videos with fewer servers and less energy. We've worked hard to minimize the environmental impact of these services so that when you use our products, you're also being good to the environment.

Helping businesses save energy in the cloud

The cloud supports many products at a time, so it can more efficiently distribute resources among many users. That means we can do more with less energy — and businesses can too. In 2013, Lawrence Berkeley National Laboratory published research indicating that moving all office workers in the United States to the cloud could reduce the energy used by information technology by up to 87%.

Related specifically to Google products, by switching to Google Apps, companies have reduced office computing costs, energy use, and carbon emissions by 65-90%. Additionally, businesses that use Gmail have decreased the environmental impact of their email service by up to 98% compared to those that run email on local servers.

Because of our energy efficiency efforts, our cloud is better for the environment. This means businesses that use our cloud-based products are greener too.

How we do it

Our data centers use much less energy than the typical data center. We raise the temperature to 80°F, use outside air for cooling, and build custom servers. We also share detailed performance data to help move the entire industry forward.

Measuring and improving our energy use

We're focused on reducing our energy use while serving the explosive growth of the internet. Most data centers use almost as much non-computing or "overhead" energy (like cooling and power conversion) as they do to power their servers. At Google we've reduced this overhead to only 11%. That way, most of the energy we use powers the machines directly serving Google searches and products. We take detailed measurements to continually push toward doing more with less — serving more users while wasting less energy.

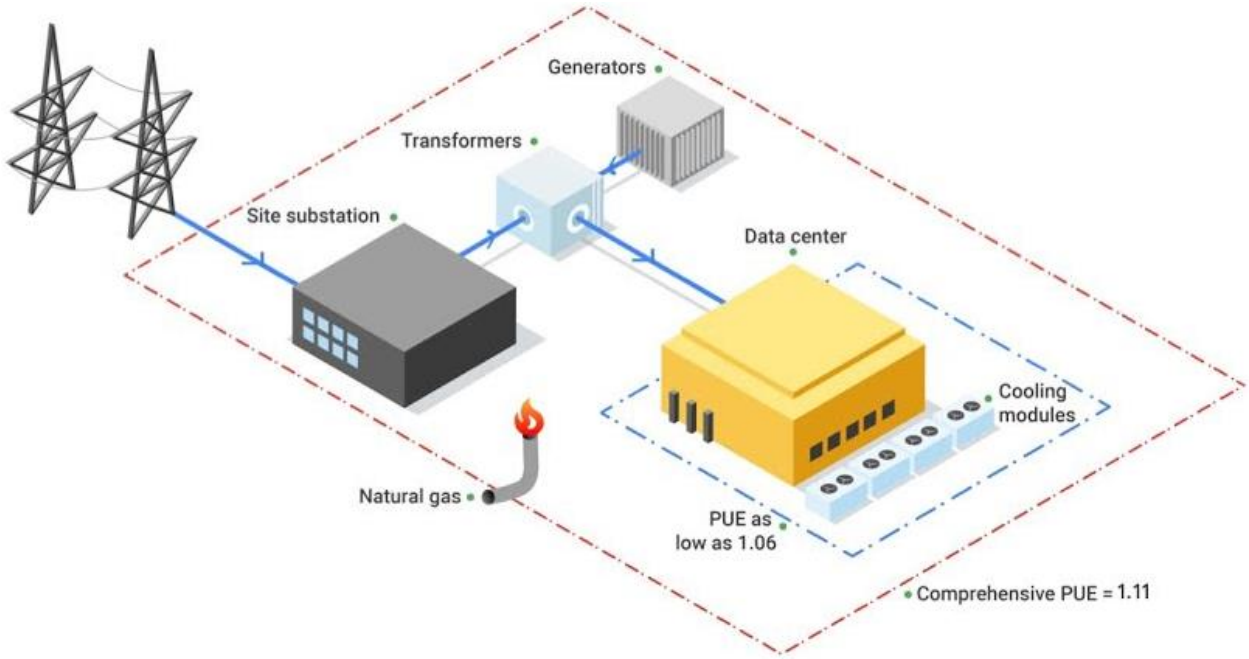


Figure 1: Google Data Center PUE measurement boundaries. The average PUE for all Google Data Centers is 1.11, although we could boast a PUE as low as 1.06 when using narrower boundaries.

We take the most comprehensive approach to measuring Power Usage Effectiveness (PUE)

Our calculations include the performance of our entire fleet of data centers around the world — not just our newest and best facilities. We also continuously measure throughout the year — not just during cooler seasons.

Additionally, we include all sources of overhead in our efficiency metric. We could report much lower numbers if we took the loosest interpretation of the Green Grid's PUE measurement standards. In fact, our best site could boast a PUE of less than 1.06 if we used an interpretation commonly used in the industry. However, we're sticking to a higher standard because we believe it's better to measure and optimize everything on our site, not just part of it. Therefore, we report a comprehensive trailing twelve-month

(TTM) PUE of 1.11 across all our large-scale data centers (once they reach stable operations), in all seasons, including all sources of overhead.

Google Data Center PUE performance

Our fleet-wide PUE has dropped significantly since we first started reporting our numbers in 2008. The TTM energy-weighted average PUE for all Google data centers is 1.11, making our data centers among the most efficient in the world.

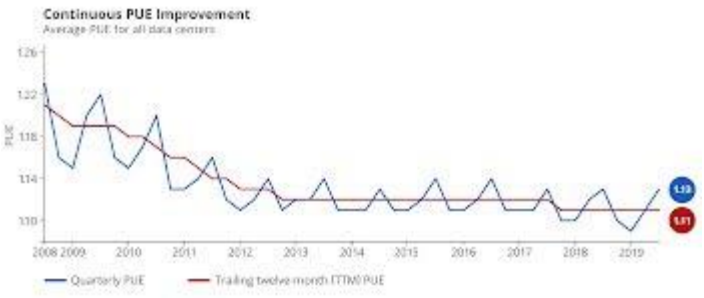


Figure 2: PUE data for all large-scale Google Data Centers

Yearly performance report

- 2019
- 2018
- 2017
- 2016
- 2015
- 2014
- 2013
- 2012

- 2011
- 2010
- 2009
- 2008
  - QUARTER 01
  - QUARTER 02
  - QUARTER 03

Fleet wide PUE	Quarterly PUE	
Fleet	1.09	1.11
Campuses	**Quarterly PUE**	
Douglas County, Georgia	1.10	1.12
Lenoir, North Carolina	1.09	1.10
Berkeley County, South Carolina	1.08	1.11
Council Bluffs, Iowa	1.12	1.12
Council Bluffs, Iowa (2nd facility)	1.08	1.09
Mayes County, Oklahoma	1.08	1.10
The Dalles, Oregon	1.11	1.12
The Dalles, Oregon (2nd facility)	1.07	*
Dublin, Ireland	1.13	1.12
St. Ghislain, Belgium	1.09	1.09

*We report individual campus TTM PUE only for campuses with at least twelve months of data.

- For Q1 2019, TTM PUE was 1.11, equal to last year's value. Quarterly PUE was 1.09, slightly less than last year's 1.10.
- For individual campuses, our lowest TTM PUE was 1.09, in Iowa, Belgium, and Finland. Our lowest quarterly PUE was 1.08, in South Carolina, Iowa, and Oklahoma.
- Our highest TTM PUE was 1.17, in Singapore. Our highest quarterly PUE was 1.15, in Singapore.

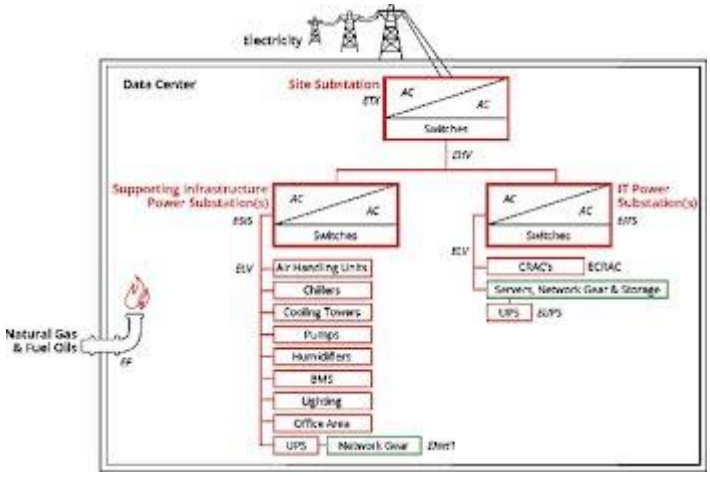
Measurement FAQ

What's the average PUE of other data centers?

Why does our data vary?

How do we get our PUE data?

What do we include in our calculations?

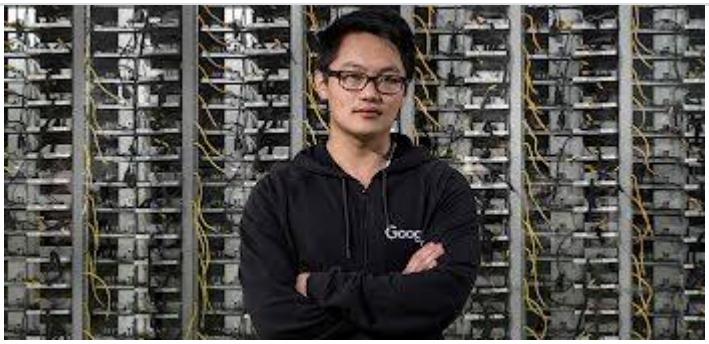


$$PUE = \frac{ESIS + EITS + ETX + EHV + ELV + EF}{EITS - ECRAC - EUPS - ELV + ENet1}$$

-

Power Usage Effectiveness (PUE)

The data center industry uses the measurement PUE, or power usage effectiveness, to measure efficiency. A PUE of 2.0 means that for every watt of IT power, an additional watt is consumed to cool and distribute power to the IT equipment. A PUE closer to 1.0 means nearly all of the energy is used for computing.

- 

Data centers and machine learning

The virtual world is built on physical infrastructure, and all those racks of humming servers use vast amounts of energy. Together, all existing data centers use roughly 2% of the world’s electricity; if left unchecked, this

energy demand could grow as rapidly as internet use. So making data centers run as efficiently as possible is a very big deal — and that's what we set out to do.

See what's new and browse our innovations

-

Google's European infrastructure

2019

Digital transformation is a defining challenge and opportunity for the European economy. Copenhagen Economics' 2019 study details Google’s efforts to support Europe's economy while remaining energy efficient.

-

Unleashing digital opportunities in Europe
2019

Plans to invest €3 billion to expand our data centers across Europe over the next two years will bring our total investment in Europe's internet infrastructure to €15 billion since 2007. Our investments generate economic activity for the region and support more than 13,000 full-time jobs in the EU every year.

-

Our biggest renewable energy purchase ever

2019

Google has made the [biggest corporate purchase of renewable energy in history](), which will also spur the construction of more than $2 billion in new energy infrastructure, including millions of solar panels and hundreds of wind turbines spread across three continents.

-

Environmental report

2019

Operating our business in an environmentally sustainable way has always been a Google core value. Our 2019 environmental report covers the 2018 financial year, in which we matched 100% of the electricity consumption of our global operations with renewable energy for the second consecutive year.

-

100% renewable energy, for the second year in a row

2019

In 2017, we first reached our longstanding goal of buying enough renewable energy to match 100% of Google's global annual electricity use. During 2018, our purchases of energy from sources like solar and wind once again matched our entire annual electricity consumption.

-

Google Data Centers and community

2018

At Google we care about the communities where we work and live. Our data centers help keep the internet up and running 24/7 while supporting local economies and job creation. Learn about their positive impact in the Oxford Economics report for the US and in the Copenhagen Economics report for Europe.

-

100% renewable energy — just the beginning

2017

In 2017 we achieved a great milestone: purchasing 100% renewable energy to match consumption for global operations, including our data centers and offices. But we're also working to make the electricity supply greener as a whole — not just for us, but for everyone.

-

Introducing the Data Center Mural Project

2016

Data centers play an important role in our everyday lives, so we asked four artists to reimagine the exterior walls of our data centers in Oklahoma, Iowa, Ireland, and Belgium. Watch videos that tell the stories of how these artists were inspired to bring a bit of the magic from the inside to the outside.

-

Virtual reality tour inside a Google Data Center

2015

People often ask what's inside a Google Data Center. In this 360° virtual tour (taken at The Dalles, Oregon), you can see where the Google Cloud lives and the numerous high security measures we use to protect not only customer data but our own custom-made infrastructure.

-

Harnessing our data to drive efficiency

2014

In our ongoing pursuit of extreme efficiency, we've developed software that optimizes our data centers by predicting and guiding our energy use. In 2014, we shared our model so that others could follow suit and enhance their data center efficiency in similar ways.

-

Increasing access to renewable energy

2013

As demand for renewable energy increases, we've been working with electricity providers on scalable solutions that increase the supply of clean energy. In 2013, we partnered with Duke Energy to pilot a new type of arrangement that lets companies buy renewable energy directly from the utility.

-

Showing what we're made of — inside and out

2012

For a rare glimpse into where we run our products – like Search, YouTube, Gmail, Drive, and Photos, browse our photo gallery of the technology, people, and places that keep Google's products running 24/7.

-

Certifying our high environmental standards

2011

In 2011, we became the first major internet services company to gain external certification of our high environmental and workplace standards throughout our US and European data centers, reinforcing our commitment to take care of the environment and our employees.

-

Adding renewable energy to our communities

2010

We aim to use as much clean energy as possible, but we're limited by a grid with few renewable options. In 2010, we started buying renewable energy from wind farms near our data centers. In doing so, we're adding new clean power for our data centers and greening the electricity where we operate.

-

Sharing what we've learned

2009

In 2009, we brought together our peers to discuss how we can all make the industry greener and more efficient. This was followed by an event on efficiency targets across geographies and climates, and a summit exploring the environmental impact and benefits of the internet.

-

Disclosing industry-leading efficiency

2008

Most companies don’t talk about their real-world efficiency. In 2008, we started sharing our efficiency performance to help push the industry forward. Since then, we've continued to publish detailed, comprehensive data for all of our data centers.

-

Becoming the first carbon neutral internet company

2007

Our data centers use much less energy than the average data center, and if you add our renewable energy use and carbon offsets, our footprint is zero. We've been carbon neutral since 2007, and we're happy to see that other tech companies have followed in our footsteps.

-

Designing efficient data centers

2003

In 2003, we designed portable data centers: shipping containers pre-packed with servers and cooling equipment, which were easier to assemble and more energy efficient. By 2006, we moved to newer designs that provided the same benefits, but with greater flexibility.

-

Pioneering highly efficient servers

2001

We've been designing our own custom servers to minimize their energy use since 2001. By using highly efficient power supplies and putting batteries directly onto the machines, we've built some of the most efficient servers in the industry.

Renewable Energy

We're sourcing clean energy for a better future.

In 2017, Google achieved a major milestone: purchasing 100% renewable energy to match our annual electricity consumption for global operations, including our data centers and offices. . Our large-scale procurement of wind and solar power is a cornerstone of our sustainability efforts, and has made Google the world's largest corporate buyer of renewable energy. To date, we've signed more than 50 long-term contract commitments to buy energy from wind and solar farms, resulting in more than $7 billion in new construction across four continents, and thousands of related jobs.

Although our 100% renewable milestone signifies that we buy enough renewable energy over the course of a year to match our annual electricity consumption, it does not mean that our facilities are matched with renewable energy in every hour of every day. To compensate for times and places in which the wind slows or sunlight fades, we currently buy a surplus of renewable energy at other times and in other places. Our ultimate goal, however, is to source enough carbon-free energy to match our electricity consumption in all places, at all times. We are actively exploring strategies to achieve 24/7 carbon-free energy for all of our data centers, and are carefully toward that aspiration.

We're breaking new ground in scaling up renewable energy.

By purchasing renewable energy from wind and solar farms, we're adding new clean power near our data centers and the communities around us.

When we sign agreements to purchase clean power, we choose projects that have a material impact in bringing new renewable energy sources to the market. By providing developers with a solid commitment, we help them get the money they need to finance new clean energy facilities. In exchange, we get clean energy at competitive prices, as well as the renewable energy certificates to help reduce our carbon footprint. Together, we make the grid a little bit greener.
In addition, we work closely with utility companies and other energy providers to pioneer new pathways for renewable energy adoption, from Tennessee to Taiwan.
LEARN HOW WE BUY RENEWABLE ENERGY

We're investing in a brighter future for everyone.

We’re going beyond investing in renewables for our own operations — we want to make carbon-free power more accessible for consumers of all types and sizes. To promote carbon-free solutions, we share technology and insights that help others learn about the potential of renewable energy for solving environmental challenges.
LEARN HOW WE’RE SOLVING FOR SUSTAINABILITY

-

The Google Environmental Report

To meet the challenges posed by climate change and the need for resource efficiency, we work to empower everyone — businesses, governments, non-profit organizations, communities, and individuals — to use Google technology to create a more sustainable world. This report is a chance for us to reflect on what we’ve achieved, the lessons we’ve learned along the way, and the opportunities and challenges we see ahead.

-

The Internet is 24x7 — carbon-free energy should be too

Our long-term challenge is to source carbon-free energy to match our electricity consumption in all places, at all times. This means realizing a future where each Google facility is always matched — around the clock — with local carbon-free power. Some of our data centers are already performing remarkably well on this front, and others are making progress but still have room to improve.

See what's new and browse our innovations

-

Google's European infrastructure

2019

Digital transformation is a defining challenge and opportunity for the European economy. Copenhagen Economics' 2019 study details Google’s efforts to support Europe's economy while remaining energy efficient.

-

Unleashing digital opportunities in Europe
2019

Plans to invest €3 billion to expand our data centers across Europe over the next two years will bring our total investment in Europe's internet infrastructure to €15 billion since 2007. Our investments generate economic activity for the region and support more than 13,000 full-time jobs in the EU every year.

-

Our biggest renewable energy purchase ever

2019

Google has made the biggest corporate purchase of renewable energy in history, which will also spur the construction of more than $2 billion in new energy infrastructure, including millions of solar panels and hundreds of wind turbines spread across three continents.

-

Environmental report

2019

Operating our business in an environmentally sustainable way has always been a Google core value. Our 2019 environmental report covers the 2018 financial year, in which we matched 100% of the electricity consumption of our global operations with renewable energy for the second consecutive year.

-

100% renewable energy, for the second year in a row

2019

In 2017, we first reached our longstanding goal of buying enough renewable energy to match 100% of Google's global annual electricity use. During 2018, our purchases of energy from sources like solar and wind once again matched our entire annual electricity consumption.

-

Google Data Centers and community

2018

At Google we care about the communities where we work and live. Our data centers help keep the internet up and running 24/7 while supporting local economies and job creation. Learn about their positive impact in the Oxford Economics report for the US and in the Copenhagen Economics report for Europe.

-

100% renewable energy — just the beginning

2017

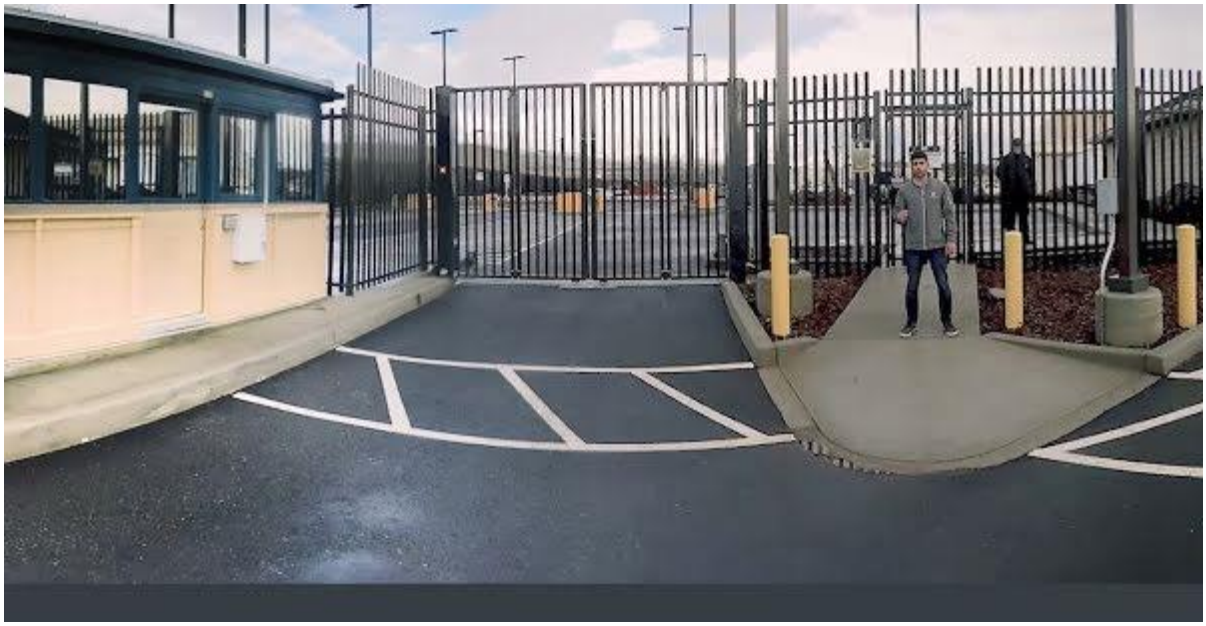
In 2017 we achieved a great milestone: purchasing 100% renewable energy to match consumption for global operations, including our data centers and offices. But we're also working to make the electricity supply greener as a whole — not just for us, but for everyone.

-

Introducing the Data Center Mural Project

2016

Data centers play an important role in our everyday lives, so we asked four artists to reimagine the exterior walls of our data centers in Oklahoma, Iowa, Ireland, and Belgium. Watch videos that tell the stories of how these artists were inspired to bring a bit of the magic from the inside to the outside.

- 

Virtual reality tour inside a Google Data Center

2015

People often ask what's inside a Google Data Center. In this 360° virtual tour (taken at The Dalles, Oregon), you can see where the Google Cloud lives and the numerous high security measures we use to protect not only customer data but our own custom-made infrastructure.

-

Harnessing our data to drive efficiency

2014

In our ongoing pursuit of extreme efficiency, we've developed software that optimizes our data centers by predicting and guiding our energy use. In 2014, we shared our model so that others could follow suit and enhance their data center efficiency in similar ways.

-

Increasing access to renewable energy

2013

As demand for renewable energy increases, we've been working with electricity providers on scalable solutions that increase the supply of clean energy. In 2013, we partnered with Duke Energy to pilot a new type of arrangement that lets companies buy renewable energy directly from the utility.

-

Showing what we're made of — inside and out

2012

For a rare glimpse into where we run our products – like Search, YouTube, Gmail, Drive, and Photos, browse our photo gallery of the technology, people, and places that keep Google's products running 24/7.

-

Certifying our high environmental standards

2011

In 2011, we became the first major internet services company to gain external certification of our high environmental and workplace standards throughout our US and European data centers, reinforcing our commitment to take care of the environment and our employees.

-

Adding renewable energy to our communities

2010

We aim to use as much clean energy as possible, but we're limited by a grid with few renewable options. In 2010, we started buying renewable energy from wind farms near our data centers. In doing so, we're adding new clean power for our data centers and greening the electricity where we operate.

-

Sharing what we've learned

2009

In 2009, we brought together our peers to discuss how we can all make the industry greener and more efficient. This was followed by an event on efficiency targets across geographies and climates, and a summit exploring the environmental impact and benefits of the internet.

-

Disclosing industry-leading efficiency

2008

Most companies don’t talk about their real-world efficiency. In 2008, we started sharing our efficiency performance to help push the industry forward. Since then, we've continued to publish detailed, comprehensive data for all of our data centers.

-

Becoming the first carbon neutral internet company

2007

Our data centers use much less energy than the average data center, and if you add our renewable energy use and carbon offsets, our footprint is zero. We've been carbon neutral since 2007, and we're happy to see that other tech companies have followed in our footsteps.

-

Designing efficient data centers

2003

In 2003, we designed portable data centers: shipping containers pre-packed with servers and cooling equipment, which were easier to assemble and more energy efficient. By 2006, we moved to newer designs that provided the same benefits, but with greater flexibility.

-

Pioneering highly efficient servers

2001

We've been designing our own custom servers to minimize their energy use since 2001. By using highly efficient power supplies and putting batteries directly onto the machines, we've built some of the most efficient servers in the industry.

Berkeley County, South Carolina

Google is proud to call South Carolina home to one of our data centers.

In 2007, we announced plans to construct a data center complex in Berkeley County, South Carolina. We announced the expansion of the facility in 2013, bringing our total investment to $1.8 billion and establishing a long-term commitment to the region and state. We've created more than 400 jobs on site, and we work hard to support the communities in which our employees live and work.

Since 2009, we've awarded more than $2 million to local schools and nonprofits. Additionally, we've helped implement free downtown WiFi networks throughout the Lowcountry.

Why did Google choose Berkeley County?

Berkeley County has the right combination of energy infrastructure, developable land, and available workforce for the data center.

An innovative cooling system

We collect rainwater as a water source for our cooling system, and we've ensured that our proposed groundwater usage will be sustainable.

Explore our photo gallery

-
-

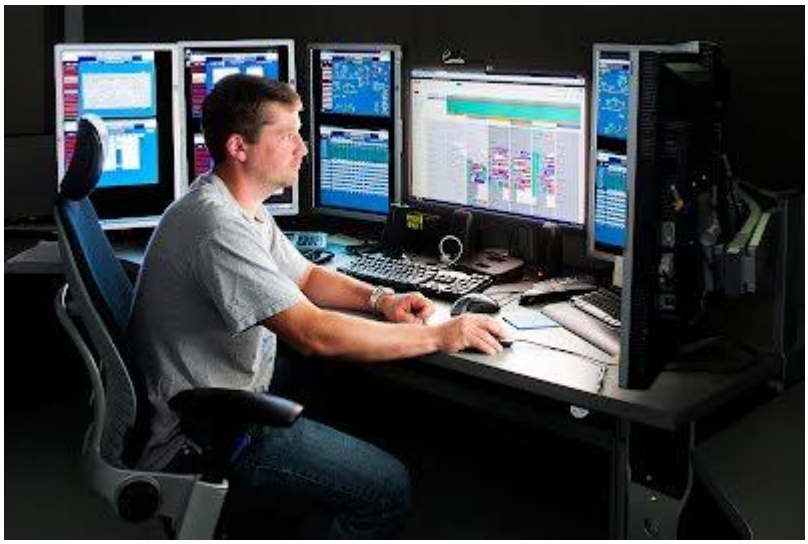
-
-

-
-

-
-

-
-

-
-

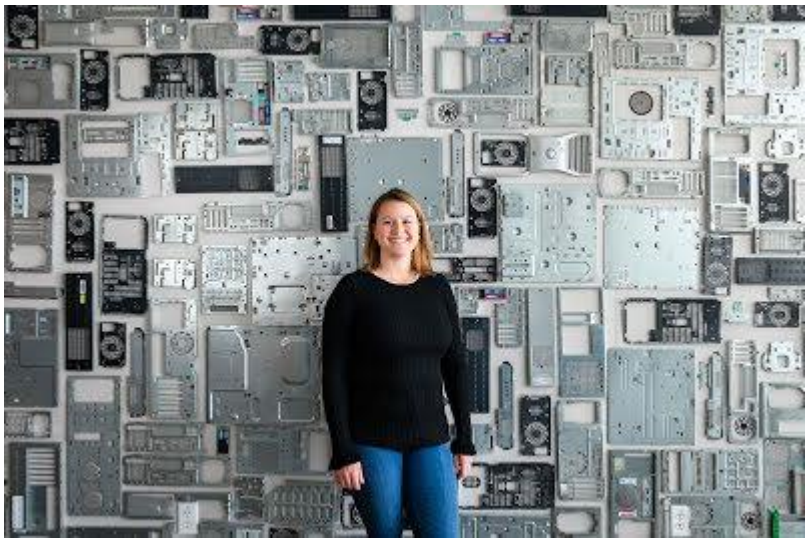
- 
-

-
-

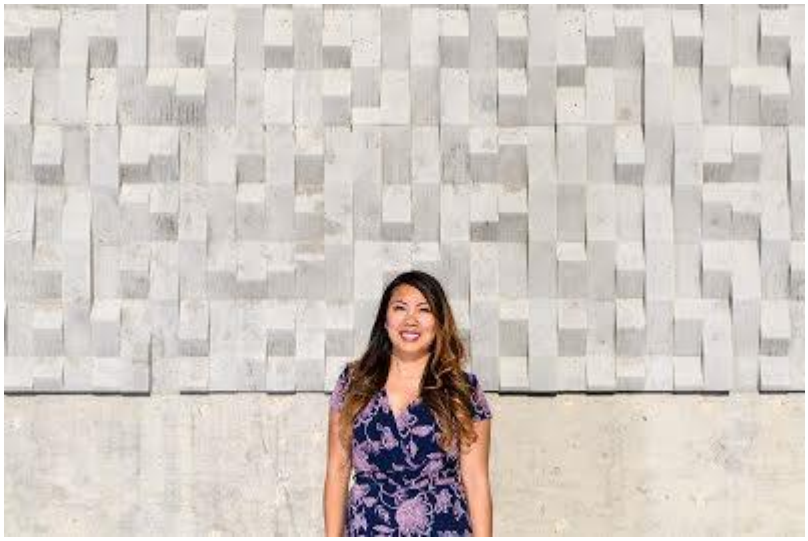
-
-

-
-

-
-

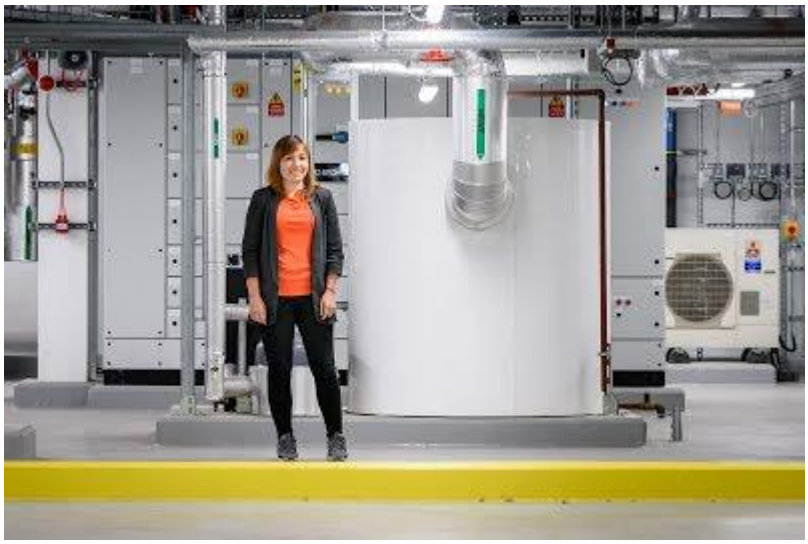
-
- 

-
-

-
- 

-
-

-
-

- 
-

-
-

-
-

-
-
PCWR

-
-

-
Google
-

-
-
2
3
4
5

-
-

-
-

-
-

-
-

-
-

-
-

-
BULK STORAGE
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
DANGER
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

-
-

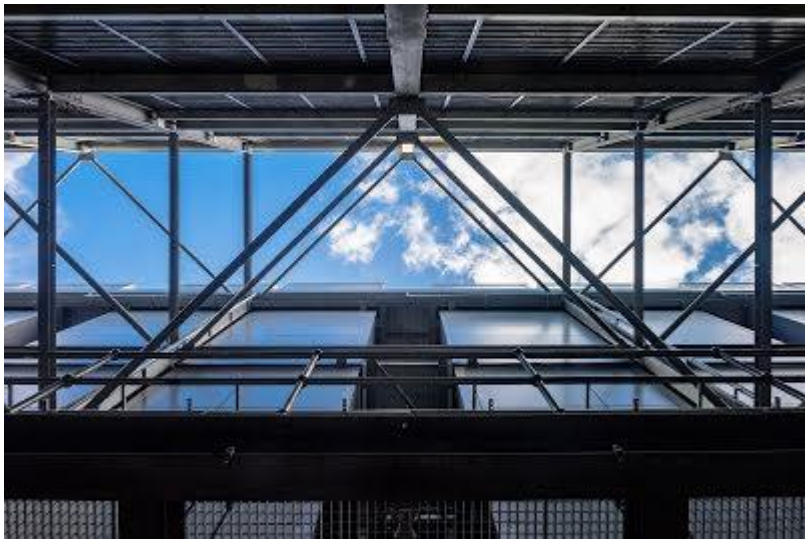
-
-

-
-

-
-

-
-

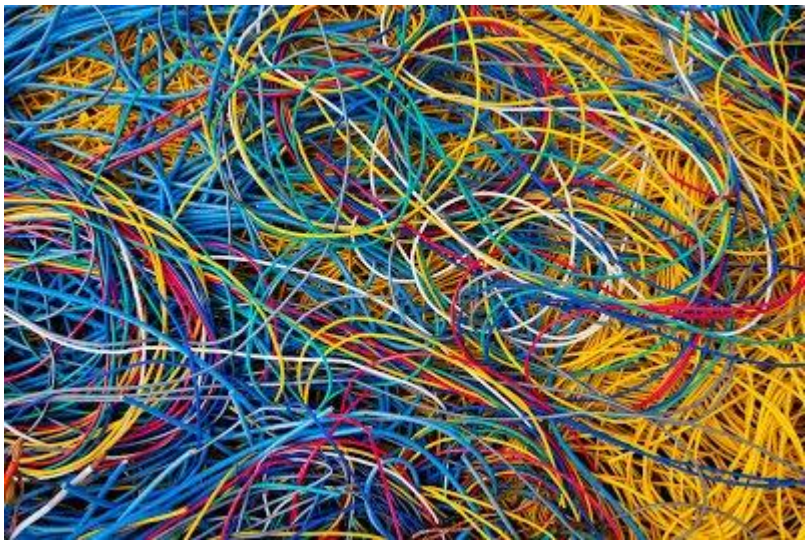
-
-

-
- 

-
-

-
-

-
-

- 
-

-
-

-
1
-

-
-

-
-

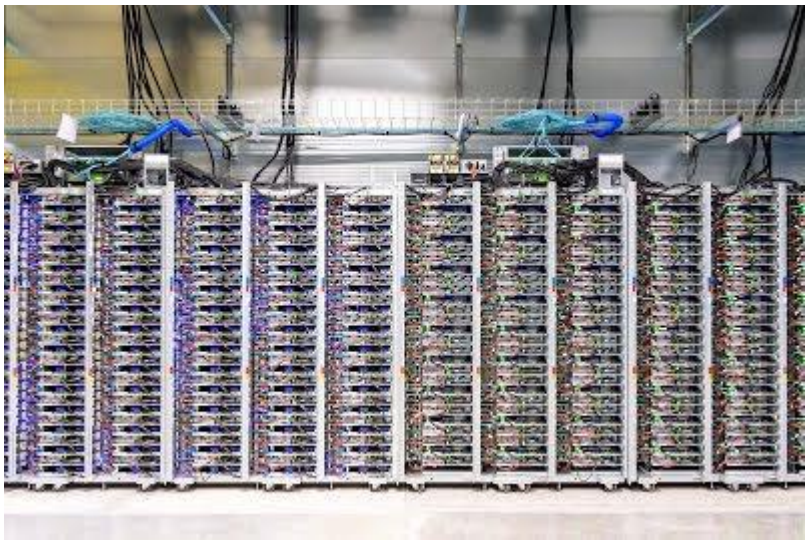
-
-

-
-

-
-

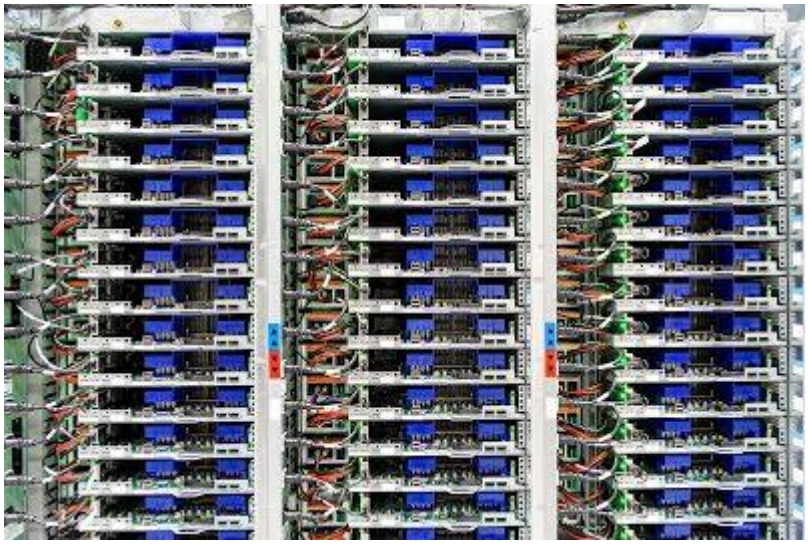
-
-

-
-

-
- 

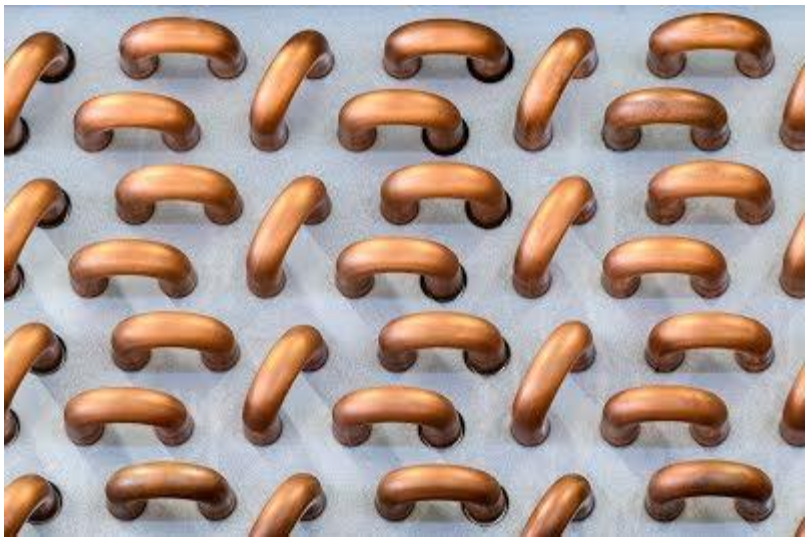
-
ROW
ROW
-

-
-

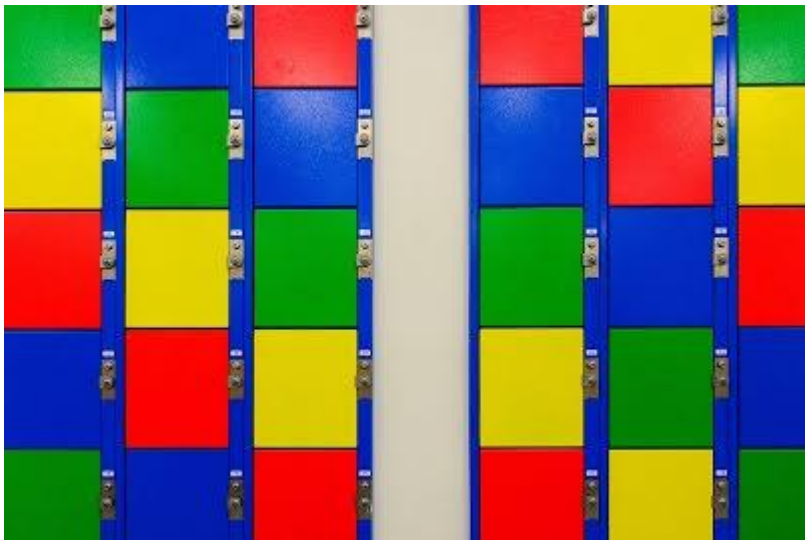
-
- 

-
-

-
-

- 
-

-
-

- 

Life at Google Data Centers

Meet the amazing people behind the technology that lets us build for everyone

Careers at Google Data Centers

Aishah

Data Center Facilities Specialist

Bachelor's degree in chemical engineering, MBA in finance

15 years’ experience in the maintenance industry

Mother of four beautiful children

Mario

Data Center Technician

Chose from over 30 college football scholarships

Former professional athlete with the Arena Football League and Buffalo Bills

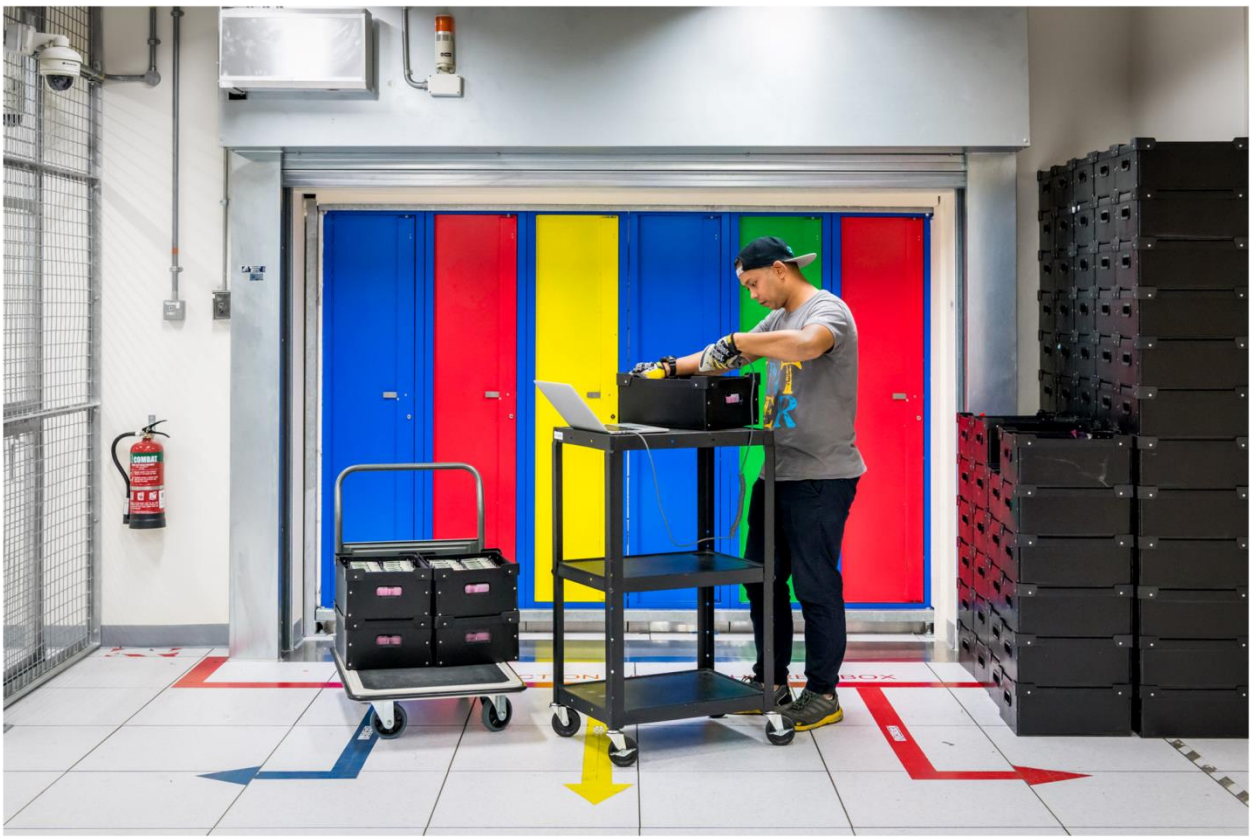
Team player with a strong work ethic and ability to adapt

Anne-France

Construction Quality Program Manager

Alain

Data Center Facilities Manager



Haikal

Data Center Technician, Hardware Operations

Has a diploma in both machine technology and information technology

Maintains servers and ensures user data security

Loves to occasionally play table tennis at work

Derek

Data Center Facilities Technician

Keeps the Google Data Center online

Trained to fix all issues promptly and efficiently

Satisfied with job well done

Neha

Director, Operations, Data Center Energy Strategy

Background in electric utilities

Leads team that buys electric infrastructure and energy

Her projects span the globe

Google Data Center FAQ

Google CEO Sundar Pichai speaks during Google I/O 2016 at the Shoreline Amphitheatre on May 19, 2016 in Mountain View, California.

Google Data Center FAQ

https://www.datacenterknowledge.com/archives/2017/03/16/google-data-center-faq

Everything you ever wanted to know (and didn't) about Google data centers but were afraid to ask

Data Center Knowledge | Mar 17, 2017

Google is the largest, most-used search engine in the world, with a global market share that has held steady at about 90 percent since Google Search launched in 1997 as Backrub. In 2017, Google became the most valuable brand in the world, topping Apple, according to the Brand Finance Global 500 report. Google's position is due mainly to its core business as a search engine and its ability to transform users into payers via advertising.

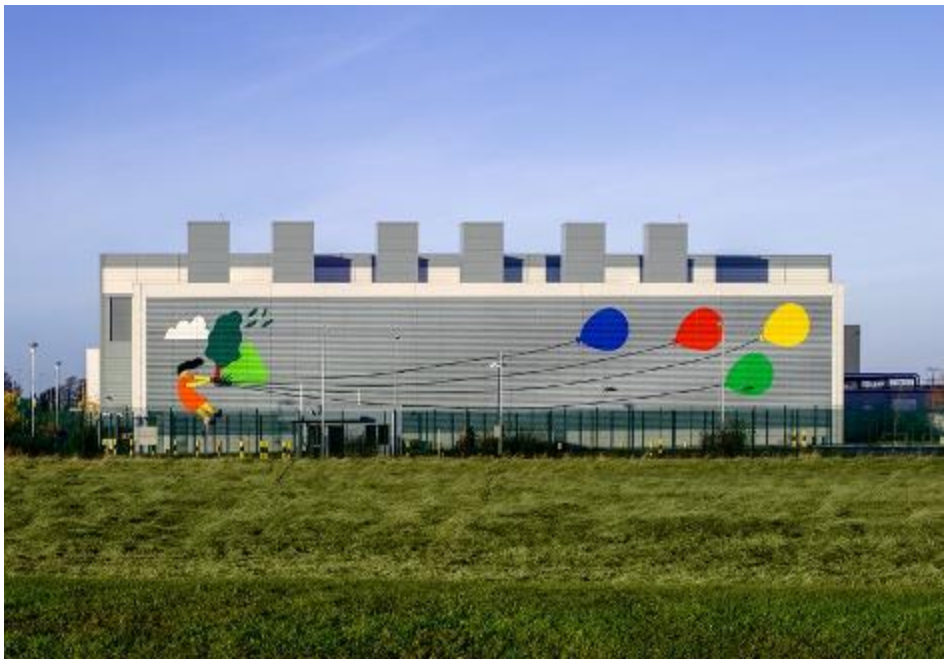
About 32 percent of Google visitors come from the US, where the company holds 63.9 percent of the search engine market, according to statista.com. Google had 247 million unique US users in November 2015. Globally, it boasts 1.5 billion search engine users and more than 1 billion users of Gmail.

Google data centers process an average of 40 million searches per second, resulting in 3.5 billion searches per day and 1.2 trillion searches per year, Internet Live Stats reports. That's up from 795.2 million searchers per year in 1999; one year after Google was launched.

In reorganization in October 2015, Google became a subsidiary of a new company it created called Alphabet. Since then, several projects have been canceled or scaled back, including the halt of further rollout of Google Fiber. Following the reorg, however, Google has placed a lot of focus (and dedicated a lot of resources) to selling cloud services

to enterprises, going head-to-head against the market giant Amazon Web Services and the second-largest player in the space, Microsoft Azure.

That has meant a major expansion of Google data centers specifically to support those cloud services. At the Google Cloud Next conference in San Francisco in March 2017, the company's execs revealed that it spent nearly $30 billion on data centers over the preceding three years. While the company already has what is probably the world's largest cloud, it was not built to support enterprise cloud services. To do that, the company needs to have data centers in more locations, and that's what it has been doing, adding new locations to support cloud services and adding cloud data center capacity wherever it makes sense in existing locations.



The largest of several murals illustrator Fuchsia MacAree painted on the walls of Google's data center in Dublin (Photo: Google)

Here are some of the most frequently asked questions about Google data centers and our best stab at answering them:

Where are Google Data Centers Located?

Google lists eight data center locations in the U.S., one in South America, four in Europe and two in Asia. Its cloud sites, however, are expanding, and Google's cloud map shows many points of presence

worldwide. The company also has many caching sites in collocation facilities throughout the world, whose locations it does not share.

This far-flung network is necessary not only to support operations than run 24/7, but to meet specific regulations (like the EU's privacy regulations) of certain regions and to ensure business continuity in the face of risks like natural disasters.

Google data centers in the Dalles, Oregon, 2006 (Photo by Craig Mitchelldyer/Getty Images)

In the works as of March 2017, are Google data centers for cloud services in California, Canada, The Netherlands, Northern Virginia, São Paulo, London, Finland, Frankfurt, Mumbai, Singapore, and Sydney.

Here are the data center sites listed by Google:

North America:

- Berkeley County, South Carolina
- Council Bluffs, Iowa
- Douglas County, Georgia
- Jackson County, Alabama
- Lenoir, North Carolina
- Mayes County, Oklahoma
- Montgomery County, Tennessee
- The Dalles, Oregon

South America

- Quilicura, Chile

Asia

- Changhua County, Taiwan
- Singapore

Europe

- Dublin, Ireland
- Eemshaven, Netherlands
- Hamina, Finland
- St Ghislain, Belgium

How big are Google Data Centers?

A paper presented during the IEEE 802.3bs Task Force in May 2014 estimates the size of five of Google's US facilities as:

- Pryor Creek (Mayes County), Oklahoma , 980,000 square feet
- Lenoir, North Carolina, 337,000 square feet
- The Dalles, Oregon, 200,000 square feet (before the 2016, 164,000 square foot expansion)
- Council Bluff, Iowa, 200,000 square feet
- Berkely County, South Carolina, 200,000 square feet.

Many of these sites have multiple data center buildings, as Google prefers to build additional structures as sites expand rather than containing operations in a single massive building.

Google itself doesn’t disclose the size of its data centers. Instead, it mentions the cost of the sites or number of employees. Sometimes, facility size slips out. For example, the announcement about the opening of The Dallas in Oregon said the initial building was 164,000 square feet. The size of subsequent expansions, however, has been kept tightly under wraps.

Reports discussing Google’s new data center in Emeshaven, Netherlands, which opened December 2016, didn’t mention size. Instead, they said the company has contracted for the entire 62 Megawatt output of a nearby wind farm and ran 9,941 miles of computer cable within the facility. The data center employs 150 people.

How Many Servers Does Google Have?

There’s no official data on how many servers there are in Google data centers, but Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers. This number, of course, is always

changing as the company expands capacity and refreshes its hardware.

Dell Lowers Annual Sales Forecast on Weak Server Demand

The move signals that geopolitical and trade tensions have dented corporate demand for servers.

Bloomberg | Nov 26, 2019

Nico Grant (Bloomberg) -- Dell Technologies Inc. lowered its annual sales forecast, signaling that geopolitical and trade tensions have dented corporate demand for servers.

Adjusted sales will be $91.8 billion to $92.5 billion for fiscal year 2020, Dell Chief Financial Officer Tom Sweet said Tuesday during a conference call with analysts. The company said in August that revenue would be $93 billion to $94.5 billion in the fiscal year. Shares fell more than 4% in extended trading.

Earlier, Dell reported adjusted revenue rose 1.2% to $22.9 billion in the fiscal third quarter, just missing analysts' average estimate of $23 billion.

Big Vendors Blame Macroeconomics for Weak Data Center Hardware Sales

HPE, Dell, Cisco report slowing hardware sales amid uncertainty caused by trade disputes among other tensions.

Yevgeniy Sverdlik | Nov 26, 2019

Corporate spending on data center servers, storage, and networking hardware has slowed down, and the largest companies that sell this equipment are faulting macroeconomic uncertainty caused, according to them, by international trade disputes and other geopolitical tensions.

Hewlett Packard Enterprise, Cisco Systems, and Dell Technologies all reported earnings for the last full quarter this month (Dell's report came Tuesday), and weak enterprise data center hardware purchasing was a common theme across the board.

Cisco started seeing "some weakness" due to global macroeconomics in the quarter before last, CEO Chuck Robbins said on the company's earnings call earlier this month. It was mainly apparent in service-provider and emerging-markets segments at the time, he said. But the weakness has now broadened to also include enterprise and commercial segments, he said.

"Ongoing global trade tensions and other geopolitical factors have created uncertainty that

contributes to an uneven demand environment," HPE CEO Antonio Neri said on his company's earnings call Monday. "The elongation of sale cycles we experienced since Q2 continues, particularly in larger deals."

HPE's overall revenue was down 9.1 percent in the quarter – to $7.22 billion, according to Bloomberg. Server revenue was down 13 percent and storage hardware dropped 12 percent.

Dell's data center hardware revenue for the reported quarter was down 6.1 percent year over year, sagging to $8.39 billion. Within the category, servers and networking products dropped by 16 percent, while storage revenue was up 6.9 percent, according to Bloomberg.

The Dell-controlled data center software company VMware did much better, reporting $2.46 billion in revenue for the quarter – a 12-percent increase.

Cisco's Infrastructure Platforms segment, which includes data center products, declined 1 percent, to

$84 million. There was revenue growth from Cisco data center switches, however, driven by Nexus 9000 Series sales, the company said.

Likewise, Cisco's HyperFlex converged infrastructure hardware drove growth, but a decrease in sales of routing products dragged the entire segment down. Cisco did not specify how much revenue growth it saw from data center switches and HyperFlex.

Execs of the big incumbent data center hardware vendors sought to reassure analysts that the strategies they had in place made them optimistic about the future.

Both Dell and HPE have been reorienting their businesses to focus on subscription-based services to take advantage of hybrid cloud, which they expect to be the architecture of choice for enterprises going forward. That means their customers will want to keep control of at least some hardware that runs their applications and stores their data – not move

everything into the giant public clouds run by Amazon, Microsoft, Google, and others.

Hoping to recreate the cloud user experience in their customers' own data centers, they are marketing on-premises hardware services, where customers order equipment through a web interface and get it delivered, installed, and managed by the vendors while paying only for the capacity they use and the time they use it for.

Dell and HPE are also positioning themselves as end-to-end infrastructure providers for the modern enterprise, from core on-prem and cloud data centers to the edge.

Customers "need a technology partner that has the expertise, the tools, and a flexible delivery model to help them harness the power of their data across all their clouds and edges," Neri said. "HPE is uniquely positioned to meet these needs and provide a consistent cloud-like experience for apps and data everywhere."

Robbins has been repositioning Cisco in hopes of making it less dependent on hardware sales and more on software and services.

In the shorter term, however, what will help the overall market is less uncertainty and more “clarity,” Robbins told analysts earlier this month:

“If you just go around the world right now and you look at what's happening in Hong Kong, you look at the China-US trade situation, you look at what's going on in DC, you've got Brexit, you've got uncertainty in Latin America.” If these issues “get resolved, then you could see some of the uncertainty removed… Business confidence just suffers when there's lack of clarity, and there's been lack of clarity for so long that I think it finally just came into play.”

ITPro Today™

Multicloud Architecture: 3 Common Performance Challenges and Solutions

Careful upfront planning is required to help ensure that your multicloud architecture doesn't create performance bottlenecks.

Christopher Tozzi | Nov 26, 2019

Multicloud architectures offer many benefits, but they also create some special performance challenges--especially if the teams who design and manage multicloud deployments don’t pay careful attention to the performance bottlenecks that can creep up within multicloud architectures.

Here’s a look at three of the most common multicloud performance difficulties, along with tips on overcoming them.

Multicloud Architecture Performance Overview

By now, many IT professionals are familiar with the advantages of a multicloud strategy. By allowing organizations to run workloads on more than one cloud at once, multicloud offers opportunities for greater cost effiency and reliability.

In addition, a multicloud strategy can in some cases help to improve overall workload performance. You might choose to deploy an application using one service from one cloud vendor and another service from another cloud vendor, because that approach offers better performance at a better price point than relying on the services of a single vendor.

Performance Challenges in Multicloud Architectures

Yet, while multicloud architectures offer some potential performance advantages, they can also lead to performance shortcomings. The more clouds you have in the mix, the easier it becomes to connect them together in ways that create performance bottlenecks, thereby slowing down workloads that would run faster if they were hosted in the same cloud.

These multicloud architecture performance problems can be avoided, but only with the proper planning:

1. Networking between clouds

First and foremost, consider network connections that span clouds.

When you have two applications or services exchanging data within the same cloud, the data typically does not need to travel over the internet; instead, it stays within the cloud provider's infrastructure. Network bandwidth and latency rates can vary depending on whether the data is traveling between different data centers or regions. In virtually all cases, however, data transferred over the network within the same cloud is going to move a lot faster than data that has to travel across the Internet from one cloud to another.

What this means is that network connections between clouds can become a serious performance bottleneck for multicloud architectures.

Given that the network is basically the only way to connect one cloud service to another, there is no way to avoid network performance bottlenecks entirely. However, IT teams can deploy a few strategies to mitigate the problem:

- Avoid multicloud architectures where large volumes of data are stored in one cloud but need to be processed in another. You might be tempted, for instance, to use one cloud provider's storage service because it is cheaper, while feeding data from that service to an application hosted in another cloud. This might save a little money, but it may not be worth the performance cost.
- When possible, compress data before it leaves one cloud and moves into another. Compressing data might increase your cloud compute bill, but it will improve performance.

- In cases where workloads are mirrored across two or more clouds to improve reliability, design the workloads in such a way that each cloud's instance of the workload can operate even if its data is not synced with the other instance's version of the data. This approach ensures that data transfer delays won't interrupt workload performance.

2. Monitoring multiple clouds

Another frequent performance challenge for multicloud architectures is the increased difficulty of monitoring multiple clouds. When monitoring your clouds becomes harder, identifying performance or availability issues within the clouds is harder, too.

The best way to avoid this pitfall is to rely on cloud monitoring tools that can monitor all of your clouds at once. These days, virtually all APM solutions support all of the major clouds, so it's easy enough to find a tool that fits the bill.

Keep in mind, however, that optimal performance monitoring for multiple clouds includes not just monitoring all of your clouds at once, but configuring the tools to understand the nuances of multicloud workloads. In other words, your tools must recognize that two services running in different clouds are connected and depend upon each other in order to be able to alert you effectively to potential problems.

3. Scaling limitations

One of the key advantages of cloud computing in general is the ability to scale resource allocations for a workload up or down quickly when demand shifts.

Within a single cloud, it's quite easy to configure auto scaling of workloads using the cloud vendor's native auto scaling tool. But when your workloads span multiple clouds, auto scaling becomes trickier. You can't use Azure's autoscaling framework to scale up the AWS-based components of your multicloud workload, and vice versa.

You can, of course, configure auto scaling on each cloud separately. That approach will work well enough, and the manual effort it requires may not be so much as to overwhelm an IT team (since auto scaling configuration is typically a set-it-and-forget-it affair).

However, in cases where auto scaling for multicloud workloads is particularly complex, teams might consider relying on a universal control plane to set up and manage their cloud environments. A universal control plane will automate scaling and load-balancing across clouds automatically, eliminating the need to configure auto scaling in each cloud individually.

Conclusion

Multicloud strategies can boost cloud performance, but only with the right architectural design and tooling. Without planning for network bandwidth bottlenecks between clouds, as well as multicloud monitoring and scaling challenges, organizations risk missing out on the performance opportunities that multicloud confers.

HPE CEO Antonio Neri

HPE Reports Revenue That Misses Estimates on Server Demand

The report signals that corporate demand for data center hardware remains in a slump amid slowing economic growth.

Bloomberg | Nov 26, 2019

Nico Grant (Bloomberg) -- Hewlett Packard Enterprise Co. reported sales that fell short of Wall Street estimates, signaling that corporate demand for data-center hardware remains in a slump amid slowing economic growth.

Revenue dropped 9.1% to $7.22 billion, the San Jose, California-based company said Monday in a statement. It marked the fourth consecutive quarter of year-over-year sales declines. Analysts, on average, projected $7.42 billion.

HPE Chief Executive Officer Antonio Neri is trying to increase sales by moving the server maker to a subscription business model. By 2022, all HPE hardware and software will be available as a service or via a pay-per-use model. The company recently bought Cray Inc.

to bolster its position in the supercomputer market. HPE has cut expenses, including the size of its workforce, as part of a restructuring meant to modernize the company and boost its profit margin.

Like other makers of servers and storage hardware, HPE has been contending with slowing economic growth and geopolitical tensions around the globe, which generally make businesses more cautious about buying new hardware. Cisco Systems Inc. said this month that customers had slowed down or reduced new purchases due to business uncertainty.

HPE generates about two-thirds of its quarterly revenue from overseas sales.

“Global trade tension is causing uncertainty, which results in longer sales cycles,” Neri said in an interview. He also pointed to other geopolitical factors causing some clients to pause orders.

“We have put the focus on particular areas where we want to grow faster,” Neri said, citing businesses including high-performance computing and the company’s GreenLake hybrid cloud-computing product that lets customers pay for services based on how much they use. “That’s why we’re confident. We have a better portfolio

and we have better coverage even though we see weakness in the market."

HPE shares declined about 4% in extended trading after closing at $17.45 in New York. The stock has increased 32% this year.

"Hewlett Packard Enterprise's sales growth will likely remain anemic for fiscal 1H amid seasonal weakness and pressured corporate capital spending," Anand Srinivasan, a Bloomberg Intelligence analyst, said in a note before the results were released.

Server sales dropped 13% to $3.23 billion and storage hardware revenue fell 12% to $848 million in the quarter ended Oct. 31.

Neri reiterated his pledge that the company will deliver 1% to 3% sales growth in the next three years.

HPE posted fiscal fourth-quarter net income of $480 million, or 36 cents a share, from a loss of $757 million, or 53 cents. Profit, excluding some items, was 49 cents a share. Analysts projected 46 cents.

Adjusted profit will be 42 cents to 46 cents per share in the current period. Analysts on average expected 43 cents, according to data

compiled by Bloomberg. The company also affirmed its recently issued fiscal 2020 outlook of $1.78 to $1.94 per share.

AUSTIN, TX - MARCH 27: Title sponsor CEO Michael Dell watches the prize-giving after the championship match of the World Golf Championships-Dell Match Play at the Austin Country Club on March 27, 2016 in Austin, Texas. (Photo by Tom Pennington/Getty Images)

Dell to Explore Sale of RSA Cybersecurity Unit

The company hopes the sale could fetch at least $1 billion, sources who asked not to be identified said.

Bloomberg | Nov 26, 2019

Liana Baker and Nico Grant (Bloomberg) -- Dell Technologies Inc. is exploring a sale of RSA Security, a cybersecurity business it hopes could fetch at least $1 billion, including debt, according to people familiar with the matter.

Discussions are at an early stage and there is no guarantee Dell will end up selling the asset, said the people, asking not to be identified because the matter is private.

A representative for Dell declined to comment.

RSA helps companies detect, investigate and respond to security risks as well as reduce intellectual property theft, fraud and cybercrime, according to Dell's annual report. RSA, which says it has more than 30,000 customers, is perhaps best known for its

software tokens, which generate random codes to enable access to corporate networks.

Dell acquired the business via its 2016 takeover of EMC Corp., which had purchased RSA for about $2.1 billion in 2006.

RSA has also had to deal with stiff competition in recent years from rivals including Okta Inc. and Ping Identity Holding Corp.

Dell, one of the largest personal computer makers in the world, reported that sales rose 2% in the quarter ended Aug. 2 from a year earlier, as growth in corporate sales of desktops and laptops overcame weakness in its server and storage business.

The Secret to a Thriving Data Center: More Diversity

We still have a long way to go before we reach gender equality in the data center, and the recent boom of "women in tech" might be misleading.

Nov 25, 2019

By Ivonne Valdes, Sales VP, Cloud and Service Provider Segment, Schneider Electric

Five years ago, some of the largest tech companies released diversity reports for the first time to demonstrate needed progress. Yet we've made little progress. In fact, we are not even close to gender neutrality.

The IT and data center industry is a subsector where the divide is especially prevalent. According to the Bureau of Labor, women filled 36 percent of IT roles in 1991. Fast forward to 2016, when, according to Deloitte, this number declined to 25 percent. We still have a long way to go before we reach gender equality in the data center, and the recent boom of "women in tech" might be misleading.

Create Opportunities to Solve an Industry-Wide Problem

Data centers face a staffing crisis, with most finding it difficult to recruit and retain staff. At the same time, the lack of women in the data center is particularly notable compared to those in other subsectors. One-quarter of facility operators surveyed under the Uptime Institute 2019 data center report had no women at all among their design, build or operations staff, and only five percent said women represented 50 percent or more of their staff. As women, we comprise 49 percent of the world population, so a greater focus on recruiting females has the potential to alleviate the staffing crisis.

The urgent need to create more opportunities for women in the data center – including positions among the C-suite, the boardroom and

middle management – starts with awareness and exposure. There is a saying: “If I can see it, I can be it.” The industry needs to invest in creating awareness campaigns to educate women on the plethora of opportunities that exist in this space.

Studies show that women will only apply for a position when they can check all the boxes, in contrast with men, who may apply for roles when they only meet a portion of the criteria. We need to create more gender-neutral job descriptions that convey the actual day-to-day job. In today’s globalized world, collaboration, teamwork, ability to multitask, and organization skills are critical success factors.

Navigate Your Own Career Journey

While there’s still plenty the industry can do to support young female professionals interested in a career in IT, there is also a significant opportunity for us to create our destiny and help change history. Here are my top recommendations for women in the data center to not only support a prosperous and fulfilling career but also pave the path for future female leaders:

- **Define your “why.”** Learn your business, define your why, and put a price tag on your worth. A clear sense of purpose enables you to focus on what matters most and enables you to push forward.

- **Take risks.** The benefit usually outweighs the risks. Success and achievements are worked toward and gained by risky behavior. If you take the shot and make it, you feel a sense of accomplishment; if you miss, you at least tried.
- **Learn, learn, learn.** Never pass on an opportunity to learn. Learning keeps you relevant in our ever-changing world and is the best job-security tool anyone can have.
- **Build a support team.** A solid stakeholder network supports progress. A support team brings new ideas, solves problems, helps build morale and creates a system to ensure that deadlines are met and high-quality work is delivered.
- **Give back.** Giving back can help boost your career in ways you can't even imagine. For example, connecting with young professionals can spark new ideas, build your personal confidence, enable you to be a mentor, and expand your network.

We are in the middle of the fourth industrial revolution – an inflection point for everything digital, AI, ML, and other connected technologies. I encourage women to approach this transformational era as a blank canvas allowing them to define, create, and tailor their careers. Today, more than ever before in history, we have the flexibility on where, how, and why we work. It's time to act on the endless opportunities that technology provides.

Opinions expressed in the article above do not necessarily reflect the opinions of Data Center Knowledge and Informa.

Industry Perspectives is a content channel at Data Center Knowledge highlighting thought leadership in the data center arena. See our guidelines and submission process for information on participating.

0 COMMENTS

Network room at a Google data center in Council Bluffs, Iowa

Google Cloud’s Bare Metal Solution Aims at Legacy Workloads with Baggage

It’s less like traditional bare-metal cloud services and more like traditional dedicated hosting services.

Yevgeniy Sverdlik | Nov 25, 2019

Google Cloud is going after those legacy enterprise workloads that are especially difficult to “lift and shift” from on-premises data centers to the cloud.

The recently announced Google Bare Metal Solution is less like traditional bare-metal cloud services and more like traditional dedicated hosting services. If you have legacy applications that must run on dedicated hardware and have strict hardware certification requirements and “complicated licensing and support agreements,” Google is offering to provide the hardware you need, provision, and manage it for you inside a colocation data center. It will connect that environment to its public cloud infrastructure, so you can run your

legacy applications and still take advantage of all the modern cloud-native services.

Oracle databases are a primary target here, mentioned twice in the Google Cloud blog post announcing the service. Enterprises store their most critical data in Oracle databases, but they are hard to move to the cloud. Oracle's own cloud-services play revolves around making it easier for customers to plug their databases into the cloud, either via migration to Oracle Cloud data centers or by connecting on-prem enterprise infrastructure to Oracle Cloud.

Microsoft Azure and Oracle recently partnered to help companies interconnect their Oracle databases and Azure cloud services.

For now, Google Cloud's Bare Metal Solution is available in the US East region only, a Google spokesperson told DCK via email. That means Google can deploy hardware for you inside one of the many collocation data centers in Northern Virginia and connect it to its own cloud data centers there.

While the spokesperson said the company had "global expansion plans" for the service next year, the rate of expansion will depend on how much traction it will get. Asked whether Google would lease new colo space for Bare Metal for customers in places where it

doesn't already have some, they said, "We are open to expanding based on the customer needs and projected demand."

Not hosting the hardware in the same data centers as its main cloud infrastructure is one big difference from traditional bare-metal cloud services. Another is the hardware itself. Google is offering OEM hardware for this service. That's standard hardware from the likes of Hewlett Packard Enterprise or Dell Technologies, not Google's custom-designed, straight-from-the-contract-manufacturer-conveyor-belt cloud hardware. The OEM hardware is certified for "multiple enterprise applications," the company said.

The server hardware comes in five different configurations, but the company is willing to customize on a case-by-case basis. Storage comes in 1TB increments, either hybrid or all-flash.

Google Cloud Bare Metal Solutions server configurations:

- Dual-socket x86 systems
  - 16 core with 384 GB DRAM
  - 24 core with 768 GB DRAM
  - 56 core with 1536 GB DRAM
- Quad-socket x86 systems
  - 56 core with 1536 GB DRAM
  - 112 core with 3072 GB DRAM

You will have access to dev/test environments “within hours of ordering the solution,” but expect four to eight weeks of lead time for larger deployments and production workloads, the spokesperson told us.

Google will bill you for bare metal once a month, integrated with your regular Google Cloud billing, but it does have a “preferred term length of 36 months,” which likely means you can’t expect Google’s remote hands to set up a single server for you in a colo unless you commit to renting it for a while.

The company won’t charge you for data egress and ingress between your bare metal infrastructure and its cloud if both are in the same region. That last part of course is moot at the moment, since the service is only available in Northern Virginia.

ITPro Today™

Virtual Machine Advantages: 4 Reasons to Keep Loving VMs

Continue to consider virtual machine advantages when evaluating options such as serverless and containers.

Christopher Tozzi | Nov 25, 2019

We're living in the age of cloud-native everything, and any virtual machine advantages tend to get dismissed. VMs are increasingly seen as a legacy technology that lacks the versatility and performance benefits of newer solutions, like containers and serverless functions. If you're deploying an application today, you are probably more inclined to do it on one of the latter types of "next-generation" platforms, rather than using boring old VMs.

To a degree, that trend is fair. Virtual machines are a less effective solution in many cases than alternative forms of technology.

Yet, this does not mean that virtual machines have totally outlived their usefulness. Just as the bare-metal environments (which virtual machines helped turn into a "legacy" technology two decades ago) still have their use cases today, there remain several good reasons to consider using a virtual machine instead of a container, a server less function or some other new-fangled type of hosting solution.

The Case against Virtual Machines

To explain why, let me first summarize the reasons why virtual machines may *not* be an ideal choice, as compared to alternative hosting technologies.

The most common reasons to avoid virtual machines, and instead choose something like containers to host your app, are the following:

- **Overhead**: Virtual machines consume more resources than a container.
- **Speed**: In some respects, virtual machines are slower. They take longer to start (possibly a minute or two, compared to seconds for a container). The apps they host may also not run as quickly due to the fact that some of the host system's resources are tied up by the virtualization hypervisor, leaving fewer available for the app to use.
- **Redundancy**: Virtual machines were designed with the assumption that each machine would reside on a single server. While it's possible to create redundancy for virtual machines by spreading them across a cluster of servers, doing so requires more effort--and is a more clunky process--than distributing an app across a cluster using containers.
- **Bulky images**: A virtual machine image that contains a host operating system (as most do) will typically consume at least several gigabytes of space, and possibly much more. In comparison, a container image could be just a few megabytes, since a container image doesn't have to package a full operating system.
- **Cloud-native**: Virtual machines are a technology that became widespread decades ago, long before the cloud age. Because of that, there is something of a cultural bias against them, spurred by the

notion that--unlike containers and serverless--virtual machines are simply not a cloud-native technology.

All of these points are true and valid. For many modern application deployments, virtual machines are not the best choice.

Reasons Why Virtual Machines Still Matter

Yet, there are still plenty of use cases where virtual machines stand out from the competition in a positive way. Consider the following reasons why you might want to keep your virtual machines, and avoid the temptation to jump on the containerized, cloud-native bandwagon.

Flexibility

Perhaps the biggest selling point for virtual machines is that, at the end of the day, they continue to offer the greatest amount of deployment flexibility. A virtual machine can be deployed almost anywhere, regardless of the operating system or configuration of its host. A Windows system can host a Linux-based virtual machine, and vice versa.

Containers offer some level of flexibility. A containerized Linux application doesn't care which Linux distribution hosts it. But you

still can't run a Linux container or Windows, or a Windows container on Linux, unless you use a virtual machine to create the additional abstraction needed.

Security and Isolation

Container security has improved immensely since the debut of Docker in 2013. But it still leaves some cause for concern; indeed, security worries are a major reason why some teams opt not to use containers.

Some of those concerns may be assuaged as container platforms continue to mature and more security tools support them fully. However, it's unlikely that containers will ever totally match virtual machines from a security perspective, due to the simple fact that containerized applications can never be isolated from each other to the same extent as virtual machines. Virtual machines don't share each other's kernel or other essential system resources in the way containers do.

Management

While it's true that containers are a more natural fit in many respects for distributed host environments, this feature can also make them harder to manage. When you have hundreds of containers spread across dozens of servers, things quickly become difficult to control. That's why you use an orchestrator like Kubernetes, of course, which automates much of the management. However, the orchestrator itself adds another layer of complexity that you have to set up, manage and secure.

Large-scale virtual machine deployments require orchestration solutions, too. But they are rarely as complex to manage as a container deployment. With virtual machines, there are simply fewer moving parts, and fewer overlapping layers of infrastructure.

Virtual Machines Are the Original Cloud

Finally, let's address the cultural bias against virtual machines. Virtual machines may predate the cloud, but that doesn't mean they're foreign to it. IaaS services based on virtual machines were the first major cloud-computing service rolled out by public cloud providers like AWS in the mid-2000s. And they remain a key part of the offerings of those providers.

Containers and other so-called cloud-native solutions may be hotter these days. But don't make the mistake of assuming that virtual machines are somehow not also a cloud-native technology. Without virtual machines, the cloud would never have become a thing in the first place.

Conclusion

For many IT teams, it's high time to stop worrying and learn to love Docker (and whichever other modern, cloud-native application hosting technologies suit your liking). But that doesn't mean abandoning virtual machines entirely. Virtual machines still have an important role to play in many clouds, and it would be a mistake to write them off based on the simple assumption that they are an "old" technology.

•

Russian tycoon Oleg Deripaska, April 2019

ENERGY

Russia’s Largest Bitcoin Mine Turns Water Into Cash

Hydropower plant feeding Soviet-era aluminium smelter sells 100MW of excess power to Bitriver for crypto mining.

Bloomberg | Nov 25, 2019

Yuliya Fedorinova and Gem Atkinson (Bloomberg) -- Bitriver, the largest data center in the former Soviet Union, was opened just a year ago, but has already won clients from all over the world, including the U.S., Japan and China. Most of them mine bitcoins.

The company rents a building near the Bratsk aluminum plant. The world’s single largest aluminum smelter was built by the USSR in 1960s along with the nearby hydropower plant as energy is the largest cost in aluminum smelting.

On top of the power supply, another thing that makes Bratsk an ideal place for crypto is the Siberian climate with its long and cold winters. Low temperatures are good for the data center equipment.

Billionaire Oleg Deripaska's team came up with the idea of building the data center in Bratsk about five years ago. En+ Group Plc and its unit United Co. Rusal, which the sanctioned businessman used to control, own the Bratsk hydropower and aluminum plants.

While Russian law doesn't recognize crypto mining, Bitriver isn't engaged in mining itself and only provides equipment at the data center and technical services, meaning its business is legal.

Deripaska's companies spent nearly 10 months under sanctions before he reached an agreement with U.S. Treasury to cut his control. The penalties were lifted in January. Continuing sanctions on En+ could have caused troubles for the cryptocurrency miners.

En+ supplies up to 100 megawatts of power to Bitriver as a way to diversify its client base and sell excess energy. Cheap and stable power is also a key ingredient for crypto mining. En+ and Bitriver also have a venture that provides computer racks to crypto miners.

Google Throws Open Doors to Its Top-Secret Data Center

A central cooling plant in Google's Douglas County, Georgia, data center.
PHOTO: GOOGLE/CONNIE ZHOU

If you're looking for the beating heart of the digital age — a physical location where the scope, grandeur, and geekiness of the kingdom of bits become manifest—you could do a lot worse than Lenoir, North Carolina. This rural city of 18,000 was once rife with furniture factories. Now it's the home of a Google data center.

Engineering prowess famously catapulted the 14-year-old search giant into its place as one of the world's most successful, influential, and frighteningly powerful companies. Its constantly refined search algorithm changed the way we all access and even think about information. Its equally complex ad-auction platform is a perpetual money-minting machine. But other, less well-known engineering and strategic breakthroughs are arguably just as crucial to Google's success: its ability to build, organize, and operate a huge network of servers and fiber-optic cables with an efficiency and speed that rocks physics on its heels. Google has spread its infrastructure across a global archipelago of massive buildings—a dozen or so information palaces in locales as diverse as Council Bluffs, Iowa; St. Ghislain, Belgium; and soon Hong Kong and Singapore—where an unspecified but huge number of machines process and deliver the continuing chronicle of human experience.

This is what makes Google Google: its physical network, its thousands of fiber miles, and those many thousands of servers that, in aggregate, add up to the mother of all clouds. This multibillion-dollar infrastructure allows the company to index 20 billion web pages a day. To handle more than 3 billion daily search queries. To conduct millions of ad auctions in real time. To offer free email storage to 425 million Gmail users. To zip millions of YouTube videos to users every day. To deliver search results before the user has finished typing the query. In the near future, when Google releases the wearable computing platform called Glass, this infrastructure will power its visual search results.

The problem for would-be bards attempting to sing of these data centers has been that, because Google sees its network as the ultimate competitive advantage, only critical employees have been permitted even a peek inside, a prohibition that has most certainly included bards. Until now.

PHOTO: GOOGLE/CONNIE ZHOU

Here I am, in a huge white building in Lenoir, standing near a reinforced door with a party of Googlers, ready to become that rarest of species: an outsider who has been inside one of the company's data centers and seen the legendary server floor, referred to simply as "the floor." My visit is the latest evidence that Google is relaxing its black-box policy. My hosts include Joe Kava, who's in charge of building and maintaining Google's data centers, and his colleague Vitally

Gudanets, who populates the facilities with computers and makes sure they run smoothly.

A sign outside the floor dictates that no one can enter without hearing protection, either salmon-colored earplugs that dispensers spit out like trail mix or panda-bear earmuffs like the ones worn by airline ground crews. (The noise is a high-pitched thrum from fans that control airflow.) We grab the plugs. Kava holds his hand up to a security scanner and opens the heavy door. Then we slip into a thunder dome of data...

Urs Hölzle had never stepped into a data center before he was hired by Sergey Brin and Larry Page. A hirsute, soft-spoken Swiss, Hölzle was on leave as a computer science professor at UC Santa Barbara in February 1999 when his new employers took him to the Exodus server facility in Santa Clara. Exodus was a collocation site, or colo, where multiple companies rent floor space. Google's "cage" sat next to servers from eBay and other blue-chip Internet companies. But the search company's array was the most densely packed and chaotic. Brin and Page were looking to upgrade the system, which often took a full 3.5 seconds to deliver search results and tended to crash on Mondays. They brought Hölzle on to help drive the effort.
It wouldn't be easy. Exodus was "a huge mess," Hölzle later recalled. And the cramped hodgepodge would soon be strained even more. Google was not only processing millions of queries every week but also stepping up the frequency with which it indexed the web, gathering every bit of online information and putting it into a searchable format. AdWords—the service that invited advertisers to bid for

placement alongside search results relevant to their wares—involved computation-heavy processes that were just as demanding as search. Page had also become obsessed with speed, with delivering search results so quickly that it gave the illusion of mind reading, a trick that required even more servers and connections. And the faster Google delivered results, the more popular it became, creating an even greater burden. Meanwhile, the company was adding other applications, including a mail service that would require instant access to many petabytes of storage. Worse yet, the tech downturn that left many data centers under populated in the late '90s were ending, and Google's future leasing deals would become much more costly.

For Google to succeed, it would have to build and operate its own data centers—and figure out how to do it more cheaply and efficiently than anyone had before. The mission was codenamed Willpower. Its first built-from-scratch data center was in The Dallas, a city in Oregon near the Columbia River.

Hölzle and his team designed the $600 million facility in light of a radical insight: Server rooms did not have to be kept so cold. The machines throw off prodigious amounts of heat. Traditionally, data centers cool them off with giant computer room air conditioners, or CRACs, typically jammed under raised floors and cranked up to arctic levels. That requires massive amounts of energy; data centers consume up to 1.5 percent of all the electricity *in the world*.
>Data centers consume up to 1.5 percent of the entire world's electricity.

Google realized that the so-called cold aisle in front of the machines could be kept at a relatively balmy 80 degrees or so—workers could wear shorts and T-shirts instead of the standard sweaters. And the "hot aisle," a tightly enclosed space where the heat pours from the rear of the servers, could be allowed to hit around 120 degrees. That heat could be absorbed by coils filled with water, which would then be pumped out of the building and cooled before being circulated back inside. Add that to the long list of Google's accomplishments: The Company broke its CRAC habit.

Google also figured out money-saving ways to cool that water. Many data centers relied on energy-gobbling chillers, but Google's big data centers usually employ giant towers where the hot water trickles down through the equivalent of vast radiators, some of it evaporating and the remainder attaining room temperature or lower by the time it reaches the bottom. In its Belgium facility, Google uses recycled industrial canal water for the cooling; in Finland it uses seawater.

The company's analysis of electrical flow unearthed another source of waste: the bulky uninterrupted-power-supply systems that protected servers from power disruptions in most data centers. Not only did they leak electricity, they also required their own cooling systems. But because Google designed the racks on which it placed its machines, it could make space for backup batteries next to each server, doing away with the big UPS units altogether. According to Joe Kava, that scheme reduced electricity loss by about 15 percent.

All of these innovations helped Google achieve unprecedented energy savings. The standard measurement of data center efficiency is called power usage effectiveness, or PUE. A perfect number is 1.0, meaning all the power drawn by the facility is put to use. Experts considered 2.0—indicating half the power is wasted—to be a reasonable number for a data center. Google was getting an unprecedented 1.2.

IMAGES OF DATA CENTER OF Google

New Google Cloud Data Center Being...

Google Looking Outside Berkeley County...

Asian investors embrace data centers to...

Building Danish data center...

Data Centers - Google

NSA, UK hacked Yahoo! and Google data...

Google Spent $7.3 Billion on its Data...

Reporter into one of its data centers...

Data center – Data Centers – Google

Never-Before-Seen Images...

Deep Mind AI Reduces Google Data Centre...

Rows and rows of server infrastructure...

Google uses AI to cut data centre...

$600 million data center in Becker...

Building new Singapore data center...

Data centers go green thanks to Italian...

Google to Invest $3.3 Billion to Grow...

Google plans $600 million expansion at...

Google Adds More Cloud Data Centers...

TPU Chip Helped It Avoid Building...

A look inside Google's data centers

Google Cloud Data Center Locations - Kinta

Data Centers - Google

Google Data Center Locations

Seeking a green data center? Facebook ...

A tour of Google's top-secret data...

$600mn Alabama data center...

Google Building Four-Story Data Centers

Google Berkeley County Data Center...

Google to invest $3.3 billion in...

Super-Efficient Belgium Data Center...

Ever wondered what a Google data centre...

Electricity Usage of Its Data Centers...

Google expanding Michigan offices...

Google uses self-aware datacenters to...

Google's Amazing Floating Data Centers...

Google data centers - Wikipedia

Google pours $600 million into Oklahoma...

Data Center In Belgium

How Google is Using AI for Data Center...

Google invests €1 billion in data...

Apple lead on green data centers...

Cloud Giants' 2019 Data Center...

Google talks efficient data centers | ZDNet

Throws some weight behind on-premises...

data center push for renewable energy ...

Data Center in Nevada

Google Buys More Land in Europe for...

Google Releases Never-Before-Seen...

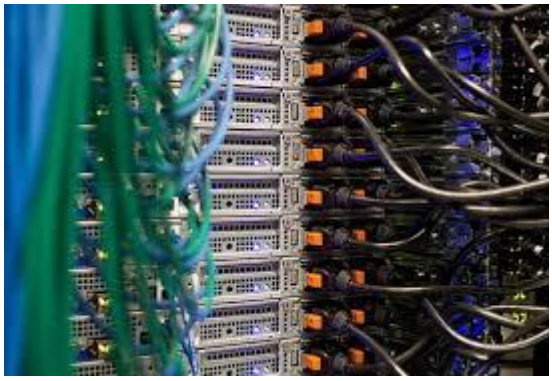
using AI to keep its data centers cool ...

Google to set up data center in Korea...

Pushing ahead with European expansion...

Google CEO announces $3.3B data center...

Parcels of land to build data...



Data Centers In Loudoun County

SG$404 mln in SG-based data centre...

Google data center site taking shape in...

Google Data Center Uses No Chillers...

Google opens regional data center in...

Invest $13B in US data centers...

Google invests $3 billion in European...

150m expanding Dublin data centre

Mesa City Council OKs deal for $1B...

Data center in Belgium...

Google on-premises Kubernetes solution...

Google data centre and green energy...

A look inside Google's data centers

Google to build €600mn data centre in...

Google's secret water-based data center...

Google plans to spend $600 million...

Data Center In New Albany...

Data center in Finland upgraded...

Google NC data center gets Street View...

Data Centers...

Google to purchase 413MW of solar power...

Google to invest USD 600 million in...

Google considers $600M data center in...

Belgium Solar | Google Sustainability

It’s Data Center Security Measures

Google gives Artificial Intelligence...

Google Datacenter: Not Everything Is As...

10 Easter eggs from inside a Google...

Better, bidder, bigger SIEM coming you’re ...

Google lost 0.000001% of users' data...

Why Data Centers Collect Big Tax Breaks...

Guards Company's Data Center Secrets...

DoJ: Google no longer challenges most...

Data Centers...

Hyper-Scale Operators Lower Data Center...

Google Unwrapped - Columbia Insight

Google Data Center Locations

Google to break ground on $600 million...

Google to Spend $13B on US Data Center...

Google dolls up its data centers with a...

Xcel partner on $600M data center...

2nd Google Green Powered Data Center

Google, Microsoft, Facebook, Apple, and...

Google Picks Mesa, Ariz., for Location...

The Agenda: Google pumping $600M into...

Google leaves data center servers in...

Data centre hosting caches for Google...

Charleston area data center...

Opens window into secretive data centers

RTL Today - Google in Luxembourg...

Data centers: How Google helps regions grow

Google growing its Oklahoma data center

Google Data Center under Construction...

Invest $13B in US data centers...

Google Cloud Platform Blog: A look...

Map of all Google data center locations...

Data Center in North Texas...

Reveals secretive data centers...

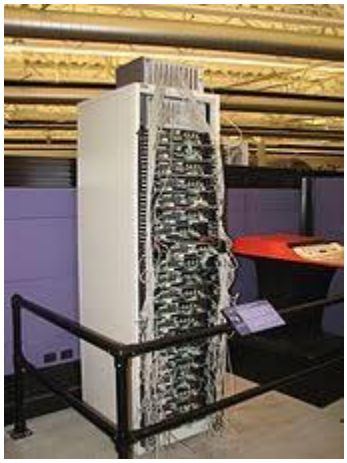
invested over $10bn in data centers ...

Google to invest in new data centers...

Spend $13B on new data centers...

Seeking a green data center? Facebook ...

India 2019 List Google Data Center



Google data centers - Wikipedia

Google uses AI to cool down its servers...

Google unveils plans to spend $600M for...

New Google Cloud Data Center Being...

Google Looking Outside Berkeley County...

Asian investors embrace data centers to...

Building Danish data center...

Data Centers - Google

NSA, UK hacked Yahoo! and Google data...

Google Spent $7.3 Billion on its Data...

Reporter into one of its data centers...

Data center – Data Centers – Google

Never-Before-Seen Images...

DeepMind AI Reduces Google Data Centre...

Rows and rows of server infrastructure...

Google uses AI to cut data centre...

$600 million data center in Becker...

Building new Singapore data center...

Data centers go green thanks to Italian...

Google to Invest $3.3 Billion to Grow...

Google plans $600 million expansion at ...

oklahoman.com

Google Adds More Cloud Data Centers...

TPU Chip Helped It Avoid Building...

A look inside Google's data centers

Google Cloud Data Center Locations - Kinsta

Data Centers - Google

Google Data Center Locations

baxtel.com

Seeking a green data center? Facebook ...

A tour of Google's top-secret data...

$600mn Alabama data center...

Google Building Four-Story Data Centers

Google Berkeley County Data Center...

Google to invest $3.3 billion in...

Super-Efficient Belgium Data Center...

Ever wondered what a Google data centre...

Electricity Usage of Its Data Centers...

Google expanding Michigan offices...

Google uses self-aware datacenters to...

Google's Amazing Floating Data Centers...

Google data centers - Wikipedia

Google pours $600 million into Oklahoma...

Data Center In Belgium

How Google is Using AI for Data Center...

Google invests €1 billion in data...

Apple lead on green data centers...

Cloud Giants' 2019 Data Center...

Google talks efficient data centers | ZDNet

Throws some weight behind on-premises...

Data center push for renewable energy...

Data Center in Nevada

Google Buys More Land in Europe for...

Google Releases Never-Before-Seen...

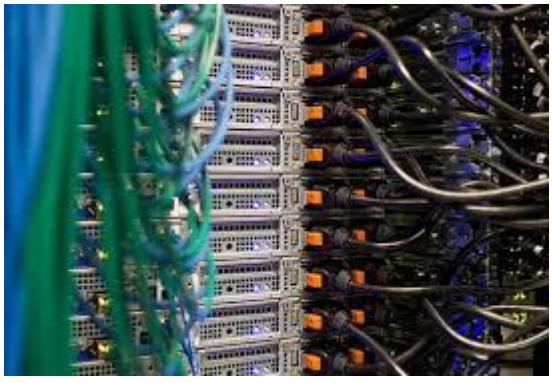
Using AI to keep its data centers cool...

Google to set up data center in Korea...

Pushing ahead with European expansion...

Google CEO announces $3.3B data center...

Parcels of land to build data...



Data Centers In Loudoun County

SG$404 mln in SG-based data centre...

Google data center site taking shape in...

Google Data Center Uses No Chillers...

Google opens regional data center in...

Invest $13B in US data centers...

Google invests $3 billion in European...

150m expanding Dublin data centre

Mesa City Council OKs deal for $1B...

Data center in Belgium...

Google on-premises Kubernetes solution...

Google data centre and green energy...

A look inside Google's data centers

Google to build €600mn data centre in...

Google's secret water-based data center...

Google plans to spend $600 million...

Data Center In New Albany...

Data center in Finland upgraded...

Google NC data center gets Street View...

Data Centers...

Google to purchase 413MW of solar power...

Google to invest USD 600 million in...

Google considers $600M data center in...

Belgium Solar | Google Sustainability

It’s Data Center Security Measures

Google gives Artificial Intelligence...

Google Datacenter: Not Everything Is As...

10 Easter eggs from inside a Google...

Better, bidder, bigger SIEM coming your ...

Google lost 0.000001% of users' data...

Why Data Centers Collect Big Tax Breaks...

Guards Company's Data Center Secrets ...

huffpost.com

DoJ: Google no longer challenges most...

Data Centers...

Hyper-Scale Operators Lower Data Center...

Google Unwrapped - Columbia Insight

Google Data Center Locations

Google to break ground on $600 million...

Google to Spend $13B on US Data Center...

Google dolls up its data centers with a...

Xcel partner on $600M data center...

2nd Google Green Powered Data Center

Google, Microsoft, Facebook, Apple, and...

Google Picks Mesa, Ariz., for Location...

The Agenda: Google pumping $600M into...

Google leaves data center servers in...

Data centre hosting caches for Google...

Charleston area data center...

Opens window into secretive data centers

RTL Today - Google in Luxembourg...

Data centers: How Google helps regions grow

Google growing its Oklahoma data center

Google Data Center under Construction...

Invest $13B in US data centers...

Google Cloud Platform Blog: A look...

Map of all Google data center locations...

Data Center in North Texas...

Reveals secretive data centers...

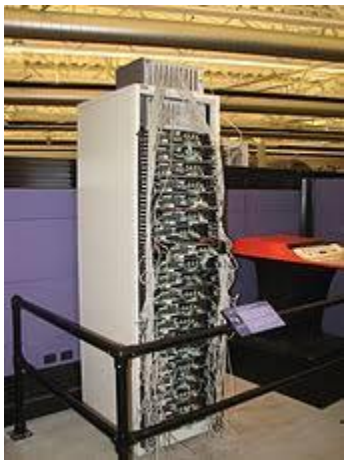
Invested over $10bn in data centers...

Google to invest in new data centers...

Spend $13B on new data centers...

Seeking a green data center? Facebook ...

India 2019 List Google Data Center



Google data centers - Wikipedia

Google uses AI to cool down its servers...

Google unveils plans to spend $600M for...

Google's Data Center Switching Fabric...

New Google data centre to employ 1,700...

Google Data Center Tour- In Containers?

The Data Center Mural Project

Google data center in Eemshaven

Google to invest €1 billion...

mega data center in Finland ...

Google plans to open second data center...

Google is opening new offices in Austin...

600m to expand its Finland data...

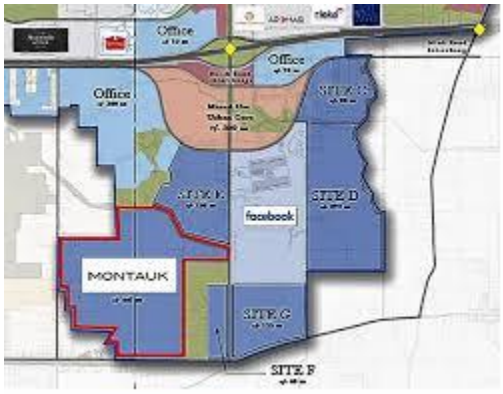
Google building $600M data center in...

Google to break ground on $600 million...

Data Center in Google cloud Platform...

Important Difference between...

NSA Targeting of International Data Centers



go ahead with $600 million data center ...

Google Data Center Wallpaper · ①...

Collocation Data Center Company Cologix ...

Data center takes shape in Ellis...

Google has a data storage solution...

Google streams data center pods to...

Invests $670 Million to Expand it’s...

Google to invest $1B in data center

The Future Of Data Centers - CB...

Data center energy...

Floating data center plan hold water...

Google Doubles Down on Sea-Cooled Data...

Google Turns To Partners In Cloud Tangle

Coal Power Plant to Data Center...

Releases the state of the data center...

Mind-blowing facts about Google...

Google's Shipping Container Server Storage

Google kicks off construction on $600M...

Economic Impact of Google Data Centers...

Google search reveals New Albany best...

Building $600 million data center in Ohio

Peeping inside Google's data centers...

Data Center in Belgium has a QR code...

Data center runs at 95 degrees...

Google building a floating data center...

Second 'floating Google data center ...

Seven years of its data center history...

Data center in Salt Lake City in 2020...

1430 Veterans Memorial Hwy Omaha

Carbon-intensive its data...

Here's How Google is Approaching AI...

Google Data Center Wallpaper (63+ images)

Google opens up new data center...

$500 million Midlothian data center...

Another AWS Server Farm (Amazon Data ...

Data center competition heats up in...

Environmental Impact of the Data-Center...

European Data Center Expansion

Google data centre construction site...

Google will spend $3.3B to expand its ...

Google uses AI to save on electricity ...

Google invests €600mn in data centre in...

Google to invest €600m in new data...

Data Center like a Google SRE...

Google Data Center Project in Uruguay...

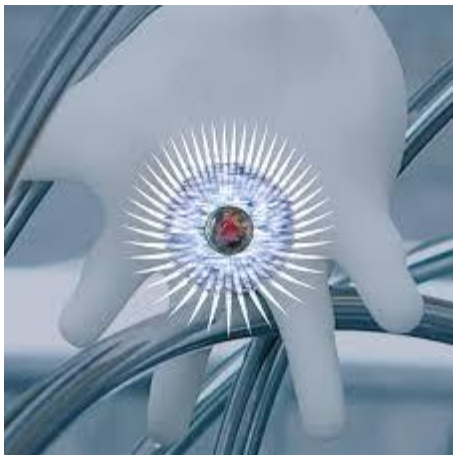
Google Data Center Pictures: Year in Review

Google to build data center in Henderson

reveal 340 illegal construction workers ...

Google Data Centers: Economic Impact...

data center cooling systems ...



Live @ Google Data Center | Sam Kiel

releases the state of the data center ...

Google Data Center - The Dales, Oregon..

Amazon, Google, Microsoft Cloud Data...

roots where Google's European data ...

Google Data Center Locations

Google, Facebook, Microsoft, and...

PPA for its new data center in Denmark...

Land in Europe for Data Center Construction

Google invests €1 billion in data...

Guided tour of a Google data center...

Hyper scale data centre number reaches ...

Google cloud services to store cloud...

Google Data Center Wallpaper ·①...

Google joins New Albany high-tech crowd...

Google to invest 3 billion Euros into...

Google affiliate's $600M kc data center...

Inside a Google data center – tam cats

Business Finland - Google to invest EUR...

Google to invest $13 billion in data...

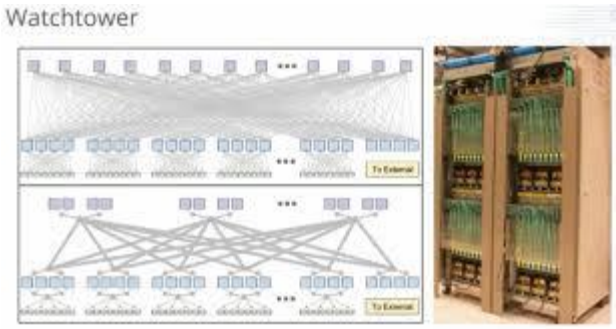
Google to invest $13B in new US offices...

Google Data Center Png & Free Google...

Google Cloud Storage Security gets an...

Google to invest $700 million in Danish...

Google rebrands cloud services 'Google ...



Google Homegrown Datacenter Networks

Google data center deals questioned in...

Oregon Data Center Opened By Google

Google solves barge mystery: 'A ...

Economic impact of data center in Pryor...

Google Building Four-Story Data Centers

Google Cloud Tests Devices for Private...

GOOGLE DATA CENTER 2079, beeple ...

Google container data center tour - YouTube

Data Center Security Uses Storm troopers...

Google shacks up with Equinox for cloud...

Google to Invest $13 Billion in New US...

Google Data Center in Finland...

Los Angeles Data Center for Curie...

Data centers will blow...

Plans $1B expansion in Council Bluffs...

Finnish data center – Beyondtech

Google to set up data center in Tainan...

Google data center in Ellis...

Release Specs for IBM Power9-Based Server

Google flushes heat from data center...

Is Google's Renewable Energy Plan What...

Google Data Center from the inside

Google's data center is a game changer..

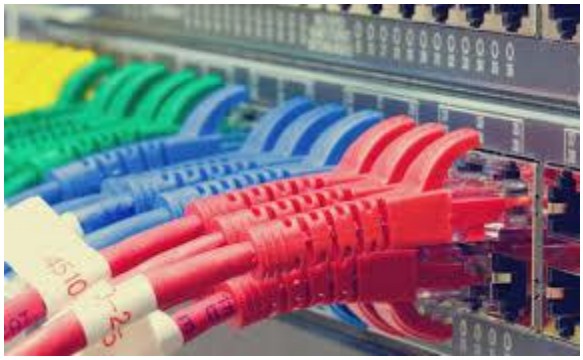
European Data Center Headlines...

Google's Photo Tour of Their Sprawling...

The Data Center Mural Project

Data center location...

Google building $600M data center in...



Google score in Greenpeace data center...

Google's Chief Sustainability Officer...

AI the reins over its data center...

Google to launch first data centre in...

Google data centre in Jurong ...

Global geography of Google Data Centers...

Google Invests $670 Million to Expand...

Technology/Google Data Center (750x1334 ...

Silly debunk: Google Mars Data Center...

Inside A Google Data Centre

Sam Kidel hosts online sound exhibition...

data center in Hamina, Finland ...

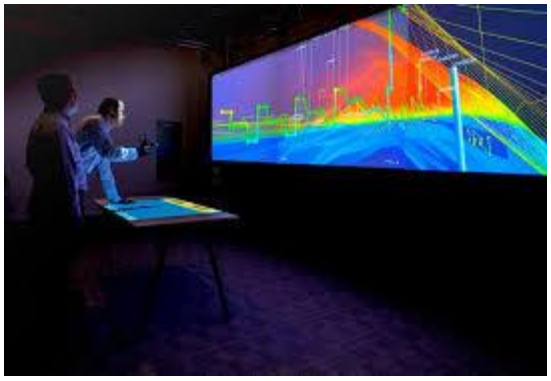
Google to invest $140 million to expand...

data center. A closer look...

Google Datacenter Eemshaven – Minerals

Google data center site taking shape in...

Explore a Google data center with...

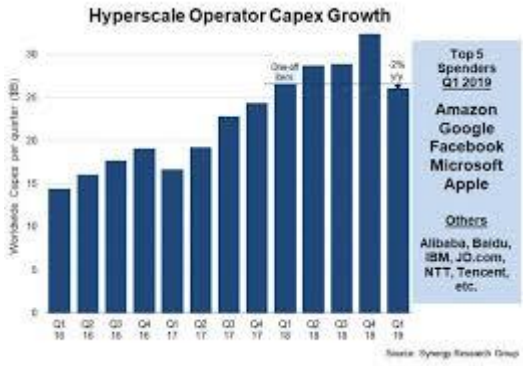


Google And Saudi Armco May Build Data...

Google's new data centers in the U.S...

Google Berkeley County Data Center...

Robocut studio | Google Data Center...



Data Center Spending Pauses In Q1 2019

Expand its Charleston-area data center...

Mesa OKs deal to bring $1B Google site...

How Google is extending data center...

An Artist and a Helicopter Capture...

Google boosts Taiwan data center plans...

Electricity Usage of Its Data Centers...

Inside Google's huge Iowa data centers...

Midlothian is worth a Google: Tech...

Solar energy project in Taiwan...

Microsoft now has one million servers...

Google invests €3 billion to expand...

Building $600M data center in Becker...

Google Data Center Wallpaper (63+ images)

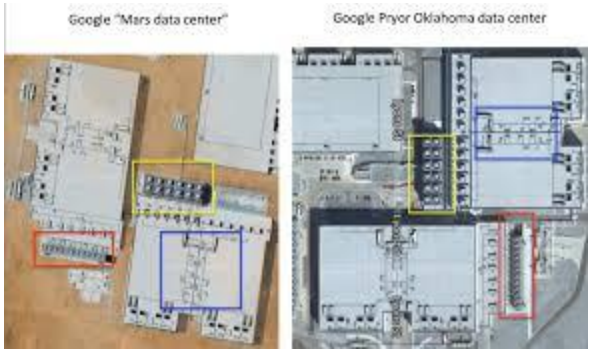
AWS Still Leads in Cloud Infrastructure...

Google Maps to find Google data centers...

Take a 360-degree tour inside gigantic...

Google data centers – how do they do it...

Data center takes shape in Ellis...



Silly debunk: Google Mars Data Center...

Data Centers Give Off a Blade-Runner Vibe

Google Maps to find Google data centers...

Google Data Center Wallpaper 63 Images...

Take a 360-degree tour inside gigantic...

Google data centers – how do they do it...

Data center takes shape in Ellis...

Silly debunk: Google Mars Data Center...

Data Centers Give Off a Blade-Runner Vibe

Google data centers have evolved, set...

A Tour through a Google Data Center...

Google lines up Netherlands data centre...

Green light for second Dublin data centre

Building a new data center in Singapore

Google reveals TAPE-TASTIC data centre...

Google Releases Never-Before-Seen...

Data Centers without Humans?

Google to build data centre in Belgium...

Google | Data Center | Community Partner

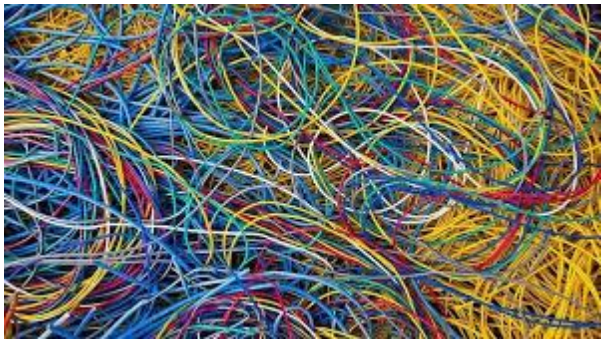
Google data center site taking shape in...

Google Datacenter near Delfzijl in the...

Data Centers - Google

Data Centers in Iowa: Are the...

Google Announces Dutch Data Centre...



top secret' data centers ...

700 millions dams un data center...

Data Centers with a Cool Mural...

Visual tour: Inside Google's data...

Data Centers on Street View...

Google Opens Data Centers...

Most creative data...

Google is beautifying its data centers...

go ahead with $600 million data center ...

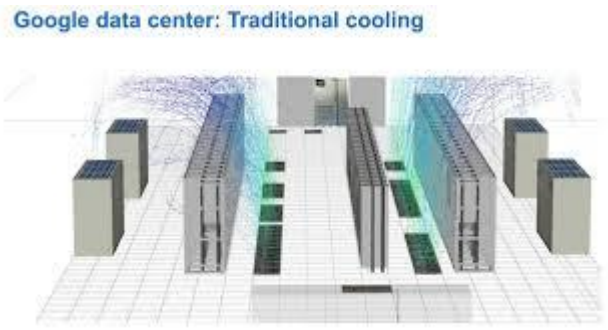
Google's Data Center Cooling Determined...

Google Cancels Hong Kong Data Center Plans

Google Datacenter, Eemshaven.Google ...

Google data center network to include...

How Big is Google Data Centre? |



Adopts Water as key to save Energy...

Why Snap's IPO Means More Data Centers...

AI the reins over its data center...

Google to launch first data centre in...

Alligator Patrols Google's Data Center...

Google Cloud to offer VMware data...

Google plans second Chilean data center...

Google – Pryor, Oklahoma: World's Top...

Data centers in Eastern Oregon...

Google Data center How they protect our ...

data center footprint across USA ...

30 billion watts and rising: balancing...

Google building $600M data center in...

Google Data Center Wallpaper ·①...

Buying a Crazy Amount of Clean Energy...

Is your Data Center Being Challenged to...

Google feeds net giants India and China...

Why, Unless You're Running with Amazon...

Photos: How Apple, Google, Facebook and...

Inside A Google Data Center GIF | Gfycat

Google to open a €600m data center in...

Google spotlights data center inner...

Google building a third data center in...

Google's secret water-based data center...

Helicopter Capture Google's Off-Limits...

How AI is going to transform the Data...

Google 2018 Environmental Web Report...

4K HD Desktop Wallpaper for 4K Ultra...

Google data center spies on IT...

Belgian data center...

Modular data center - Wikipedia

Google Midlothian TX Data Center

Why Google technicians wear shorts...

mind-blowing facts about Google ...

Applications of Artificial Intelligence...

Microsoft now has one million servers...

Google Data Center Wallpaper (63+ images)

UK justifies wholesale Google and...

Google is beautifying its data centers...

Google plans $600 million investment in...

Computer sketch, Art...

Google Releases Never-Before-Seen...

Iowa's Data Center Cluster Keeps Growing

Google to Build Three Data Centers in...

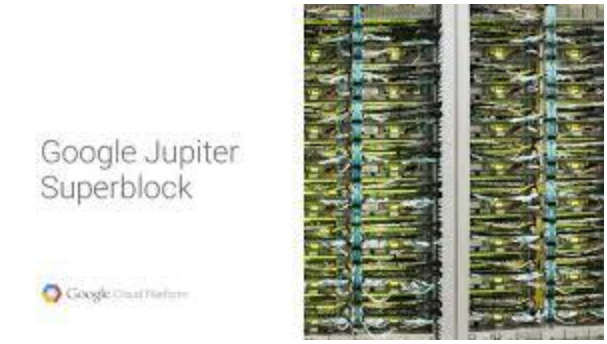
Google building $600M data center in...

Extreme targets data center automation...

Data Center With Toilet Water...

Google builds out its data center...

Google chooses Midlothian for new Data...



Pulls Back Curtain On Its Data Center...

Google announces $13 billion investment...

Second 'floating Google data center ...

Project Wizard: Sarpy County may get...

Important Difference between...

Apple Denmark Aabenraa Kassø Data Center

Google spotlights data center inner...

Deed for data center land...

Google Data Center Wallpaper (63+ images)

Future Data Centers: Distributed, Grid...

Utah Considers Cutting off Water to the...

30 billion watts and rising: balancing...

Google Releases Never-Before-Seen...

Microsoft Azure Data Center

Azure global infrastructure

Achieve global reach and the local presence you need

Go beyond the limits of your on-premises datacenter using the scalable, trusted, and reliable Microsoft Cloud. Transform your business and reduce costs with an energy-efficient infrastructure spanning more than 100 highly secure facilities worldwide, linked by one of the largest networks on earth.

Deliver services confidently with a cloud you can trust

Scale globally

Reach more locations, faster, with the performance and reliability of a vast global infrastructure.

Safeguard data

Rely on industry-leading data security in the region and across our network.

Promote sustainability

Help build a clean-energy future and accelerate progress toward your sustainability goals.

Gain global reach while retaining local presence

Reduce the cost, the time, and the complexity of operating a global infrastructure while meeting your local data residency needs. Azure lets you deliver services at the scale you need to reach your customers and partners, wherever they are.

Let Azure keep your data secure

Azure safeguards data in facilities that are protected by industry-leading physical security systems and are compliant with a comprehensive portfolio of standards and regulations.

Join a sustainable IT future

Microsoft is committed to driving innovation that helps people and organizations improve the environment. We achieved carbon neutrality in 2012 and have met our goal of averaging 1.125 power usage effectiveness (PUE) for any new datacenter—exceeding the industry average by more than 30 percent.

Join us in reducing energy use and transitioning to a more carbon-neutral grid through projects and business models that harness next-generation technologies.

Sovereign regions

Azure Government

Azure Government regions provide the exclusivity required for US federal, state, local, and tribal governments and their partners to dedicated instances operated by screened US citizens. Azure Government offers the most certifications of any other cloud provider to simplify critical compliance requirements. Azure compliance services include FedRAMP and DoD compliance certifications, CJIS state-level agreements, the ability to issue HIPAA Business Associate Agreements, and support for IRS 1075.

Regions

A region is a set of datacenters deployed within a latency-defined perimeter and connected through a dedicated regional low-latency network.

With more global regions than any other cloud provider, Azure gives customers the flexibility to deploy applications where they need to. Azure is generally available in 52 regions around the world, with plans announced for 2 additional regions.

Geographies

Geography is a discrete market, typically containing two or more regions, that preserves data residency and compliance boundaries.

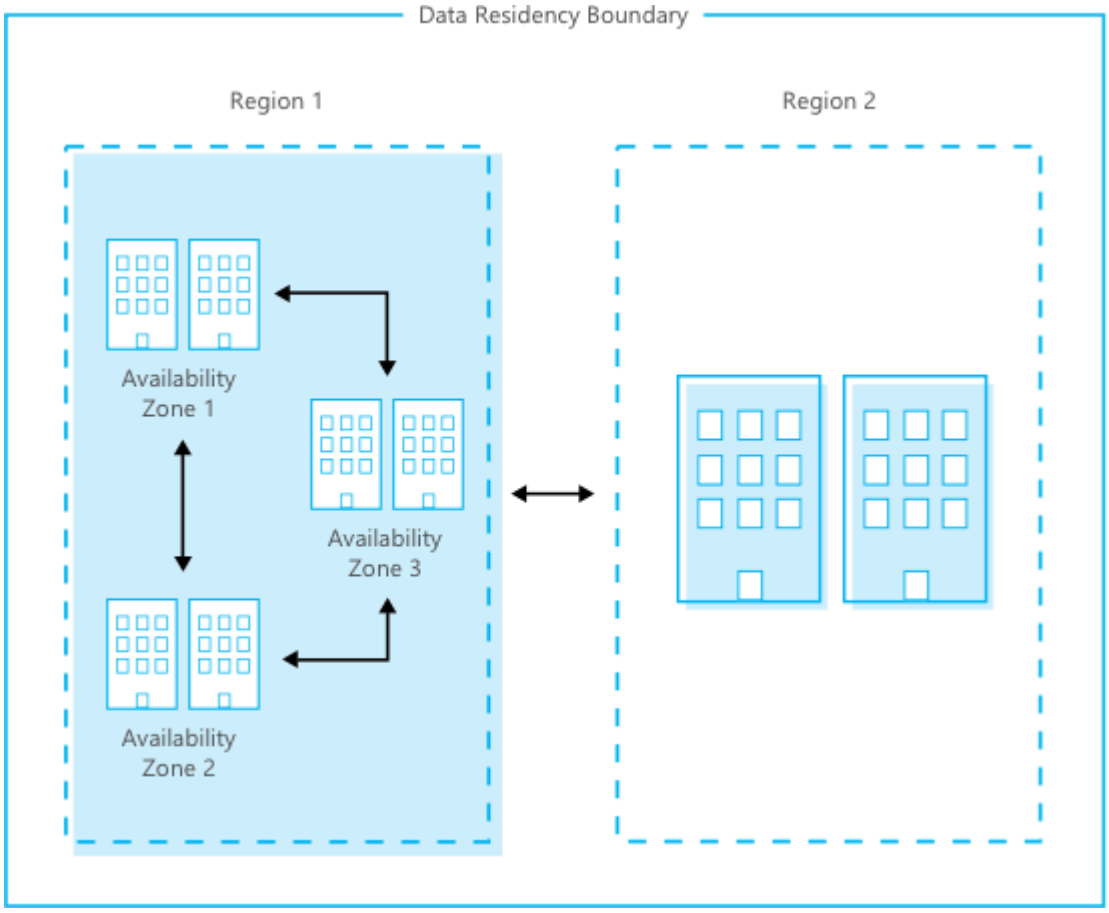
Geographies allow customers with specific data-residency and compliance needs to keep their data

and applications close. Geographies are fault-tolerant to withstand complete region failure through their connection to our dedicated high-capacity networking infrastructure.

Availability Zones

Availability Zones are physically separate locations within an Azure region. Each Availability Zone is made up of one or more datacenters equipped with independent power, cooling, and networking.

Availability Zones allow customers to run mission-critical applications with high availability and low-latency replication.



Achieve full resiliency with data residency.

Pair region and Availability Zones within the same data residency boundary for high availability, disaster recovery, and backup.

Resiliency to build with confidence

Azure regions, geographies, and Availability Zones form the foundation of our global infrastructure—providing customers high availability, disaster recovery, and backup.

Resiliency to build with confidence

Azure regions, geographies, and Availability Zones form the foundation of our global infrastructure—providing customers high availability, disaster recovery, and backup.

Which Azure region is right for you?

Azure China

Azure is now available in China through a unique partnership between Microsoft and 21Vianet, one of the country's largest internet providers. Operated and sold by 21Vianet,

Azure brings reliable, flexible, and cost-effective cloud computing technologies to customers there, while propelling technological development forward. Microsoft is the first global company to make public cloud services available in China.

Azure regions

Azure Regions

Azure has more global regions than any other cloud provider—offering the scale needed to bring applications closer to users around the world, preserving data residency, and offering comprehensive compliance and resiliency options for customers.

54

Regions
worldwide

140

Available in
140 countries

Which Azure region is right for you?

Consider these factors when choosing your region.

Compliance and data residency

Do you or your customers have specific compliance or data-residency requirements?

Service availability

Are the Azure services you want available in the region you're considering?

Pricing

Is cost an important factor in your decision making?

Azure solutions

Find the solution to meet the needs of your application or business

Solve your business problems with proven combinations of Azure services and related products. Whether you're just beginning in the cloud, or have years of experience developing cloud-based applications, we'll help you get started with sample architectures, documentation, and partner resources.

174. **Internet of Things**

Power your digital transformation, collect untapped data, and find new insights by connecting your devices, assets, and sensors

175. **AI platform**

AI productivity for every developer and every scenario

176. **SAP on Azure**

Bring cloud scale and agility to your mission-critical SAP workloads

177. Block chain

Quickly develop and deploy distributed apps on the block chain of your choice

178. Hybrid cloud applications

Maximize productivity by empowering developers to build and deploy applications the same way, whether your app runs on Azure or Azure Stack

179. Oracle on Azure

Run your Oracle® database and enterprise applications on Azure

180. DevOps

Bring together people, processes, and products to enable continuous delivery of value to your customers and coworkers

181. Mobile

Reach your customers everywhere, on every device, with a single mobile app build

182. E-commerce

Give customers what they want with a personalized, scalable, and secure shopping experience

183. Azure governance

Ensure compliance using the cloud governance capabilities built into Azure

184. Confidential computing

Protect your data and code while it's in use in the cloud

185. Dynamics on Azure

Fuel business growth by bringing together enterprise resource planning (ERP) and cloud services

186. LOB applications

Modernize your internal line of business (LOB) apps to meet today's IT challenges

187. Development and test

Simplify and speed up the process of building and testing applications across every platform

188. Business intelligence

Drive better, faster decision making by analyzing your data for deeper insights

189. Big data and analytics

Make the most informed decision possible by analyzing all the data you need in real time

190. **Modern data warehouse**

Handle exponential data growth without leaving security, scalability, or analytics behind

191. **Business SaaS apps**

Use business insights and intelligence from Azure to build software as a service (SaaS) apps

192. **Backup and archive**

Protect your data and applications no matter where they reside to avoid costly business interruptions

193. **Disaster recovery**

Protect all your major IT systems while ensuring apps work when you need them most

194. **Digital marketing**

Connect with customers worldwide with digital campaigns that are personalized and scalable

195. **Digital media**

Deliver high-quality videos to your customers anywhere, anytime, on any device

196. **High-performance computing**

Tap into unlimited resources to scale your high-performance computing (HPC) jobs

197. Micro service applications

Deliver scalable, reliable applications faster to meet the ever-changing demands of your customers

198. Gaming

Build, quickly launch, and reliably scale your games across platforms, and refine based on analytics

199. Server less computing

Build apps faster, focusing on innovation instead of infrastructure management

"Using Azure frees us to focus on where we add value—developing great software and delivering great customer experiences."

Mike Mellor: Senior Director, Technical Operations, Adobe

Industries

Find answers to the business challenges you face with an Azure solution that brings together everything you need—related products, services, and third-party applications.

Government

Build secure solutions that help government agencies improve how they serve and protect their citizens.

Financial services

Reimaging the customer experience, empower employees, optimize risk management, and transform your product's mission.

Retail

Deliver personal, seamless, and differentiated experiences with data-based customer insights.

Manufacturing

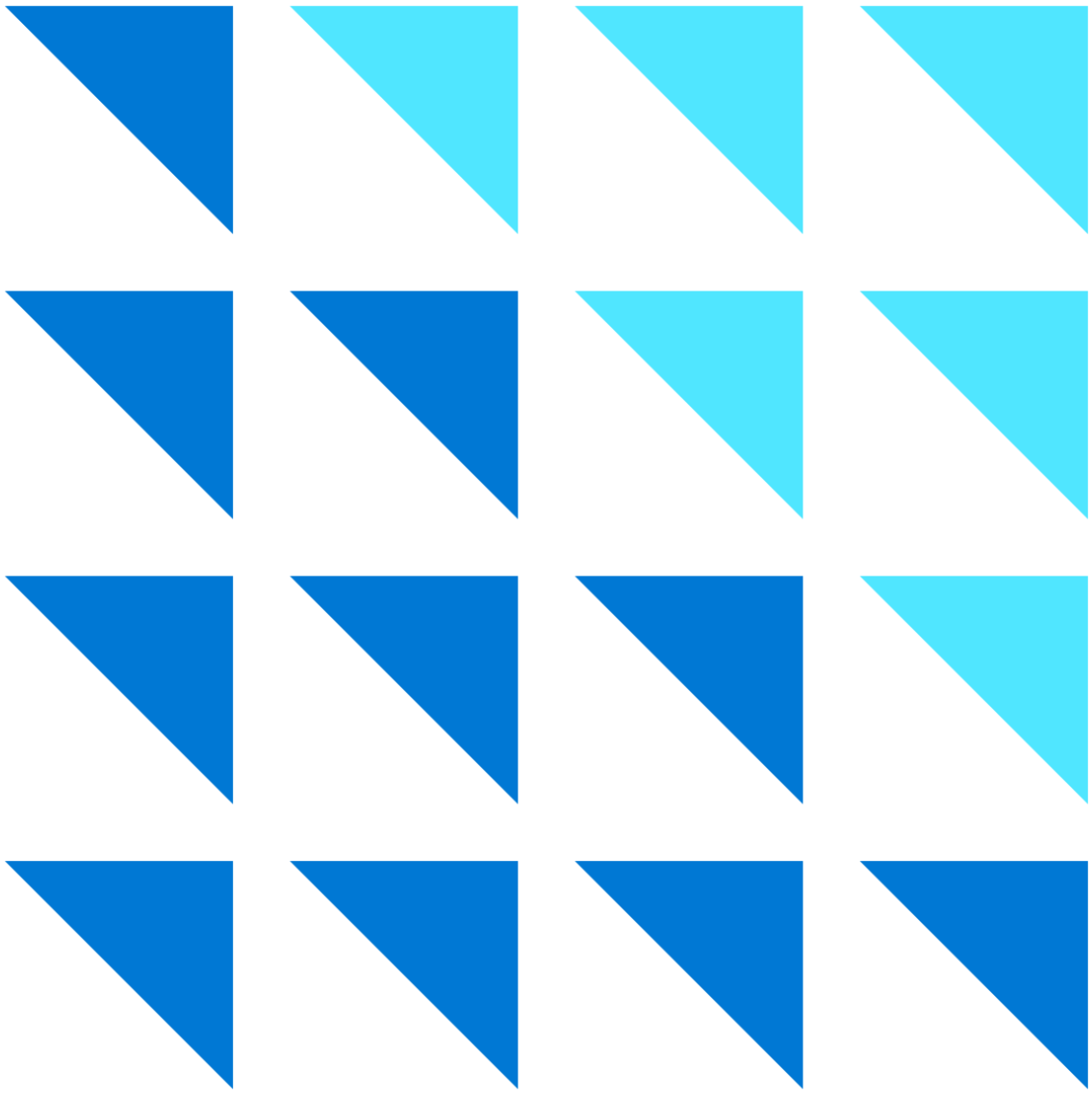
Gain actionable insights and respond quickly to customer feedback and market trends.

Health and life sciences

Develop solutions to improve health-based outcomes and drive efficiency, while harnessing technology innovation.

Energy

Develop secure solutions that optimize your digital operations, empower your workforce, deliver customer outcomes and create a sustainable future.



Migrate your workloads to Azure

Make your cloud migration faster and easier with the Azure migration center. Find proven tools and processes that guide you in migrating your on-premises applications, workloads, and databases to the cloud.

Introduction to Cloud Computing

Lower cost

Cloud computing is when you access computing services—like servers, storage, networking, software—over the internet ("the cloud") from a provider like Azure. For example, instead of storing personal documents and photos on your personal computer's hard drive, most people now store them online: that's cloud computing.

Cloud computing platforms, like Azure, tend to be less expensive and more secure, reliable, and flexible than on-premises servers. With the cloud, equipment downtime due to maintenance, theft, or damage is almost non-existent. You can scale your compute and storage resources—up or down—almost instantly when your needs change on Azure. Also, you typically pay only for the services you use, which provides a level of convenience and cost-control that's almost impossible to achieve with on-site infrastructure.

LoweLower cost

Cloud computing eliminates the expense of setting up and running on-site datacenters, which often have added costs such as employing staff and buying and maintaining land, buildings, and computer hardware. The cloud allows businesses to access the computer resources they need in real time to match their business needs on demand.

Higher securitygher security

Security is a key focus of cloud providers, who invest huge sums of money into securing their infrastructure. Cloud providers typically also offer a broad set of policies, compliance, technologies, and controls that strengthen your security posture by protecting your data, apps, and infrastructure from threats.

Increased productivityeased

More efficiently develop and manage your applications with nearly unlimited cloud computing resources. Cloud providers continuously update their datacenter networks with the latest-generation hardware, providing you with fast, efficient computing resources that never go obsolete and would be more costly to implement in a single datacenter.

Global scalebal scale

Cloud computing runs on data centers around the world, providing overall resiliency and reliability by allowing your data to be backed up in more than one geographic location. This also allows your IT resources to be delivered from specific geographic locations when required.

What is Azure?

Azure is an ever-expanding set of cloud computing services to help your organization meet its business challenges. Azure gives you the freedom to build, manage, and deploy applications on a massive, global network using your preferred tools and frameworks. With Azure, you can:

Be future

Be future-ready

Continuous innovation from Microsoft supports your development today—and your product visions for tomorrow. Build on the latest advancements in the cloud, including more than 1,000 new capabilities released in the last year.

Operate hybrid seamlessly

On-premises, in the cloud, and at the edge—Azure meets you where you are. Integrate and manage your environments with tools and services designed for hybrid cloud.

Build on your termsld on your terms

You have choices—with Azure's commitment to open source and support for all languages and frameworks, you're free to build how you want and deploy where you want.

Trust your cloud

Get security from the ground up—backed by a team of experts and proactive, industry-leading compliance that's trusted by enterprises, governments, and startups.

Azur is secure

Security is a given in the cloud industry, but Azure's proactive approach to security, compliance, and privacy is unique. Microsoft leads the industry in establishing and consistently meeting clear security and privacy requirements.

Between its industry-leading compliance and privacy certifications to built-in security controls and unique threat intelligence, Azure has everything you need to identify and protect against rapidly evolving threats.

Start with a secure foundation

Reduce costs and complexities with a highly-secure cloud foundation that takes advantage of multi-layered security provided by Microsoft.

Streamline your compliance and enable business transformation

Use built-in controls, configuration management tools, implementation and guidance resources, and third-party audit reports to simplify your compliance needs.

Detect threats earlythreats early

Identify new threats and respond quickly with unique services informed by real-time global cybersecurity intelligence delivered at cloud scale.

AAzure is global

With datacenters in more regions than any other cloud provider, Azure provides a global reach with local presence that many businesses and organizations need, allowing them to reduce the cost, the time, and the complexity of operating a global infrastructure while meeting local data residency needs.

The advantages of Azure over AWS

Organizations all over the world recognize Microsoft Azure over Amazon Web Services (AWS) as the most trusted cloud for enterprise and hybrid infrastructure for many reasons:

Competitive pricing

AWS is 5 times more expensive than Azure for Windows Server and SQL Server. Azure matches AWS pricing for comparable services.

Achieve more with open source on Azureore

Use any open-source OS, languages, and tools on Azure. Azure made the most contributions to GitHub in 2017 and it's the only cloud with integrated support for Red Hat.

Enhanced proactive security and compliance

Compare AWS and Azure and you'll find that Azure's compliance offerings—including 70+ compliance certifications—and are more comprehensive.

Get more value from your existing Microsoft investment

Keep using your organization's existing tools and knowledge: get a consistent experience across your on-premises and cloud technologies by integrating them with Azure Active Directory.

Azure is the future

The advent of the cloud and smart technologies is revealing new scenarios that were simply not possible until now. Smart sensors and connected Internet of Things (IoT) devices now allow us to capture new data from industrial equipment: from factories to farms, from smart cities to homes. And whether it's a car or even a refrigerator, new devices are increasingly cloud connected by default.

Simultaneously, hybrid cloud is evolving from being the integration of on-premises datacenters with the public cloud, to becoming units of computing that are accessible even in the world's most remote destinations.

By bringing these two new realities together—and with artificial intelligence running across all systems—we have entered the era of the intelligent cloud and intelligent edge.

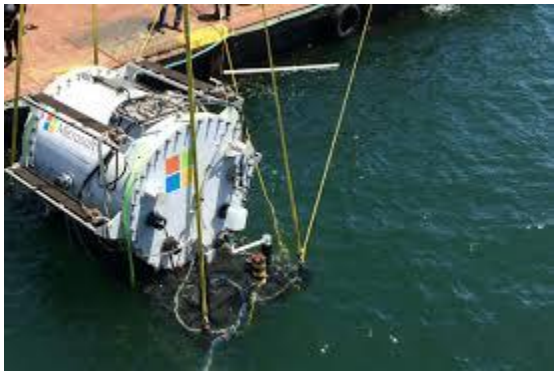
The intelligent edge is a continually expanding set of connected systems and devices that gather and analyze data—close to your users, the data, or both. Users get real-time insights and experiences, delivered by highly responsive and contextually aware apps. The Azure platform is built to provide an agile and secure experience across the intelligent cloud/intelligent edge in alignment with these enduring principles:

PHOTO GALLERY OF MICROSOFT DATA CENTERS

Microsoft tests a datacenter...

Data Centers Located...

Microsoft Inspire Data Center...



Underwater data centers...

Microsoft Center of Excellence Update...

Generation 5 Data Center Design

Microsoft sinks data centre off Orkney...

Microsoft's Underwater Data Center Is...

Why Microsoft wants to put data centers...

Microsoft Azure now 'pay-as-you-go ...

With LinkedIn Deal, Microsoft Expands...

Microsoft tests a datacenter...

Microsoft's deep sea data center is now ...

engadget.com

Microsoft data center in Chicago ...

researchgate.net

Microsoft beats Google and Amazon to ...

venturebeat.com

Microsoft now has one million servers ...

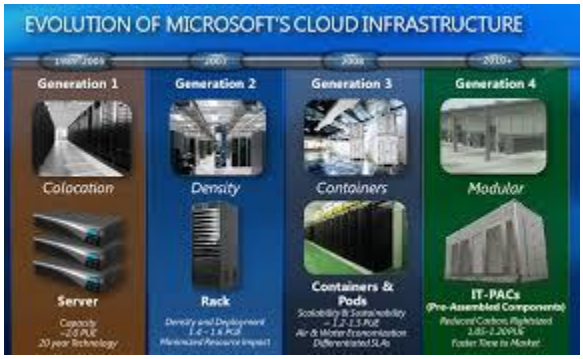
extremetech.com

Moves Away from Data Center Containers ...

datacenterknowledge.com

Microsoft announces two new datacenter ...

mspoweruser.com



Microsoft Shares Video Tour of its ...

blogs.technet.microsoft.com

Microsoft partners with Canberra Data ...

zdnet.com

Microsoft sinks data centre off Orkney ...

bbc.com

Microsoft ouvre ses data centers en ...

news.microsoft.com

Microsoft buys $80 million data center ...

mysanantonio.com

90 MINUTES makes! Microsoft ...

data-economy.com

Microsoft Data Center Quincy - Cascade ...

cascadedoorandhardware.com

Project Alluvion' data ...

geekwire.com

Microsoft to build cloud data centres ...

cbronline.com

Future of the Cloud Lies Beneath the ...

engineering.com

Microsoft Just Put a Data Center on the ...

vice.com

Microsoft Datacenter Tour - YouTube

youtube.com

Microsoft deploys underwater data ...

newatlas.com

Inside Microsoft's Quincy data center - DCD

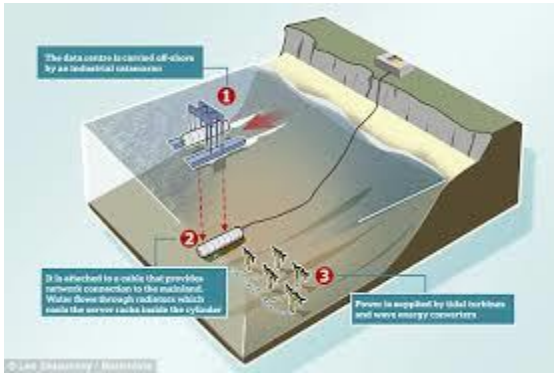
datacenterdynamics.com

ClickDimensions Launches in Microsoft's ...

blog.clickdimensions.com

Microsoft Azure Datacenter Architecture ...

youtube.com



Microsoft's underwater cloud: First ...

dailymail.co.uk

Microsoft Azure Data Center Locations ...

joranmarkx.wordpress.com

Microsoft Plans 2 New Data Centers in ...

baxtel.com

Redesigning Datacenters for an Advanced ...

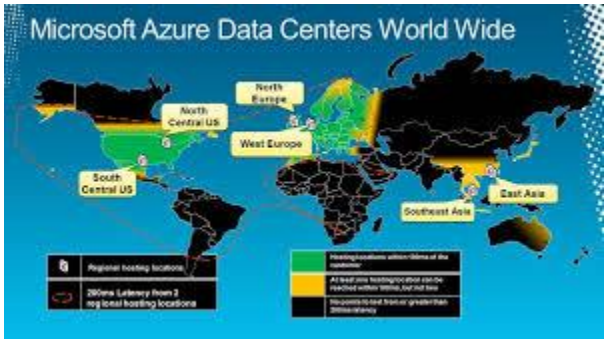
blogs.microsoft.com

data center now has a webcam with fish ...

theverge.com

Microsoft sinks data centre off Orkney ...

bbc.com



Windows Azure Data Center Locations ...

joranmarkx.wordpress.com

Microsoft Data Center Locations, News ...

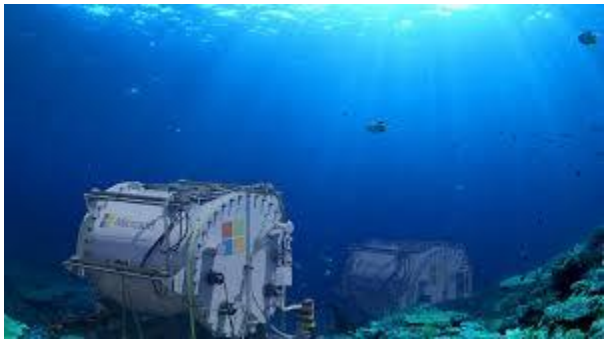
baxtel.com

Microsoft research project puts cloud ...

news.microsoft.com

BBCC & Quincy Data Centers Partner for ...

bigbend.edu



first underwater data center

outthere.fr

cooling a breath of fresh air ...

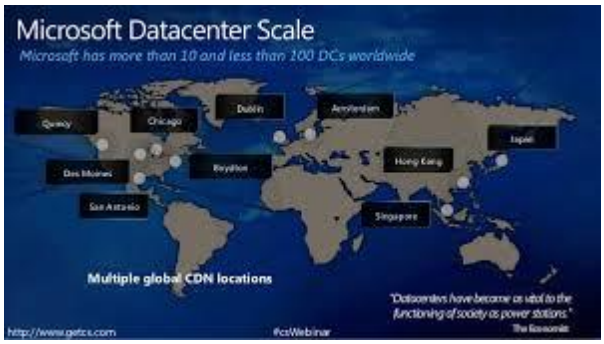
gcn.com

Azure Cloud Outage ...

datacenterknowledge.com

U.S. v. Microsoft: Is Your Data and ...

bytebacklaw.com



Traffic Safety Resource Center

trafficsafetystore.com

Canberra data centres ...

which-50.com

deploys an underwater data center ...

ciobulletin.com

Gateway to the Cloud

datacenterfrontier.com

Dublin data centres ...

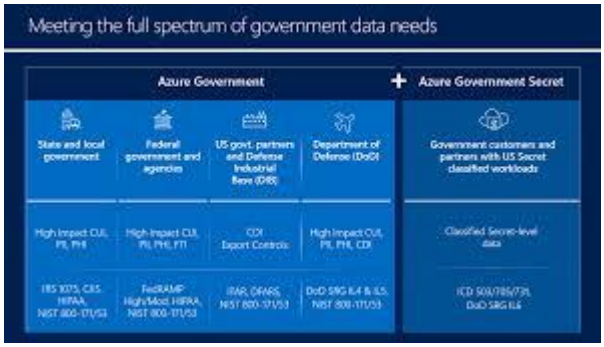
independent.ie

Three gas-powered data centers planned ...

datacenterdynamics.com

Inside Microsoft's Chicago Data Center ...

datacenterknowledge.com



Azure Government Secret Data Centers

crn.com

invest $2.7 billion in a new Dutch data ...

windowscentral.com

Microsoft: Company Profile, Data Center ...

datacenters.com

Microsoft Moves a Data Center Under the ...

nextgov.com

deliver the intelligent cloud ...

news.microsoft.com

Microsoft announces plans for new data ...

windowscentral.com

Data Center Fueled By Waste Gases ...

technologyreview.com

Unpacking Office 365: Exploring ...

avepoint.com

deliver the intelligent cloud ...

news.microsoft.com

Photos: Inside a Microsoft data center

cnet.com

Super-Efficient Underwater Data Center ...

greentechmedia.com

Microsoft Data Center Locations, News ...

baxtel.com

Microsoft Datacenter Security - YouTube

youtube.com

Microsoft beats Amazon to open the ...

engadget.com

Data Center Turns to Natural Gas ...

groebner.com

Azure data centers across ...

windowscentral.com

Microsoft now has one million servers ...

extremetech.com

TECHNOLOGY: Data center or power plant ...

eenews.net

Cloud Move to Secret Microsoft Data ...

bloomberg.com

Microsoft Corp.'s data center was down ...

bizjournals.com

Data Centers In The UK ...

digitaltrends.com

W.D.M. data center will be Microsoft's ...

desmoinesregister.com

South Central US data center went down ...

onmsft.com

Microsoft opens first African data ...

techradar.com

Energy Practices in Quincy Under Fire ...

datacenterknowledge.com

Microsoft opens first African data ...

techradar.com

Driving Data Center Modernization ...

blogs.cisco.com

Tour Of Microsoft's Hyperscale Datacenters

nextplatform.com

Middle Eastern data center regions ...

menabytes.com

Microsoft Launches Second Subsea Data ...

oilprice.com

sunk a data center to the ocean floor

cnbc.com

Tech's High-Stakes Arms Race: Costly ...

wsj.com

Microsoft Sinks Data Center in the Sea ...

interestingengineering.com

Microsoft submerges a data center off ...

techspot.com

Microsoft Plumbs Ocean's Depths to Test ...

nytimes.com

Why Microsoft wants to put data centers ...

techcrunch.com

Microsoft commits to running data ...

computerworld.com

Biggest Data Centers ...

westhost.com

data center regions in Middle East ...

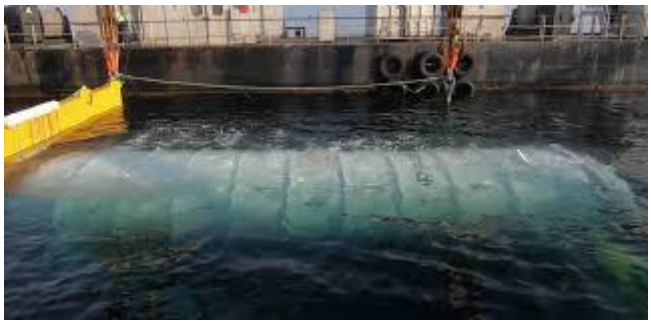
fiercetelecom.com

building a massive data center ...

techrepublic.com

Inside Azure datacenter architecture ...

youtube.com



Microsoft's Project Natick criticized ...

mspoweruser.com

Microsoft sites in metro Phoenix will ...

azcentral.com

Microsoft's data center in Quincy ...

pinterest.com

Microsoft, Daimler to use fuel cells to...

Apple Steals Microsoft Employee To Run...

Microsoft plans two SA data centers ...

africatimes.com

Tour a Microsoft Datacenter to learn ...

itprotoday.com

Microsoft Cloud to set up shop in...

Cloud Data Centers in Germany...

Microsoft opens Azure data centers in...

data centers in Norway to power ...

neowin.net

New Microsoft data center in Iowa ...

blogs.seattletimes.com

Microsoft Data Center Locations, News ...

baxtel.com

Microsoft Accelerates its Cloud Growth

datacenterfrontier.com

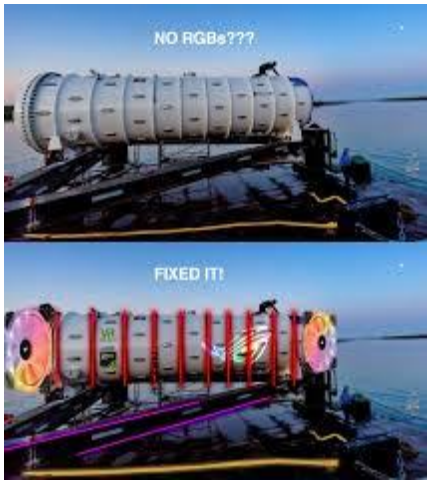
Data Center Locations | Greg Olsen ...

yellowduckguy.wordpress.com

Microsoft Brings Two More Mega Data ...

blogs.technet.microsoft.com

Microsoft opens first data centers in ...



Microsoft Submerges Ocean Cooled Data...

Might Come Up With Underwater Data Centers

medium.com

Microsoft sites in metro Phoenix will ...

azcentral.com

Microsoft: Chicago Northlake Data Center

Facebook meet hyper scale computing...

Microsoft brings 3 data centers, 1,100...

PROVIDER	REGION & SUBREGION
AWS US	US East (N Virginia)
AWS US	US West (N California)
AWS US	US West (Oregon)
AWS	GovCloud (Oregon)
AWS	South America (Sao Paulo)
AWS	EU (Ireland)
AWS	Asia Pacific (Singapore)
AWS	Asia Pacific (Tokyo)
AWS	Asia Pacific (Sydney)
Google	Central US (Council Bluffs, IA)
Google	Central US (Pryor Creek, OK)
Google	Europe (Europe)
Microsoft	Azure North-central US (Chicago, IL)
Microsoft	Azure South-central US (San Antonio, TX)
Microsoft	Azure West US (California)
Microsoft	Azure East US (Boydton, Virginia)
Microsoft	Azure East Asia (Hong Kong, China)
Microsoft	Azure South East Asia (Singapore)
Microsoft	Azure Northern Europe (Dublin, Ireland)
Microsoft	Azure West Europe (Amsterdam, Netherlands)

Cloud datacenter locations...

Fuel Cells to Power Data Centers

Watch Live as Fish Swim by Microsoft's...

Microsoft Azure South Africa datacenter...

Microsoft out on arrogance in Quincy ...

datacenterdynamics.com

Strong cloud computing sales indicate ...

networkworld.com

Wales lost out on Microsoft deal as it ...

bbc.co.uk

Apple to build data center in Iowa ...

venturebeat.com

Microsoft Windows Server 2019 ...

mychoicesoftware.com

Microsoft Data Center Academy ...

innovationia.com

Microsoft Data Center - Apstra

apstra.com

Microsoft Announces $1.1 Billion Data ...

iowapublicradio.org

Data Centers & Renewable Energy ...

geo-tel.com

Microsoft signs new wind energy deal ...

technologyrecord.com

A Field Trip to the Microsoft ...

redmondmag.com

Microsoft Silver Datacenter Competency ...

cl8.com

Microsoft setting up data centres in ...

s-ge.com

Microsoft plans for more data centers ...

bizjournals.com

Microsoft Sinks Data Center in the Sea ...

interestingengineering.com

cloud provider ...

mashable.com

Microsoft gets green light from Qatar ...

reuters.com

cooling its data centers ...

arstechnica.com

Data Centers Launch in Switzerland ...

zuehlke.com

Clouds under the sea: Microsoft deploys ...

geekwire.com

Microsoft Azure Cloud ...

infochola.com

Microsoft Data Center Tour (Amsterdam ...

itsdone.at

DataON Announces New Data Center Plus ...

aithority.com

data center set for West Des Moines

desmoinesregister.com

5 days ago

Government Secret Data Centers Amid ...

sdxcentral.com

data center goes completely off grid ...

treehugger.com

Way Around Data Center Traffic Jams ...

nytimes.com

Microsoft and the Software Defined Data ...

zdnet.com

Microsoft data centres globally ...

securitas.uk.com

Microsoft increases European cloud data ...

networkworld.com

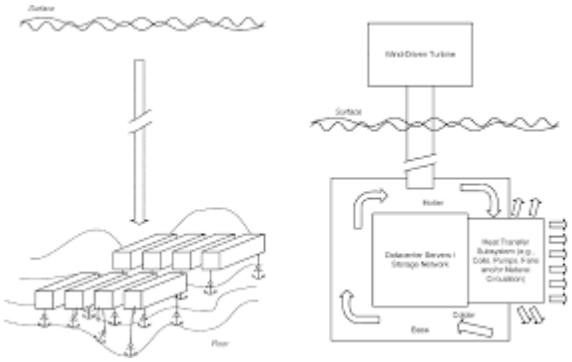
Microsoft to open two data centers in ...

Microsoft tests underwater data centre ...

bbc.com

Africa for Its Data Center Expansion ...

marketrealist.com



The Future Of Data Centers - CB ...

cbinsights.com

data centers in UAE for Azure ...

unlock-bc.com

A Smarter Commercial Cloud | SIGNAL ...

afcea.org

HPE Microsoft Storage Solutions

brighttalk.com

Microsoft sank a data center in the ...

treehugger.com

Data Center Outage

environmentalleader.com

Microsoft Azure to Eastern Europe ...

Microsoft

Microsoft expanding data center at the...

Microsoft launches two new cloud ...

cio.com.au

Microsoft: Quincy MWH Data Center

datacenters.com

Microsoft Azure Datacenter Architecture ...

neroblanco.co.uk

Data Centers, Offers Preview of Azure

governmentciomedia.com

Amazon.com: Microsoft Windows Server ...

amazon.com

data center energy ...

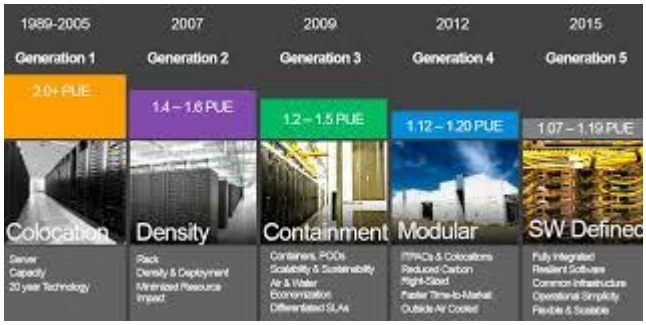
smartenergydecisions.com

Microsoft Azure Datacenter ...

itprotoday.com

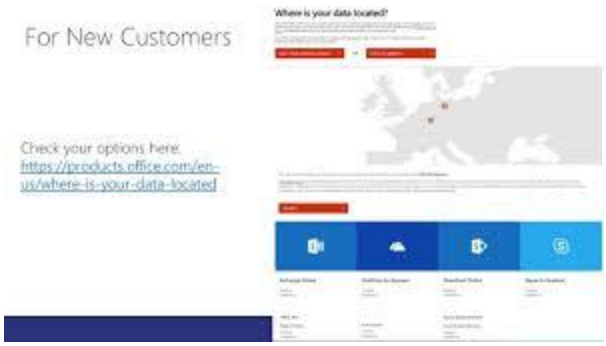
380m Dublin data centre gets go ...

independent.ie



Tour Of Microsoft's Hyperscale Datacenters

nextplatform.com



Microsoft Data Center Locations ...

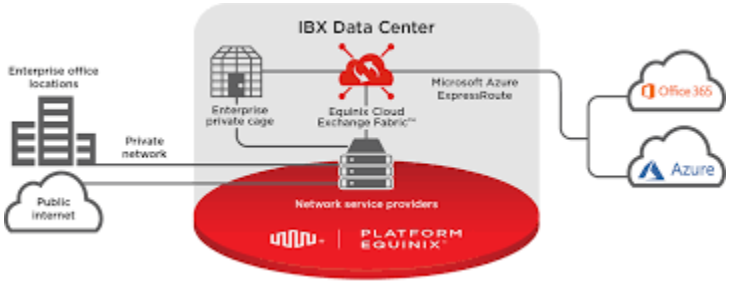
recordpoint.com

launches Middle East data centers ...

medium.com

Microsoft: Can't wait for ARM to power ...

theregister.co.uk



Microsoft Azure - Platform Partners ...

equinix.com

10 Fascinating Data Centers ...

digitalinformationworld.com

build two massive Azure data centers ...

dailyhostnews.com

Microsoft Azure Datacenter Architecture ...

neroblanco.co.uk

Pros & Cons of an Underwater Data Center

hostdime.com

Photos: Inside a Microsoft data center ...

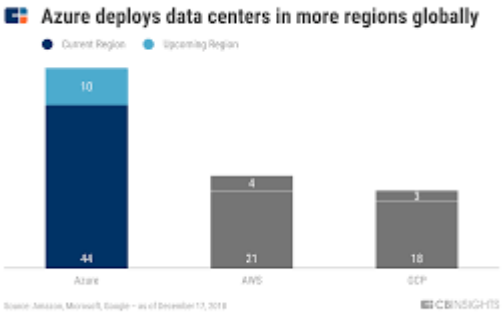
cnet.com

Shortages Force Gas-Fueled Power Plant

racksolutions.com

data centers in South Africa

born2invest.com



Microsoft Strategy Teardown

data center off Scottish island ...

dailystar.com.lb

Microsoft Partner

Microsoft

Gold Datacenter

Microsoft System Center 2019

comparex-group.com

Microsoft

Expand Southern Virginia Data Center ...

capremedia.com

Equinor partners with Microsoft for ...

equinor.com

soon open its first two data centers ...

techcrunch.com

10 Fascinating Data Centers ...

digitalinformationworld.com

Microsoft Data Center Evolution

florence20.typepad.com

Deploys Underwater Data Center ...

subtelforum.com

Microsoft catapults FPGAs to new ...

rambus.com

Windows Server Datacenter (Discounted)

techsoup.org

Microsoft data centers

ciobulletin.com

Microsoft to build new data center ...

usaherald.com

Data Center in San Antonio

cpexecutive.com

Microsoft Seattle data centre to be ...

businessgreen.com

Azure at new Korean data centers ...

geekwire.com

Lenovo lights up Microsoft Ignite 2016 ...

blog.lenovo.com

Azure Stack Lets You Put The Microsoft ...

prescientsolutions.com

Microsoft Data Center Waterproofing ...

chamberlinltd.com

Microsoft Adds Private Azure Links at ...

itprotoday.com

Inside Azure Datacenter Architecture ...

channel9.msdn.com

Azure North Europe downed by the curse ...

theregister.co.uk

Data Chain ...

americanbuildersquarterly.com

Microsoft Azure Datacenter Architecture ...

neroblanco.co.uk

Computational Demand Response ...

greentechmedia.com

Overheating brings down Microsoft data ...

datacenterdynamics.com

power its Singapore datacenter services ...

hardwarezone.com.sg

Data Center Power ...

insights.dice.com

Microsoft Windows Server 2019 ...

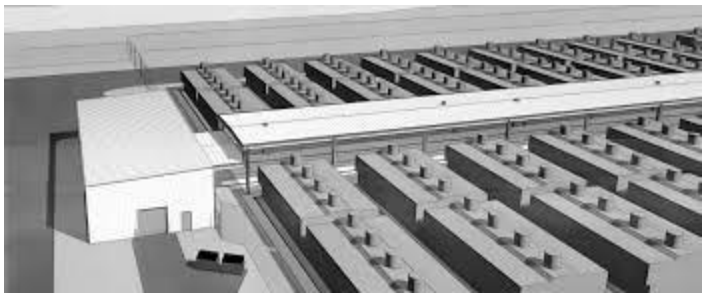
consogame.com

Inside a hyperscale data center (how ...

computerworld.com

Microsoft Data Center Locations, News ...

baxtel.com



Microsoft BN6 Data Center | Lithko

lithko.com

First Solar planning 150-MWac Arizona ...

solarpowerworldonline.com

Photos: How Apple, Google, Facebook and ...

techrepublic.com

Generation 5 Data Center Design

datacenterfrontier.com

More Microsoft Data Centers ...

connect.media

Microsoft data centers go beyond the ...

cnet.com

Microsoft puts a data center on the sea ...

foxnews.com

Microsoft wasted 'millions of watts of ...

theverge.com

Microsoft: Amsterdam Data Center

datacenters.com

Cloud Data Center In India

trak.in

Microsoft to build Azure data centres ...

techcentral.co.za

Microsoft cloud infrastructure

techgenix.com

Data Center Design Evolution | The ...

esciencegroup.com

Data Center in Central Iowa

channelfutures.com

underwater data centre ...

eandt.theiet.org

Office 365 Datacenter Region ...

office365itpros.com

Microsoft's Bringing New Data Centers ...

petri.com

Microsoft Azure Data Center Tour ...

youtube.com

UBS Said to Prepare Cloud Move to ...

advisorhub.com

Windows Server 2019 Datacenter OEM 24 ...

scan.co.uk

Opening Data Centers in the Middle East ...

marketrealist.com

Using Biogas Fuel Cell to Power Wyoming ...

treehugger.com

Interactive Microsoft Datacenter Tour ...

blog.nicholasrogoff.com

Data Center in Iowa ...

wired.com

square foot Microsoft data ...

pinterest.com

successfully tests underwater data center

rcrwireless.com

storing Teams data for South Korea ...

onmsft.com

Cloud-Data Centers in Middle East ...

bloomberg.com

plans massive new data center in Quincy ...

geekwire.com

Hyper-Scale Operators Lower Data Center ...

crn.com

Google to launch first data centre in ...

livemint.com

Fuel Cell Datacenter project ...

hostingjournalist.com

ran Azure from an undersea datacenter ...

zdnet.com

Having flogged off its data centers, AT ...

theregister.co.uk

saves energy with cloud storage under water

betterworldsolutions.eu

Microsoft UK Data Centres | What does ...

chorus.co

Microsoft Building Cloud Data Centers ...

racksolutions.com

Plans $1.13 Billion Data Center Project ...

hothardware.com

IHS Markit: Global data center ...

fiercetelecom.com

Dell EMC broaden support for Microsoft ...

techobserver.in

Microsoft Data Center - Rocky Mountain ...

rmecinc.com

Data Centers: Microsoft's Underwater ...

facilitiesnet.com

Microsoft data centers in Middle East ...

ciobulletin.com

wasted energy at data center to avoid ...

engadget.com

Microsoft data center ...

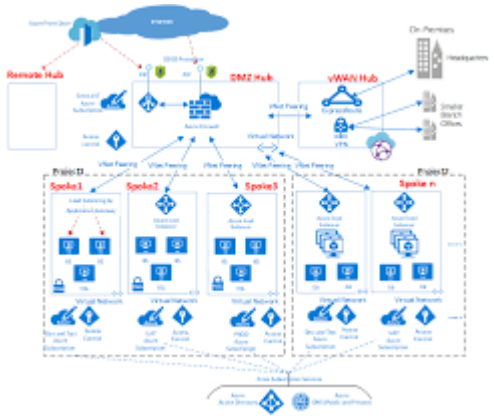
datacenternews.asia

778-microsoft-datacentre-dublin - Core ...

coreelectrical.com

Borssele 3&4 to green Microsoft data ...

renews.biz



Microsoft Azure virtual datacenter HUB ...

mountainss.wordpress.com

Microsoft doubles Amazon by opening ...

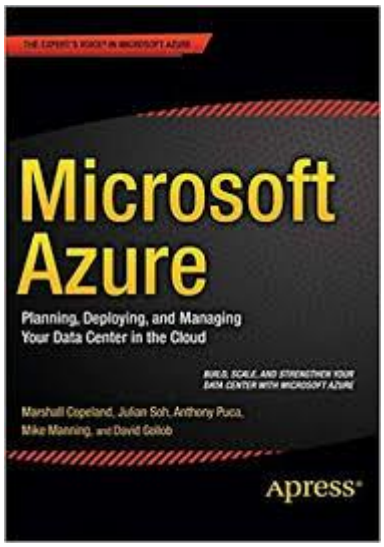
afrikatech.com

invests $34 million in Westover Hills ...

bizjournals.com

Data centers use a ton of water. Is ...

azcentral.com



Microsoft Azure: Planning, Deploying ...

amazon.com

Data Center Campuses ...

thearizonatribune.com

First Solar Project To Power Microsoft ...

solarindustrymag.com

Underwater Data Centers May Cost Less

Greenpeace: Google, Microsoft & Other...

Tour Of Microsoft's Hyperscale Datacenters

nextplatform.com

Microsoft invents an underwater data ...

Dell EMC and Microsoft Deliver for the ...

blog.dellemc.com

Cloud Data Centers In South Africa ...

relocationafrica.com

Microsoft Datacenter Tour - Digital ...

f2bservices.com

2bn Netherlands data center revealed ...

datacenterdynamics.com

Microsoft to build 3 data centers in ...

abc15.com

Microsoft Azure Cloud Computing Amazon ...

favpng.com

The 6 Largest Data Centers in the World

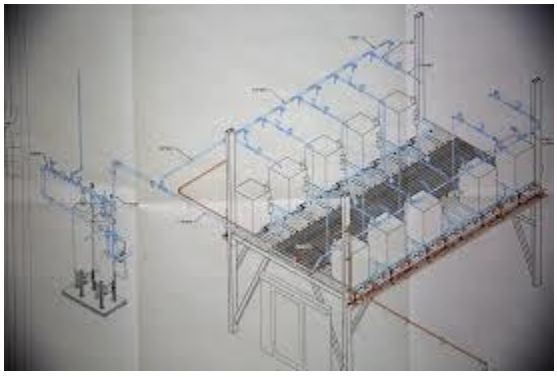
cbtnuggets.com

What Is Microsoft Azure Stack and How ...

blog.5nine.com

My PUE is Getting Better Than Your PUE ...

treehugger.com



crazy' bet on fuel cells to feed power ...

seattletimes.com

Cloud Move to Secret Microsoft Data Centers

finance.yahoo.com

Microsoft Azure Adding Two Data Centers ...

itprotoday.com

Underwater Data Center

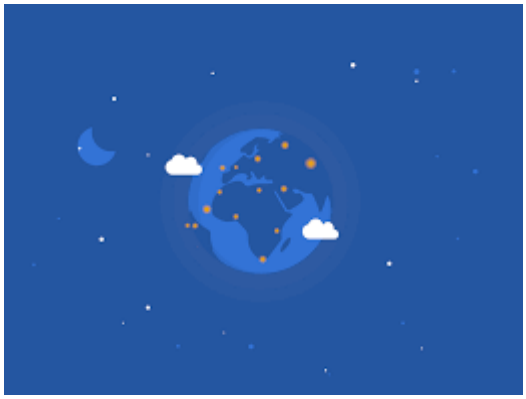
thefreshtoast.com

CloudJumper Joins Microsoft DCO Program ...

aithority.com

data center investment in Wyoming ...

techrepublic.com



Data Centers by EJ Hassenfratz on Dribbble

dribbble.com

Microsoft is witnessing a strong uptake ...

techradar.com

Microsoft successfully loaded ...

pangnote.com

Microsoft Azure Datacenter Architecture ...

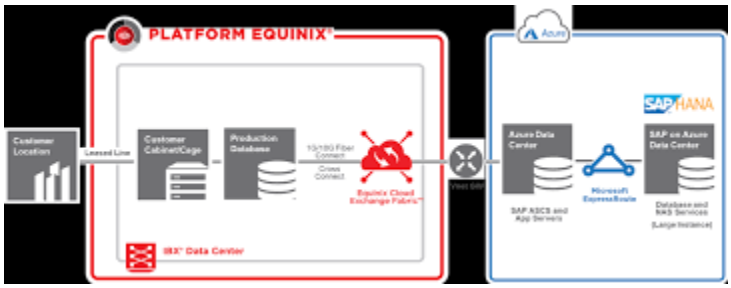
neroblanco.co.uk

Microsoft Sinks Data Center in the Sea ...

interestingengineering.com

Microsoft's Data Center of the Future ...

greenbiz.com



Migrating Workloads to Microsoft Azure

blog.equinix.com

Microsoft – Data Center ...

siemensusa.synapticdigital.com

Microsoft forced to build Dublin power ...

independent.ie

Undersea Data Center: Second Test by ...

opticalcloudinfra.com

Cloud Computing Software-defined Data ...

imgbin.com

microsoft-datacenter-100058814-orig ...

dctalk.com.bd

Microsoft to spend $1.1bn to build Iowa ...

theregister.co.uk

Project Natick' puts Microsoft's data ...

neowin.net

HCI to the Azure Stack for Cloud Customers

channelfutures.com

Amazon, Google, Microsoft Cloud Data ...

channele2e.com

896 Xeon Cores in One PC: Microsoft's ...

anandtech.com

Microsoft Generation 4.0 Data Center ...

youtube.com

CDW + Microsoft Data Center

studylib.net

data center on the bottom of the sea ...

cablingtek.com

Microsoft launches UK data centres for ...

definetomorrow.co.uk

Data center sector gets more crowded ...

ipwatchdog.com

Google, Facebook, Microsoft, and ...

theverge.com

Microsoft Windows Server 2019 ...

digitalsoftwaremarket.com

Uber Buys Microsoft Data center and ...

digitalo2.com.au

Microsoft Plumbs Ocean's Depths to Test ...

nytimes.com

Microsoft Roofless DataCenter - Natural ...

500eco.com

Windows Server 2019 Datacenter – 2 Core ...

trustedtechteam.com

data centers draw other businesses ...

roselawgroup.com

underwater datacenter Project Natick ...

menosfios.com

Microsoft Windows Server 2016 ...

mychoicesoftware.com

data centres in the UK ...

itproportal.com

Datacenter Community Meetup ...

ec-mea.com

Microsoft Windows Server 2016 ...

softvire.com.au

Microsoft Data Center San Antonio Eyeball

cryptome.org

Invest Half-Billion Dollars in Va ...

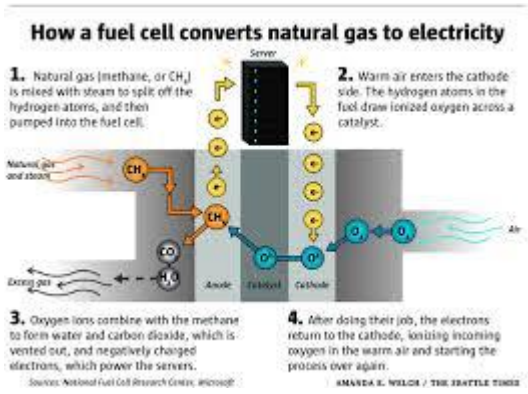
greenbiz.com

Buzzfeed: The 10 Largest Data Centers ...

blog.rackmountsolutions.net

two data centers in Kenya, Nigeria

kbc.co.ke



crazy' bet on fuel cells to feed power ...

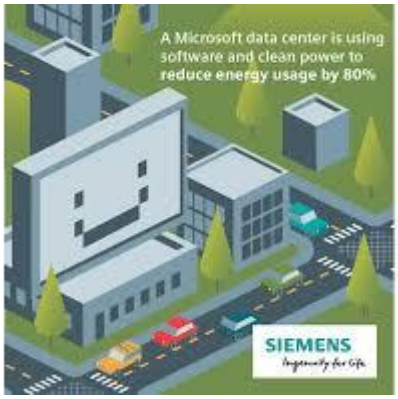
seattletimes.com

Microsoft Data Center Operations Compaq ...

ebay.com

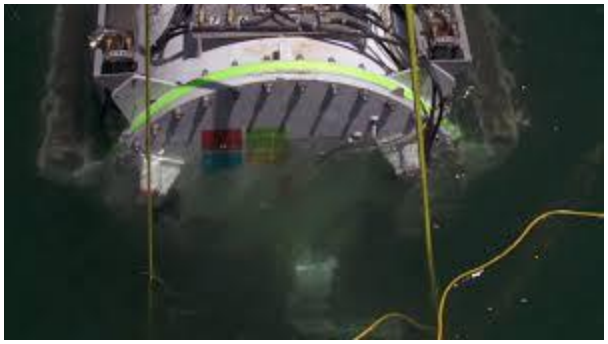
Microsoft Windows Server 2019 ...

softwarekeep.com



Microsoft Data Center

siemensusa.synapticdigital.com



Microsoft is putting a Datacenter on ...

menosfios.com

22 hours ago

Service IT Infrastructure Data Center ...

favpng.com

data centres in South Africa

mybroadband.co.za

Microsoft Tweets About Major Outage

channelpartnersonline.com

With onsite biogas and fuel cells ...

cnbc.com

Emerging Data Center Trends.ppt ...

slideshare.net

Microsoft Research paper proposes using ...

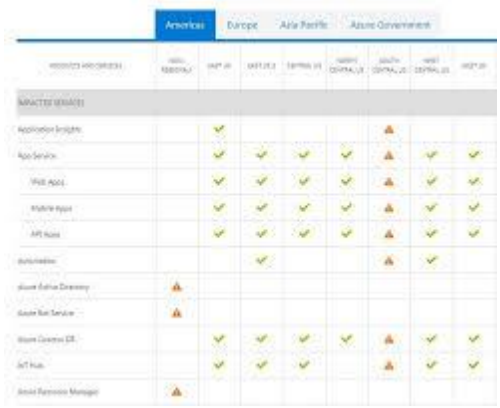
newatlas.com

Westpac and Microsoft launch a data ...

mashviral.com

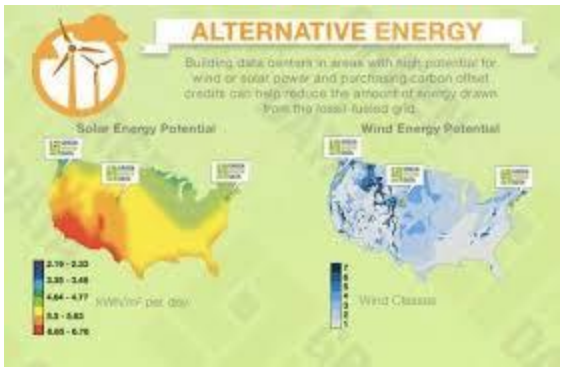
Microsoft Looking at Ultra-Modular Data ...

treehugger.com



Microsoft South Central US datacenter ...

zdnet.com



Google, Facebook and Microsoft data ...

techrepublic.com

Microsoft Submarine Data center: A new ...

blog.unisecure.com

Microsoft Data Center Outage ...

fiber.net

Windows Server 2019 | Datacenter ...

softvire.co.nz

Microsoft datacenters open in SA ...

crimsonline.co.za

Documents: Microsoft Goodyear project a ...

bizjournals.com

Microsoft sinks data centre to ...

mobilesyrup.com

Opening A $1.1 Billion Data Center In Iowa

eyeonwindows.com

UK data centres for cloud computing ...

telegraph.co.uk

data center entirely on wind power ...

networkworld.com

heating your home with Data Furnaces ...

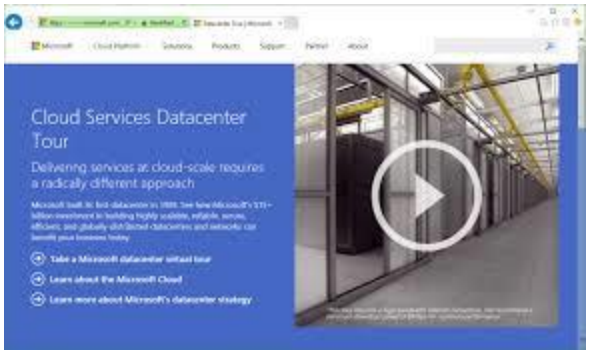
extremetech.com

Microsoft set to establish Korean data ...

techerati.com

Construct Another Data Center in Quincy ...

portofquincy.org



Microsoft Azure Virtual Datacenter Tour ...

thomasmaurer.ch

renewable energy for US data centers

datacenternews.us

Microsoft data center taking shape in Iowa

kcci.com

Deutsche Telekom to Open Azure Cloud ...

thefastmode.com

Stacks Up Against Data ...

fastcompany.com

Photos: Inside a Microsoft data center ...

cnet.com

Microsoft brings Data Center to Canada ...

synergyit.ca

Microsoft adds Azure cloud data centers ...

geekwire.com

New Microsoft data center to offer ...

desmoinesregister.com

Apple Poaches Microsoft's Datacenter GM ...

macstories.net

Webinar: Microsoft and Dell EMC Extend ...

sdxcentral.com

Cloud Data Centers Are Coming to the ...

datacenterfrontier.com

SNJ T1993 | “Microsoft's deep sea data ...

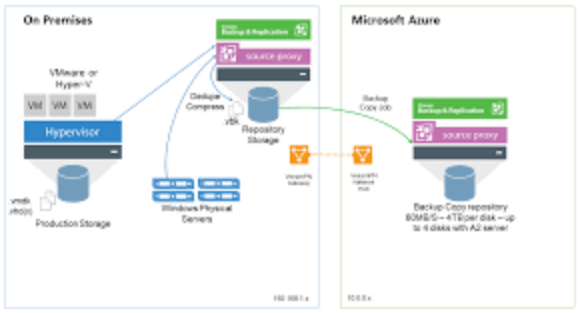
smitanairjain.blog

Microsoft® Olympus 42U Computing Rack ...

opencompute.org

Microsoft Core Infrastructure Server ...

Microsoft-Data-Center-#3-2014.10rs...



Veeam: Backup and restore workloads to ...

PC partner Intel seeks lift off

yahoo.com

Trials Underwater Natick Data Center ...

news.filehippo.com

Microsoft to open UK data centres - BBC ...

bbc.co.uk

Data centers - Washington State ...

ecology.wa.gov

Microsoft Windows Server 2008 R2...

The First Deep Sea Data Center

FEDERAL BERUE INVESTIGATION AGENCE (FBI)
Washington, D.C.
FBI National Press Office
(202) 324-3691

FBI

November 18, 2019

FBI Cuts Ribbon for New Data Center in Idaho

Today, the FBI and its partners at the Department of Justice (DOJ) participated in a ribbon-cutting ceremony for a new data center at the FBI's facility in Pocatello, Idaho. This endeavor has been years in the making and is a significant milestone in the FBI's broader information technology transformation initiative. The new data center will optimize infrastructure, information, and services by consolidating almost 100 data centers throughout DOJ. It substantially expands the footprint of the FBI's existing Pocatello facility, which has been maintained for more than three decades.

This expansion provides the FBI and DOJ with approximately 140,000 square feet of additional space, accommodating both data center and office space capacity. This consolidation will enable the FBI to reduce operational costs, create efficiencies (most notably through energy reduction), and modernize the technological architecture.

The consolidation of dozens of data centers will strengthen the cybersecurity posture for all DOJ components using the facility. It will also enhance collaboration, information sharing, and the ability to

execute advanced analytics. The new data center will provide the flexibility needed to focus on and rapidly respond to mission requirements.

The FBI appreciates the support of the Pocatello community in helping to bring this project to fruition.

1761

A data hall inside the FBI's new data center in Pocatello, Idaho, along with exterior side and front views of the facility.

The FBI and DOJ Opens $100m Idaho Data Centre Hub

By Abigail Opiah **PUBLISHED:** 12:05, 19 November, 2019 **UPDATED:** 12:05, 19 November, 2019

The data centre in Idaho

The Federal Bureau of Investigation (FBI) and the Department of Justice (DOJ) opened the doors to its data centre facility in Pocatello, Idaho.

Members of the FBI and the DOJ attended a ribbon-cutting ceremony for the new data centre yesterday, and revealed that the facility is set to optimize infrastructure information and services by consolidating almost 100 data centers throughout the DOJ.

“This Endeavour has been years in the making and is a significant milestone in the FBI’s broader information technology transformation initiative,” said the FBI in a written statement.

“It substantially expands the footprint of the **FBI’s existing Pocatello facility**, which has been maintained for more than three decades.”

The expansion provides the FBI and the DOJ with approximately 140,000 sq ft of additional space, accommodating both data centre and office space capacity.

The FBI added that the consolidation will enable the FBI to reduce operational costs, create efficiencies (most notably through energy reduction), and modernize the technological architecture.

09:11 - 18 Nov 2019

“The consolidation of dozens of data centres will strengthen the cybersecurity posture for all DOJ components using the facility,” added the FBI.

“It will also enhance collaboration, information sharing, and the ability to execute advanced analytics. The new data centre will provide the flexibility needed to focus on and rapidly respond to mission requirements.”

East Idaho News reported that the FBI invested $100m into the renovations of the data centre facility, and there will now be room for 250 more people to work in the building, joining the 300 employees already working in there.

In 2017, the FBI broke grounds on the Pocatello data centre, and revealed that it has had a presence in Pocatello for three decades, gradually increasing personnel and capabilities at the facility over time.

Fabienne Dulac Deputy CEO of the Group, CEO of Orange France

Orange France has begun constructing its new data centre facility in the municipality of Amilly in France.

The data centre called ‘Center-Val de Loire’ is scheduled to be completed by the end of 2020, and will cover an area of 16,000 m², including 5,000 m² of computer rooms.

The data centre will be used to host data and services of Orange’s corporate and consumer customers.

The company has invested over €100m for the cumulative construction of the Center-Val de Loire data centre and the Normandy 2 data centre.

The company began the construction of its Normandy 2 data centre facility in May this year, and the two data centers will be connected by a network of synchronous replication of data.

“This data centre is a perfect illustration of Orange’s major strategic challenges: it will enable us to support the explosion of our customers’ digital uses in a context of trust, while reducing our energy and environmental footprint,” said Fabienne Dulac, Deputy Chief Executive Officer of the group and CEO of **Orange France**.

“Its establishment in Amilly testifies to our desire to involve ourselves in all territories and the quality of our cooperation with the local elected representatives of the city that I wish to salute.”

The company also added that the Center-Val de Loire region was chosen to host the data centre because it benefits from several assets, including ideal weather conditions for the use of ‘free cooling’.

“I welcome the installation of this data centre on the territory,” said Jean-Pierre Gorges, Mayor of Chartres and President of Chartres Métropole.

“Chartres Métropole confirms its position on the digital scene in France and I am delighted. Our territory becomes a little more an economic lung for the centre region and an attractive node 100 km from Paris. In the age of the data, this equipment is major for the development of our territory.”

The FBI's Edgar J. Hoover Building in Washington, DC. 2017

FBI Data Center Construction Begins in Idaho

DOJ expects consolidation initiative will strengthen its cybersecurity posture

Construction crews broke ground on a site in Pocatello, Idaho, home to one of the data centers into which the FBI plans to consolidate

existing data centers. The bureau held a groundbreaking event in Pocatello together with local officials last week.

FBI infrastructure plans are part of the broader Department of Justice Data Center Transformation Initiative, which in turn is part of the government-wide effort to consolidate data centers, reduce the amount of redundant software, and whenever possible move workloads from on-premises facilities to cloud services.

The DOJ's plan is to consolidate dozens of data centers to three "core enterprise facilities" in Pocatello; Clarksburg, West Virginia; and Washington, D.C., according to FCW.

For a variety of reasons, government agencies and departments have had a difficult time achieving goals or meet deadlines set by the initiative.

One of the biggest constraints has been funding. Officials from the DOJ's office of the CIO were cited in a recent Government Accountability Office report on Federal data center consolidation efforts that funding constraints could delay moving applications to the cloud or hardware to the target consolidation facilities.

The DOJ expects data center consolidation to strengthen all its components' cybersecurity posture. The Pocatello data center -- expected to cost $100 million, according to Idaho State Journal, will serve all those components.

Components of nine FBI divisions are located in the existing Pocatello facilities, their responsibilities including investigations, intelligence analysis, IT, travel, and records management. The bureau expects this presence to expand when the new data center is completed.

FBI launches Pocatello 'Grey Wolf' data center

The unpredicted Mueller report gets a new home

November 19, 2019

By

Sebastian Moss

The FBI has opened its data center in Pocatello, Idaho, two years after *DCD* first reported on the facility.

Known as 'Grey Wolf,' the 100,000 square foot (9,300 sq m) facility is part of a $100m expansion that includes a 40,000 sq ft (3,700 sq m) office addition, both built by JE Dunn Construction.

Release the wolf

The FBI Pocatello facility– FBI

In the procurement documents, the FBI states that the Tier III facility has 25,000 sq ft (2,300 sq m) of data floor space with an optional 8,000 sq ft expansion.

"The initial IT installation will be 2.7MW with [the] capability to grow to 5.4MW of IT power," one document notes. "The facility shall be designed and constructed such that the expansion of the data center IT operations, as well as the power and cooling supporting infrastructure, will not experience an interruption of service during this period.

"The MEP systems will support growth of cabinet densities and changing IT configurations without impacting ongoing IT operations. A typical pod will consist of approximately 32 cabinets (16 per side) with hot aisle containment at 4'-0" and 5'-0" cold aisle. Each pod will be sized for a total of 340kW, but contain a mix of low, medium and high density cabinets. Initial cabinet densities will be a mix of 2 – 20kW/cab, with the ability to scale upwards to 30 – 40kW/cab in the future."

The data center will also be used by the US Marshal Service, the District Attorney, the Bureau of Prisons, and the Treasury Department. "This building, while in Pocatello, Idaho, will be the center and epicenter of what we are doing in Washington, DC," Jeffrey Sallet, FBI associate executive assistant director of finance and facilities, said at a groundbreaking attended by East Idaho News.

The FBI also operates data centers in Clarksburg, West Virginia and Sterling, Virginia, after the DOJ consolidated nearly 100 data centers (including small server rooms) across the Department.

"The consolidation of dozens of data centers will strengthen the cybersecurity posture for all [Department of Justice] components using the facility," the FBI said in a statement.

POCATELLO, Idaho (AP) — The FBI and U.S. Department of Justice held a ribbon-cutting ceremony for a new data center in eastern Idaho.

Officials say the expansion celebrated Monday adds 140,000 square feet (13,000 square meters) to the FBI's facility in Pocatello.

Officials say the expansion is part of the FBI's technology transformation initiative and will consolidate nearly 100 data centers throughout the Justice Department.

The expansion is expected to bolster cybersecurity for the Justice Department, enhance collaboration and the use of advanced analytics.

Construction on the $100 million data center expansion began in late 2017.

The FBI has had a facility in Pocatello for more than three decades.

The center provides telecommunications and database support for agency field services throughout the country.

FBI expands data center in eastern Idaho

By

The Associated Press

The Associated Press

POCATELLO, Idaho (AP) — The FBI and U.S. Department of Justice held a ribbon-cutting ceremony for a new data center in eastern Idaho.

Officials say the expansion celebrated Monday adds 140,000 square feet (13,000 square meters) to the FBI's facility in Pocatello.

Officials say the expansion is part of the FBI's technology transformation initiative and will consolidate nearly 100 data centers throughout the Justice Department.

The expansion is expected to bolster cybersecurity for the Justice Department, enhance collaboration and the use of advanced analytics.

Construction on the $100 million data center expansion began in late 2017.

The FBI has had a facility in Pocatello for more than three decades.

The center provides telecommunications and database support for agency field services throughout the country.

Established the International Bomb **Data Center** Sub-Special Interest Group ... dispute between the **FBI** and ATF over the roles of their Bomb **Data Centers**

EXPLOSIVES INVESTIGATION COORDINATION BETWEEN THE FEDERAL BUREAU OF INVESTIGATION AND THE BUREAU OF ALCOHOL, TOBACCO, FIREARMS AND EXPLOSIVES

U.S. Department of Justice
Office of the Inspector General
Audit Division

Audit Report 10-01
October 2009

Training

A lack of training by ATF has also hindered state and local bomb squads from using BATS. The BATS user manual and Memorandum of Understanding make user agencies ultimately responsible for the training of both current and any replacement personnel. In May 2008 ATF began hosting training seminars for state and local bomb squads. The presentation walks users through specific steps to document and enter example incidents into the system using a terminal connected to BATS. As of March 2009, ATF had provided BATS training to approximately 1,132 bomb technicians and investigators at 39 training seminars.

Based on our state and local bomb squad commander interviews and survey results, we determined that the extent of future BATS use by local first responders depends on ATF's ability to make BATS more user-friendly and to provide effective training. Several commanders indicated that ATF had already made revisions to BATS that improved its ease of use, although they said that the extent of the information required remains a hindrance.[46] To help enhance state and local bomb squads use of BATS, we recommend that ATF prioritize remedial and new user BATS training for federal, state, and local users. We also recommend that ATF review ways to more efficiently provide user names and passwords, and reassess the need to require the current extent of incident information.

DOJ Review of Explosives Databases

The 2004 Memorandum required the DOJ Office of the Chief Information Officer (OCIO) to examine and report to the Deputy Attorney General on the feasibility of consolidating DOJ's arson and explosives databases. DOJ's OCIO conducted a feasibility study in November 2004 and concluded that DOJ needs to develop a more coordinated program for access to arson and explosives information, resources, and technologies. However, DOJ has not adequately addressed the proliferation of explosives related databases within the FBI and ATF.

Competing Databases and Bomb Data Centers

The FBI and ATF maintain Bomb Data Centers that provide technical information to the domestic and international bomb technician community on explosives devices. Additionally, these centers maintain separate international explosives incident databases. The FBI Bomb Data Center

[46] Bomb Commanders indicated that revisions made it easier to enter incidents into BATS.

established the International Bomb Data Center Sub-Special Interest Group (IBDC) in 2006 as a clearinghouse for information on international bombing incidents and new render-safe technologies. Likewise, ATF's Bomb Data Center maintains an international explosives incident database called DFuze to collect, analyze, report, and share explosives device and explosives incident data with its international partners. But we found that operation of separate international databases can cause confusion. For example, a recent request by a foreign government for information on components typically used in certain types of IEDs resulted in the FBI and ATF sending separate and uncoordinated responses.

The 2007 ERG report to Deputy Attorney General McNulty highlighted the dispute between the FBI and ATF over the roles of their Bomb Data Centers and included white papers outlining each agency's position. ATF contended that its U.S. Bomb Data Center had all relevant bomb incident data and merited full DOJ support as the only Bomb Data Center maintained by the federal government. The FBI contended that its counterterrorism mission would be adversely affected if its explosives information-sharing responsibilities were transferred or diluted and argued that only the FBI should use the Bomb Data Center name and interact internationally on explosives matters. The Office of the Deputy Attorney General never resolved this dispute.

Moreover, in 2006 the ATF Laboratory began using an IED Component Database (IED DB) designed to assist laboratory technicians and investigators in the description, storage, and retrieval of IED components. However, other databases could potentially perform this function, including the FBI's Expert Reference Tool (EXPeRT) forensic laboratory database and possibly ATF's BATS.[47] As shown in Exhibit 3-5, the FBI and ATF each maintain disparate databases for managing laboratory forensic reports, incident reporting, and technical explosives-related information and intelligence.

[47] Using the IED-DB, investigators can perform queries of specific explosives incidents using descriptive attributes of the components within an IED. However, we note that BATS already provides a forum for the description of simple components of an IED as well as events surrounding explosives incidents.

EXPLOSIVES INVESTIGATION COORDINATION BETWEEN THE FEDERAL BUREAU OF INVESTIGATION AND THE BUREAU OF ALCOHOL, TOBACCO, FIREARMS AND EXPLOSIVES

EXECUTIVE SUMMARY

The Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) and the Federal Bureau of Investigation (FBI) share jurisdiction for investigating federal explosives crimes.[1] Despite attempts at coordination and division of jurisdiction, these components have historically developed separate and often conflicting approaches to explosives investigations and related activities such as explosives training, information sharing, and forensic analysis.

OIG Audit Approach

This Office of the Inspector General (OIG) audit examined the Department of Justice (DOJ) oversight and the FBI's and ATF's operations of explosives-related activities from fiscal year (FY) 2003 through April 2009. The objective of this review was to evaluate the level of coordination between the FBI and ATF in explosive investigations, including the effectiveness of explosive incident protocols to determine lead agency jurisdiction, the extent of information sharing and consolidation of explosives data, the degree of training coordination, and the use of laboratory resources for explosives analysis.[2]

We conducted interviews with over 100 ATF and FBI employees, including personnel from both components' headquarters and in 8 locations with both ATF and FBI field divisions or offices. Our FBI interviews included officials in the Counterterrorism Division, Criminal Investigative Division, Laboratory Division, and Critical Incident Response Group. ATF interviews included officials from the Office of Field Operations, National Center for Explosive Training and Research, Explosive Enforcement Division, U.S. Bomb

---

[1] See Appendix III for a more detailed description of FBI and ATF explosives-investigation authorities.

[2] Explosives incidents include: actual or attempted bombings with improvised (i.e., homemade), misused commercial or military, or incendiary explosives devices (e.g., gasoline bombs); recovered explosives and explosives components; render-safe operations to dispose of or destroy recovered explosives; accidental detonations of commercial explosives; threats to bomb; and the use of hoax devices.

Data Center, and Laboratory. We also conducted interviews with staff in the Office of the Deputy Attorney General and the DOJ Office of the Chief Information Officer.

We reviewed policies and procedures for explosives investigations and interagency interactions, Special Agent training records, and explosives-related course curricula. Additionally, we analyzed case data for FY 2003 through mid-FY 2008 from both components' case management systems and the Bombing and Arson Tracking System (BATS).[3] We also developed and sent survey questionnaires to all ATF and FBI Field Division Special Agents-in-Charge, ATF and FBI Special Agent Bomb Technicians, ATF Certified Explosive Specialists, ATF Explosives Enforcement Officers, and the commanders of each accredited state and local bomb squad in the country.

Appendix I contains a more detailed description of our audit objectives, scope, and methodology.

Results in Brief

Our audit found that the FBI and ATF are not adequately coordinating explosives-related operations and have developed similar technical abilities to respond to explosives incidents. Conflicts continue to occur throughout the country regarding which agency is the lead agency for federal explosives investigations and their differing explosives-handling techniques. These disputes can cause confusion for local first responders about the roles of the FBI and ATF during explosives-incident responses and delays in conducting investigations. Disputes between ATF and FBI personnel have affected working relationships, and in some locations have resulted in their racing to crime scenes to determine which agency leads an investigation. The disputes have also resulted in the two agencies declining to work and train together.

In 2004, DOJ attempted to define the roles of the FBI and ATF in explosives investigations and related activities through an Attorney General Memorandum.[4] However, this 2004 memorandum contained ambiguous directives for determining lead agency authority for explosives matters. In addition, DOJ, the FBI, and ATF did not implement the memorandum's procedures for explosives information sharing and database consolidation, training, and laboratory resources.

---

[3] BATS is ATF's automated incident reporting system developed to streamline the gathering, retrieving, reporting, and archiving of investigative information of fires and the criminal misuse of explosives.

[4] See Appendix IV for the 2004 Attorney General Memorandum.

While the Report addresses coordination challenges, it is important to recognize some of the successes and joint efforts between the ATF and FBI. From 2003 through 2008, ATF and FBI jointly investigated and recommended for prosecution 192 explosives related cases involving 397 defendants. In addition, prior to the audit period, the ATF recognized on its own some of the highlighted issues and began a process to improve the use and function of the Bombing and Arson Tracking System (BATS). In the past year, over 3,000 bomb technicians and investigators have received in person BATS training, and the numbers of agencies and individual users registered in BATS have increased significantly thus facilitating greater information sharing.

It is also important to note that the Joint Program Office (JPO), which is comprised of both the FBI and ATF, has been successful in resolving the types of issues raised in this Report. For example, the JPO coordinated the development of community-wide consensus standards for uniform training of explosive-detection canine teams, which will be published in a guidelines document for implementation nationwide. Another example of joint coordination is the Terrorist Explosive Device Analytical Center (TEDAC). Through TEDAC, the leadership of the FBI and ATF laboratories meets regularly to address inter-component issues. Although the FBI and ATF laboratories each use their own platforms to manage their laboratory's forensic reports and explosives reference material, the systems have been adapted so that both FBI and ATF laboratory information is available to TEDAC partners.

The Report suggests that the Department should create, or at least study further, the creation of a single "One DOJ" data environment to serve as the repository of all automated information about explosives. With regard to the issue of consolidated databases, it is important to point out that a "one size fits all" approach is not necessarily the best utilization of resources when it comes to how the Department carries out its explosives responsibilities. The decision to consolidate any database should be a business decision driven by the value such a consolidation will provide to the user community at large. If DOJ senior leadership, agents, analysts and other related users believe their law enforcement work is better supported by separate databases, improvements should be evaluated to make the existing databases accessible to all users and the information therein comprehensive.

In conclusion, based upon a review of the Report, the ODAG agrees in concept with all 15 recommendations directed to the ATF, FBI, ODAG and OCIO and is in the process of formulating measures to resolve many of the identified issues. The ODAG appreciated the professionalism exhibited by your staff in working jointly with our combined representatives to complete this audit.

Please feel free to contact me should you have any questions.

APPENDIX VIII

CONSOLIDATED DOJ RESPONSE TO AUDIT REPORT RECOMMENDATIONS

U.S. Department of Justice

Office of the Deputy Attorney General

Associate Deputy Attorney General

Washington, D.C. 20530

October 9, 2009

MEMORANDUM

TO: Raymond J. Beaudet
Assistant Inspector General for Audit
Office of the Inspector General

FROM: Edward N. Siskel
Associate Deputy Attorney General
Office of the Deputy Attorney General

SUBJECT: Joint Response to the OIG's Draft Report: "Explosives Investigation Coordination between the Federal Bureau of Investigation and the Bureau of Alcohol, Tobacco, Firearms, and Explosives"

The Office of the Deputy Attorney General (ODAG) very much appreciates the opportunity to review and respond to the Office of Inspector General's draft audit report entitled, "Explosives Investigation Coordination between the Federal Bureau of Investigation and the Bureau of Alcohol, Tobacco, Firearms, and Explosives" (hereinafter, "Report"). As per your office's request, this memorandum will serve as a joint response to the Report on behalf of the Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF), the Federal Bureau of Investigation (FBI), the Office of the Chief Information Officer (OCIO), and the ODAG.

The Department of Justice recognizes the critical importance of a well-coordinated and effective response to explosives incidents. Equally important is the need to adequately train our personnel and ensure effective information sharing with all appropriate components and law enforcement partners.

The Report documents the Department's challenges concerning the most efficient application and balance of its explosives enforcement assets and responsibilities and offers some remedies to those challenges. The Department agrees with the concept of the recommendations that are reflected in the body of the Report and is taking steps to address each of those recommendations. At the same time, we recognize, and have discussed with your staff, that we may have modifications to how we go about implementing those recommendations in order to achieve the most successful and efficient resolution to the matters under review.

APPENDIX VII

MEMORANDUM FROM A UNITED STATES ATTORNEY

Memorandum

U.S. DEPARTMENT OF JUSTICE

Eric F. Melgren
United States Attorney
District of Kansas

Subject	Date
Attorney General's Memorandum on Explosives Investigations	September 10, 2004

To	From
S.A.C. Mark James S.A.C. Kevin Stafford cc: Shawn Stroud Tom Atteberry	Eric F. Melgren

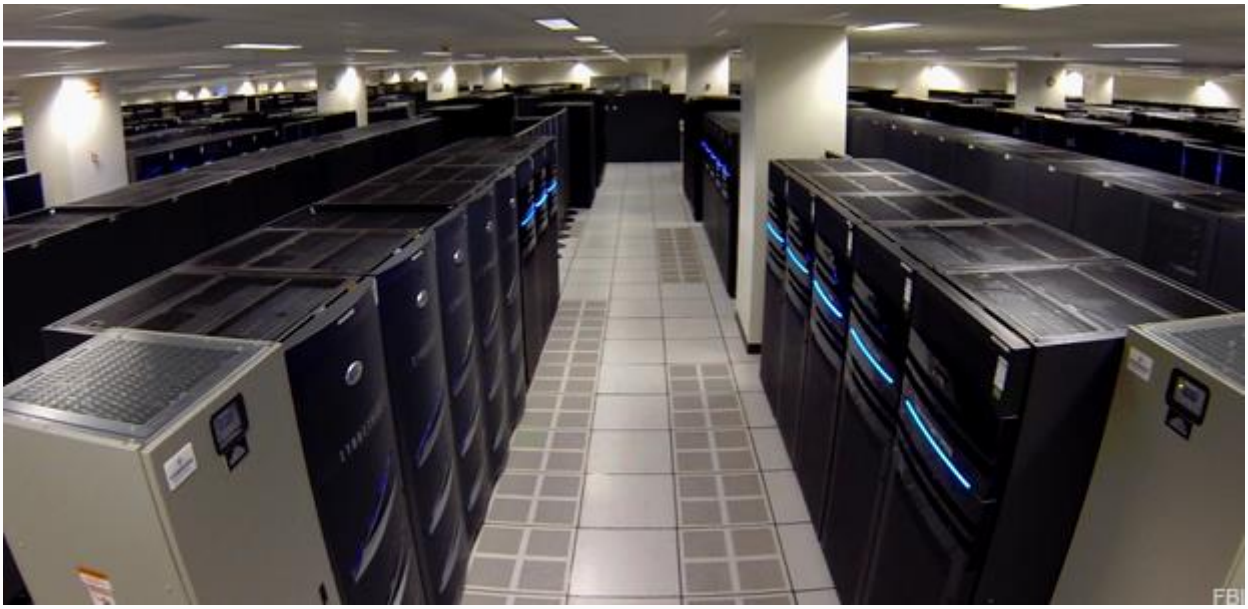
Because my meeting with Stroud and Attebury was less than successful in resolving misunderstandings between the agencies, pursuant to paragraph 6 of the memorandum I contacted Deputy Attorney General James Comey for guidance. Uttam Dhillon of his office, who had principal responsibility for implementing this memorandum, visited with me about these issues. His direction was as follows:

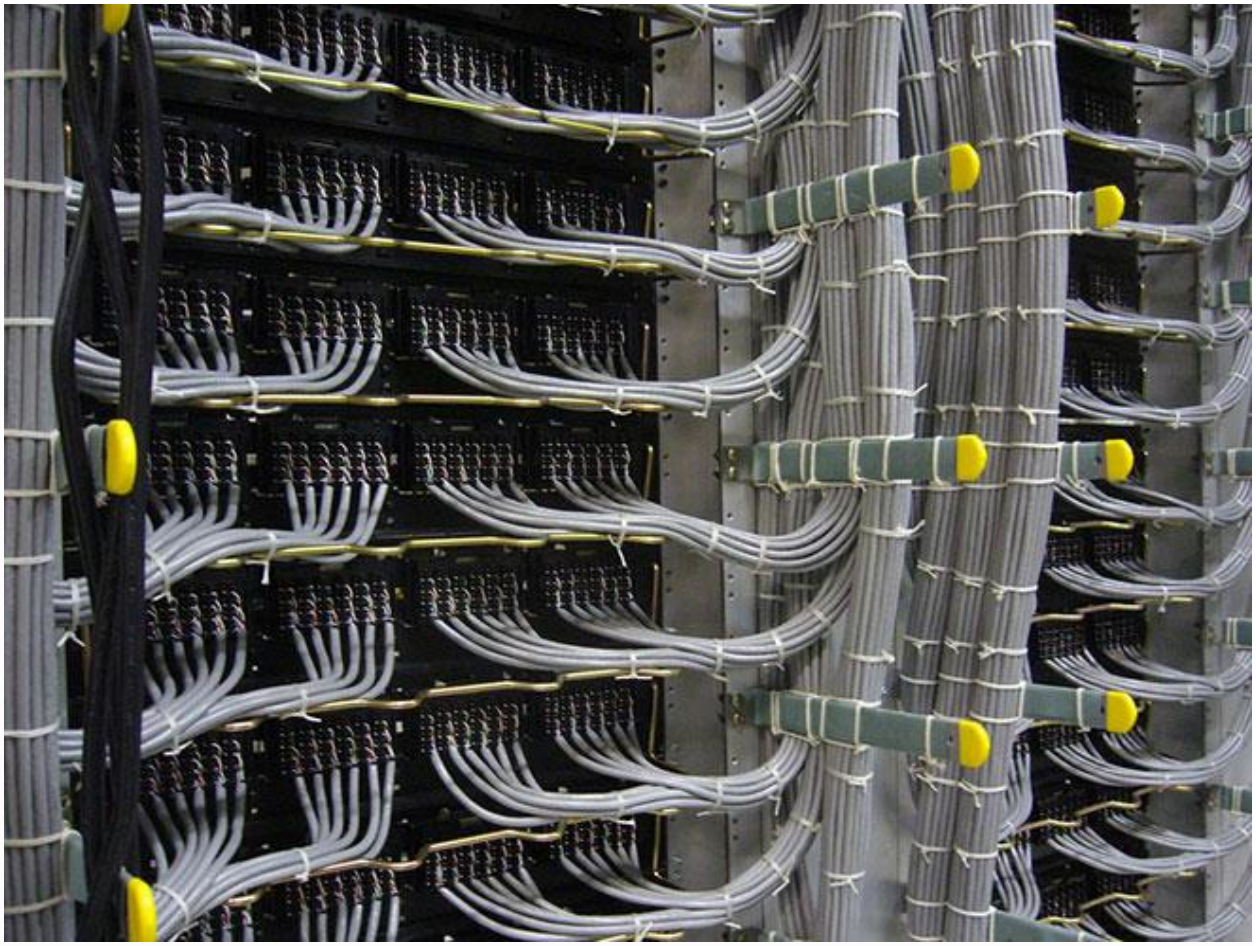
1. Pursuant to the memorandum, the ATF has jurisdiction over explosives issues unless one of the stated exceptions applies. It is more accurate to say that this means the ATF has jurisdiction unless and until it is determined that the explosives incident is related to terrorism; than it is to say that the FBI or the JTTF has jurisdiction unless and until it is determined that the incident is not related to terrorism. As Mr. Dhillon expressed it to me, the FBI must assert itself in order to acquire jurisdiction over the incident. My understanding, with which he agreed, is that "default" jurisdiction is with the ATF.

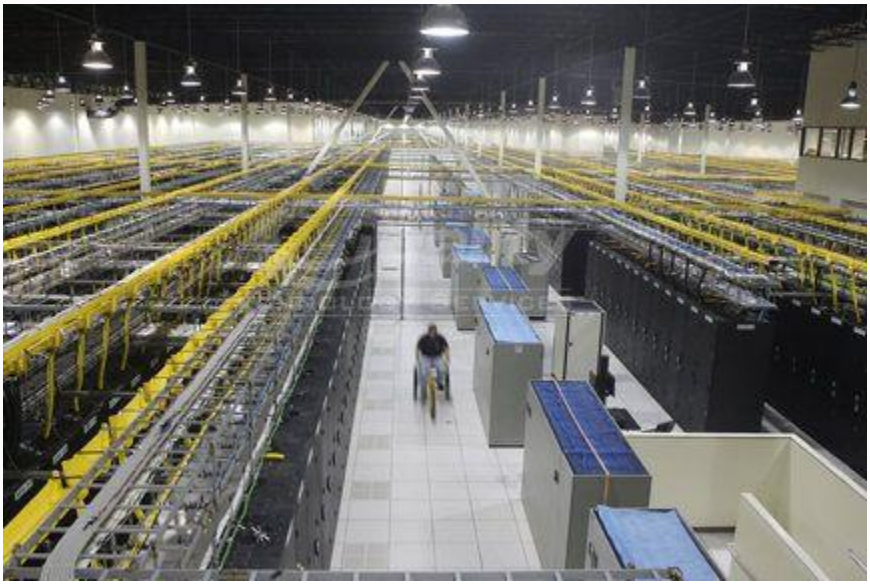
2. I inquired about the language in paragraph 1.a. stating that the "JTTF determines" the incident is related to terrorism. Practically speaking, I asked, doesn't this mean the FBI; since the FBI is the lead agency for the JTTFs? He responded that the JTTFs were selected to ensure that the decision was made by a joint task force with primary responsibility for investigation terrorism cases.

Prior to circulating this memo to you, I vetted it by Mr. Dhillon, who has agreed that it accurately represents his direction.

FBI DATA CENTERS PHOTO GALLERY







1788

5698TFD-PWR

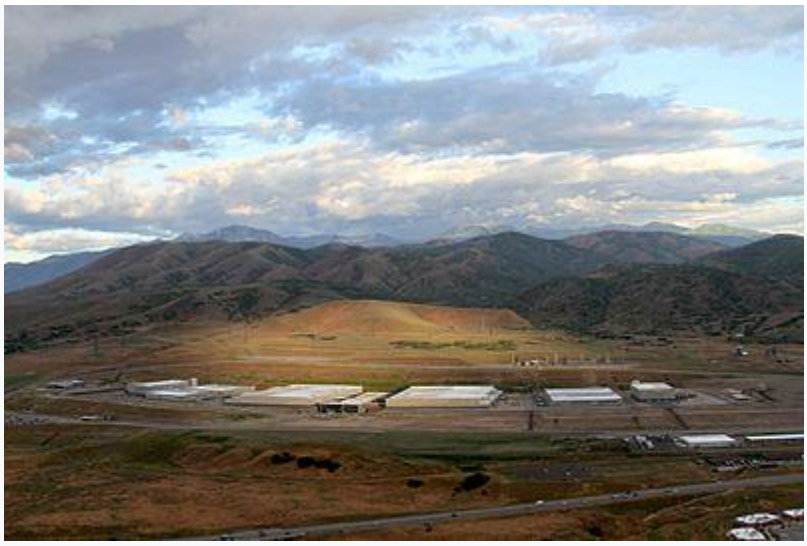
Bates St

CENTRAL INTELEGENCE AGENCEY (CIA)

Utah Data Center

From Wikipedia, the free encyclopedia

Jump to navigationJump to search



NSA's Utah Data Center

The **Utah Data Center** (UDC), also known as the **Intelligence Community Comprehensive National Cybersecurity Initiative Data Center**,[1] is a data storage facility for the United States Intelligence Community that is designed to store data estimated to be on the order of exabytes or larger.[2] Its purpose is to support the Comprehensive National Cybersecurity Initiative (CNCI), though its precise mission is classified.[3] The National Security Agency (NSA) leads

operations at the facility as the executive agent for the Director of National Intelligence.[4] It is located at Camp Williams near Bluffdale, Utah, between Utah Lake and Great Salt Lake and was completed in May 2019 at a cost of $1.5 billion.[5]

The Utah Data Center, code-named Bumblehive, is the first Intelligence Community Comprehensive National Cyber-security Initiative (IC CNCI) data center designed to support the US intelligence community. The "massive data repository" is designed to cope with the large increase in digital data that has accompanied the rise of the global internet.[6]

Purpose

The data center is able to process "all forms of communication, including the complete contents of private emails, cell phone calls, and Internet searches, as well as all types of personal data trails — parking receipts, travel itineraries, bookstore purchases, and other digital 'pocket litter'."[7] In response to claims that the data center would be used to illegally monitor email of U.S. citizens, in April 2013 an NSA spokesperson said, "Many unfounded allegations have been made about the planned activities of the Utah Data Center, ... one of the biggest misconceptions about NSA is that we are unlawfully listening in on, or reading emails of, U.S. citizens. This is simply not the case."[4] This statement was made two months prior to the document leak that revealed the existence of the PRISM program (below). Some members of the public question the credibility of official statements like Bamford's, above, in light of the revelations made in the following two paragraphs herein, and in light of the revelations made in the following two months after his statement was made.

Utah Data Center, Bluffdale, Utah (United States)

In April 2009, officials at the United States Department of Justice acknowledged that the NSA had engaged in large-scale overcollection of domestic communications in excess of the United States Foreign Intelligence Surveillance Court's authority, but claimed that the acts were unintentional and had since been rectified.[8]

In August 2012, *The New York Times* published short documentaries by independent filmmakers titled *The Program*,[9] based on interviews with former NSA technical director and whistleblower William Binney. The project had been designed for foreign signals intelligence (SIGINT) collection but, Binney alleged, after the September 11, 2001, terrorist attacks, controls that limited unintentional collection of data pertaining to U.S. citizens were removed, prompting concerns by him and others that the actions were illegal and unconstitutional. Binney alleged that the Bluffdale facility was designed to store a broad range of domestic communications for data mining without warrants.[10]

Documents leaked to the media in June 2013 described PRISM, a national security computer and network surveillance program operated by the NSA, as enabling in-depth surveillance on live Internet communications and stored information.[11][12] Reports linked the data center to the NSA's controversial expansion of activities, which store extremely large amounts of data. Privacy and civil liberties advocates raised concerns about the unique capabilities that such a facility would give to intelligence agencies.[13][14] "They park stuff in storage in the hopes that they will eventually have time to get to it," said James Lewis, a cyberexpert at the Center for Strategic and International Studies, "or that they'll find something that they need to go back and look for in the masses of data." But, he added, "most of it sits and is never looked at by anyone."[15]

The UDC was expected to store Internet data, as well as telephone records from the controversial NSA telephone call database, MAINWAY, when it opened in 2013.[16]

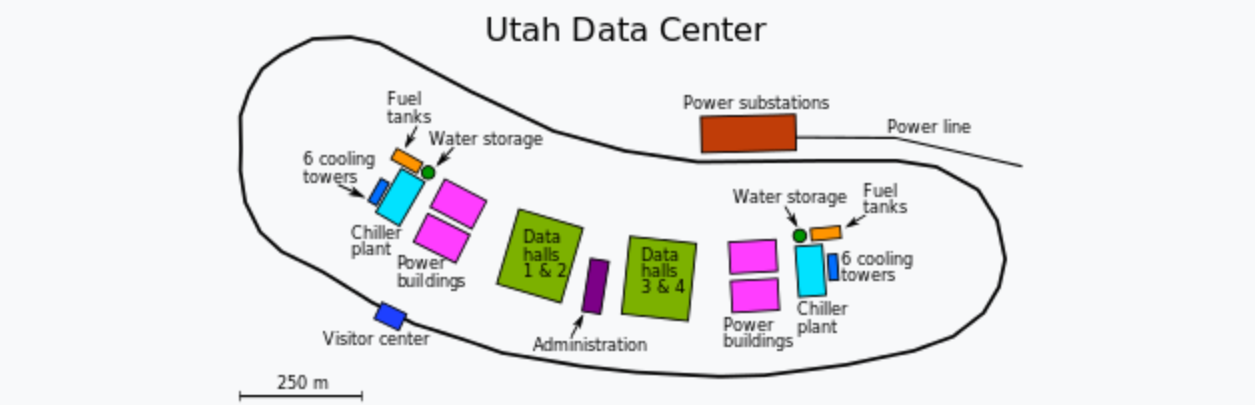
In light of the controversy over the NSA's involvement in the practice of mass surveillance in the United States, and prompted by the 2013 mass

surveillance disclosures by ex-NSA contractor Edward Snowden, the Utah Data Center was hailed by *The Wall Street Journal* as a "symbol of the spy agency's surveillance prowess".[17]

Binney has said that the facility was built to store recordings and other content of communications, not only for metadata.[18]

According to an interview with Snowden, the project was initially known as the Massive Data Repository within NSA, but was renamed to Mission Data Repository due to the former sounding too "creepy".[19]

Structure



Utah Data Center area layout

The planned structure provides 1 to 1.5 million square feet (90,000–140,000 m^2),[20][21][22] with 100,000 square feet (9,000 m^2) of data center space and more than 900,000 square feet (84,000 m^2) of technical support and administrative space.[7][20] It is projected to cost $1.5–2 billion.[3][7][20][23][24] A report suggested that it will cost another $2 billion for hardware, software, and maintenance.[20] The completed facility is expected to require 65 megawatts of electricity, costing about $40 million per year.[7][20] The facility is expected to use 1.7 million gallons (6,435 m^3) of water per day.[25] An article by *Forbes* estimates the storage capacity as between 3 and 12 exabytes in the near term, based on analysis of unclassified blueprints, but mentions Moore's Law, meaning that advances in technology could be expected to increase the capacity by orders of magnitude in the coming years.[2]

Toward the end of the project's construction it was plagued by electrical problems in the form of "massive power surges"[26] that damaged equipment.[17] This delayed its opening by a year.[26]

References

1. ^ *"NSA Utah Data Center". Facilities Magazine. 14 September 2011. Retrieved 24 April 2013.*
2. ^ Jump up to:[a] [b] *Kashmir Hill (July 24, 2013). "Blueprints Of NSA's Ridiculously Expensive Data Center In Utah Suggest It Holds Less Info Than Thought". Forbes. Retrieved 2013-07-29.*
3. ^ Jump up to:[a] [b] *Fidel, Steve (6 January 2011). "Utah's $1.5 billion cyber-security center under way". Deseret News. Retrieved 29 March 2013.*
4. ^ Jump up to:[a] [b] *Shalal-Esa, Andrea (15 April 2013). "U.S. agency denies data center to monitor citizens' emails". Reuters. Retrieved 16 April 2013.*
5. ^ *Office of the Under Secretary of Defense for Acquisition, Technology and Logistics (September 17, 2014). "MilCon Status Report - August, 2014 - Under Secretary of Defense for AT&L". Archived from the original on December 10, 2014. Retrieved April 16, 2015.*
6. ^ *"NSA Utah Data Center - Serving Our Nation's Intelligence Community". nsa.gov1.info. Retrieved 2019-02-14.*
7. ^ Jump up to:[a] [b] [c] [d] *Bamford, James (15 March 2012). "The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)". Wired. Retrieved 5 April 2012.*
8. ^ *Eric Lichtblau and James Risen (April 15, 2009). "Officials Say U.S. Wiretaps Exceeded Law". The New York Times. Retrieved September 5,2012.*
9. ^ *Poitras, Laura (August 22, 2012). "The Program". The New York Times. Retrieved August 5, 2016.*
10. ^ Lawson, Kent, "What Does the NSA Know About You?", Private WiFi, August 27, 2012
11. ^ *Gellman, Barton; Poitras, Laura (June 6, 2013). "US Intelligence Mining Data from Nine U.S. Internet Companies in Broad Secret Program". The Washington Post. Retrieved June 11, 2013.*
12. ^ *Greenwald, Glenn; MacAskill, Ewen (June 9, 2013). "Edward Snowden: The Whistleblower Behind Revelations of NSA Surveillance". The Guardian. Hong Kong. Retrieved June 9, 2013.*
13. ^ *James Risen & Eric Lichtblau (June 8, 2013). "How the U.S. Uses Technology to Mine More Data More Quickly". The New York Times. Retrieved June 11, 2013.*
14. ^ *Howard Berkes (June 10, 2013). "Amid Data Controversy, NSA Builds Its Biggest Data Farm". National Public Radio. Retrieved June 11, 2013.*
15. ^ Scott Shane and David E. Sanger, "Job Title Key to Inner Access Held by Snowden", *New York Times*, June 30, 2013

16. ^ *Thomas Burr (June 6, 2013). "Phone records could end up at NSA's Utah Data Center". Salt Lake Tribune. Retrieved June 11, 2013.*
17. ^ Jump up to: [a] [b] *Siobhan Gorman (7 October 2013). "Meltdowns Hobble NSA Data Center". The Wall Street Journal. Retrieved 19 October 2013. The Utah facility, one of the Pentagon's biggest U.S. construction projects, has become a symbol of the spy agency's surveillance prowess, which gained broad attention in the wake of leaks from NSA contractor Edward Snowden.*
18. ^ *"NSA Collects 'Word for Word' Every Domestic Communication, Says Former Analyst". PBS NewsHour. 1 August 2013.*
19. ^ *James Bamford (August 2014). "Edward Snowden: The Untold Story". Wired.*
20. ^ Jump up to: [a] [b] [c] [d] [e] *Kenyon, Henry (Jan 7, 2011). "New NSA data center breaks ground on construction -- Defense Systems". Defense Systems. Retrieved 11 August 2011.*
21. ^ *"NSA to store yottabytes in Utah data centre". CNET Networks. Retrieved 11 August 2011.*
22. ^ *Bamford, James. "Who's in Big Brother's Database? by James Bamford". The New York Review of Books. Retrieved 11 August 2011.*
23. ^ *LaPlante, Matthew D. (July 2, 2009). "New NSA center unveiled in budget documents". Salt Lake Tribune. MediaNews Group. Retrieved 2009-07-05.*
24. ^ *LaPlante, Matthew D. (July 2, 2009). "Spies like us: NSA to build huge facility in Utah". Salt Lake Tribune. MediaNews Group. Retrieved 2009-07-05.*
25. ^ *Adams, Andrew (July 12, 2013). "New Utah NSA center requires 1.7M gallons of water daily to operate". Retrieved 2013-07-25.*
26. ^ Jump up to: [a] [b] *"NSA data center 'crippled' by huge power surges". DatacenterDynamics. October 8, 2013. Retrieved April 27, 2014.*

The Details about the CIA's Deal with Amazon

A $600 million computing cloud built by an outside company is a "radical departure" for the risk-averse intelligence community.

The intelligence community is about to get the equivalent of an adrenaline shot to the chest. This summer, a $600 million computing cloud developed by Amazon Web Services for the Central Intelligence Agency over the past year will begin servicing all 17 agencies that make up the intelligence community. If the technology plays out as officials envision,

it will usher in a new era of cooperation and coordination, allowing agencies to share information and services much more easily and avoid the kind of intelligence gaps that preceded the Sept. 11, 2001, terrorist attacks.

For the first time, agencies within the intelligence community will be able to order a variety of on-demand computing and analytic services from the CIA and National Security Agency. What's more, they'll only pay for what they use.

The vision was first outlined in the Intelligence Community Information Technology Enterprise plan championed by Director of National Intelligence James Clapper and IC Chief Information Officer Al Tarasiuk almost three years ago. Cloud computing is one of the core components of the strategy to help the IC discover, access and share critical information in an era of seemingly infinite data.

For the risk-averse intelligence community, the decision to go with a commercial cloud vendor is a radical departure from business as usual.

In 2011, while private companies were consolidating data centers in favor of the cloud and some civilian agencies began flirting with cloud variants like email as a service, a sometimes contentious debate among the intelligence community's leadership took place.

As one former intelligence official with knowledge of the Amazon deal told Government Executive, "It took a lot of wrangling, but it was easy to see the vision if you laid it all out." The critical question was would the IC, led by the CIA, attempt to do cloud computing from within, or would it buy innovation? Money was a factor, according to the intelligence official, but not the leading one.

The government was spending more money on information technology within the IC than ever before. IT spending reached $8 billion in 2013, according to budget documents leaked by former NSA contractor Edward Snowden. The CIA and other agencies feasibly could have spent billions of dollars standing up their own cloud infrastructure without raising many eyebrows in Congress, but the decision to purchase a single commercial solution came down primarily to two factors.

"What we were really looking at was time to mission and innovation," the former intelligence official said. "The goal was, 'Can we act like a large enterprise in the corporate world and buy the thing that we don't have, can we catch up to the commercial cycle? Anybody can build a data center, but could we purchase something more?

"We decided we needed to buy innovation," the former intelligence official said.

A Groundbreaking Deal

The CIA's first request for proposals from industry in mid-2012 was met with bid protests to the Government Accountability Office from Microsoft and AT&T, two early contenders for the contract. Those protests focused on the narrow specifications called for by the RFP. GAO did not issue a decision in either protest because the CIA reworked its request to address the companies' complaint.

In early 2013, after weighing bids from Amazon Web Services, IBM and an unnamed third vendor, the CIA awarded a contract to AWS worth up to $600 million over a period of up to 10 years. The deal, handled in secret, was first reported by FCW in March 2013, sending ripples through the tech industry.

A month after the deal became public, IBM filed a bid protest with GAO that the watchdog eventually upheld in June, forcing the CIA to reopen bids to both companies for the contract. A legal struggle between Amazon and Big Blue ensued, and AWS filed a lawsuit against the federal government in July 2013, claiming the GAO sustainment was a "flawed" decision.

In October, U.S. Court of Federal Claims Judge Thomas Wheeler sided with Amazon and overturned GAO's decision to force the CIA to rebid the contract. Big Blue went home, AWS claimed victory under the deal's original financial specs,

and nearly 18 months after the procurement was first released, the CIA and Amazon went to work.

It is difficult to underestimate the cloud contract's importance. In a recent public appearance, CIA Chief Information Officer Douglas Wolfe called it "one of the most important technology procurements in recent history," with ramifications far outside the realm of technology.

"It's going to take a few months to bring this online in a robust way, but it's coming," Wolfe said. "And I think it's going to make a big difference for national security."

Securing New Capabilities

The Amazon-built cloud will operate behind the IC's firewall, or more simply: It's a public cloud built on private premises.

Intelligence agencies will be able to host applications or order a variety of on-demand services like storage, computing and analytics. True to the National Institute of Standards and Technology definition of cloud computing, the IC cloud scales up or down to meet the need.

In that regard, customers will pay only for services they actually use, which is expected to generate massive savings for the IC.

"We see this as a tremendous opportunity to sharpen our focus and to be very efficient," Wolfe told an audience at AWS' annual nonprofit and government symposium in Washington. "We hope to get speed and scale out of the cloud, and a tremendous amount of efficiency in terms of folks traditionally using IT now using it in a cost-recovery way."

Many agencies within the IC already have identified applications to move to the cloud. In a recent report, National Reconnaissance Office Chief Information Officer Donna Hansen said her agency had picked five applications, including its enterprise resource planning software, to migrate to the IC cloud. As with public clouds, the IC cloud will maximize automation and require standardized information, which will be shared through application programming interfaces, known as APIs. Amazon engineers will oversee the hardware because AWS owns the hardware and is responsible for maintaining it just as they do in the company's public data centers.

Whenever Amazon introduces a new innovation or improvement in cloud services, the IC cloud will evolve. Company officials say AWS made more than 200 such incremental improvements last year, ensuring a sort of built-in innovation to the IC cloud that will help the intelligence community keep pace with commercial advances. Wolfe said AWS' capacity to bring commercial innovation from places like Silicon Valley to the IC is one of the contract's greatest

benefits. Whenever AWS introduces new products, the CIA will be able to implement them.

"The biggest thing we were trying to do—the visionary folks a couple years ago—was answer the question, 'How do we keep up?'" Wolfe said. "The mission we have is important. The pace and complexity is really not [diminishing], in fact, it may be increasing. We feel it is very important to deliver the best IT and best products and services we can to our customers in the IC."

What of the data, though? Intelligence agencies are drowning in it, collecting and analyzing an amalgamation of information from sensors, satellites, surveillance efforts, open data repositories and human intelligence, among other sources. Is that data really secure in the cloud?

The CIA is convinced it is.

The IC cloud "will be accredited and compliant with IC standards," says a senior CIA official familiar with the IC cloud. It will, for example, be able to handle Sensitive Compartmented Information, a type of classified information. "Security in the IC cloud will be as safe as or safer than security on our current data centers," the senior CIA official says. Because the IC cloud will serve multiple tenants—the 17 agencies that comprise the IC—administrators will be able to restrict access to information based on the identity of the individual seeking it. The idea is

to foster collaboration without compromising security. Visually, the IC cloud can be thought of as a workspace hanging off the IC's shared network—a place where data can be loaded for a variety of tasks like computing or sharing. The IC cloud gives agencies additional means to share information in an environment where automated security isn't a barrier to the sharing itself. This could prove vital in situations reminiscent of 9/11, in which national security is an immediate concern.

Cloud vendors, including Amazon, have argued that cloud infrastructures can be more secure than traditional data centers because there are fewer points of entry, but the leaks by Snowden illustrate the potential threat from inside an organization. Snowden was able to access and download classified information intelligence officials said he shouldn't have been able to access.

To access information within the IC cloud, analysts must have the proper permissions. In addition, the standardized environment and automation means all activity within the cloud is logged and can be analyzed in near real-time.

Some government officials view cloud computing as inherently less secure than computing on locally controlled servers, but the CIA's acceptance of commercially developed cloud technology "has been a wake-up call" to those who balk at it, according to John Pirc, a former CIA cybersecurity researcher who is now chief technology officer at NSS Labs, a security research firm.

"You hear so many people on the fence about cloud, and then to see the CIA gobble it up and do something so highly disruptive, it's kind of cool," says Pirc. "To me, this removes the clouded judgment that cloud isn't secure. Their moving forward with this should send a message to the rest of the industry that cloud is something you shouldn't be afraid of."

Pirc is no stranger to disruptive technologies. At the CIA's research labs in the early 2000s, he recalls virtualization—a technology that allows multiple operating systems to run simultaneously on the same servers, allowing for far more efficient computing—before it became an integral component of many IT enterprises. Intelligence agencies use commercial off-the-shelf technology all the time, but to Pirc, the importance of the cloud capabilities the CIA gets through leveraging Amazon Web Services' horsepower is best exemplified in computing intelligence data. Scalable computing is critical for fostering shared services and enhanced collaboration between disparate intelligence agencies.

"What it allows them to do is spin up servers and add more [computing power] fast, and when you're computing intelligence data, the more compute power you have, the faster you can react," Pirc says. "In the private sector, compute is all about money and profit, but from my viewpoint when I worked for the agency, you're working with extremely time-sensitive information. Being able to have that

compute power, something that might have taken a couple of hours might instead take a few seconds. Profits aren't lost when you make mistakes in the intelligence community—people die when you make mistakes."

A test scenario described by GAO in its June 2013 bid protest opinion suggests the CIA sought to compare how the solutions presented by IBM and Amazon Web Services could crunch massive data sets, commonly referred to as big data.

Solutions had to provide a "hosting environment for applications which process vast amounts of information in parallel on large clusters (thousands of nodes) of commodity hardware" using a platform called MapReduce. Through MapReduce, clusters were provisioned for computation and segmentation. Test runs assumed clusters were large enough to process 100 terabytes of raw input data. AWS' solution received superior marks from CIA procurement officials, according to GAO documentation, and was one of the chief reasons the agency selected Amazon.

Limited Details

The CIA declined to comment when Government Executive asked about the extent of the IC cloud's capabilities or that of the National Security Agency's cloud. Amazon also declined to describe the IC cloud's technical capabilities.

It is a good bet, though, that the AWS-built cloud for the IC will have capabilities at least equal to existing capabilities Amazon has already implemented across government.

For example, the company provides the cloud bandwidth for the Securities and Exchange Commission's collection of more than 1 billion trade records and more than a terabyte of new data per day through its Market Information Data Analytics System. This example may be prescient given that now-public surveillance efforts indicate the IC collects billions and perhaps trillions of pieces of metadata, phone and Internet records, and other various bits of information on an annual basis. The potential exists for the CIA to become one of AWS' largest customers.

Within the intelligence community, examples abound where the cloud's capabilities could significantly boost the mission.

As the geospatial hub of the community, the National Geospatial-Intelligence Agency ingests, analyzes, metatags and reports all geo-intelligence and multisource content in its flagship program called Map of the World. Geospatial data's importance to the IC has increased in recent years, as evidenced by NGA's nearly $5 billion budget and its staff nearly doubling in size since 2004. For intensive applications like ingesting or analyzing geospatial data, scalable computing could have a significant impact on mission performance. The cloud also could improve the way the agency shares its large data sets.

What the IC has done with cloud is not easily replicable, according to American Council for Technology President Rick Holgate, but it is worth paying attention to.

"The IC has a model other agencies should look to and aspire to in terms of transforming the way they think about delivering services across a large enterprise," Holgate says. "They are looking to common platforms and service delivery models across an entire enterprise, and not just gaining cost efficiencies, but to provide foundational capabilities to really allow it to operate."

Whether or not the IC cloud serves as an example for the rest of government, the CIA's quest to buy innovation will loom large for years to come.

Time to Share

The Intelligence Community Information Technology Enterprise lays out a vision for the IC's 17 agencies to securely discover, access and share information. The goal: greater mission success.

In essence, ICITE (pronounced EYEsite) changes the business model for intelligence agencies by requiring that they share services.

Cloud computing—behind the IC's firewall—is one element of that vision. Think of it as the intelligence community

sharing information behind a walled castle apart from the rest of the world operating on the Internet.

The Central Intelligence Agency and the National Security Agency are leading IC cloud development; the NSA's privately hosted cloud was launched in 2013. Importantly, it was made available to users on new and legacy systems so that personnel on any of its systems could use it. The cloud built by Amazon Web Services at the behest of the CIA will be a shared resource for services on an as-needed basis. Intelligence agencies can order a variety of services like storage, computing, analytics, database services or application hosting. Both clouds will work in complementary fashion, according to senior CIA officials.

Other ICITE components include a common desktop, an IC-wide applications mall, and network requirement and engineering services.

The Defense Intelligence Agency and National-Geospatial Intelligence Agency have piloted shared desktop capabilities across their agencies for several thousand users. Those capabilities eventually will spread throughout the IC.

In essence, ICITE changes the business model for intelligence agencies, mandating shared services. Instead of each agency building out its own systems, select agencies—either one or two of those with larger budgets—are responsible for governing its major components.

The Elephant in the Room

In cloud computing, there's Amazon Web Services and then there's everybody else.

IT research firm Gartner had to rescale its famed infrastructure-as-a- service "Magic Quadrant" study in 2013 to accommodate Amazon Web Services' enormous competitive lead.

The quadrant ranks cloud providers based on their ability to execute operations and the comprehensiveness of their vision. For several years there hasn't been even a close challenger to AWS. Gartner's 2014 quadrant shows that AWS captures 83 percent of the cloud computing infrastructure market.

In the combined cloud markets for infrastructure and platform services, hybrid and private clouds—worth a collective $131 billion at the end of 2013—Amazon's revenue grew 67 percent in the first quarter of 2014, according to Gartner.

While the public sector hasn't been as quick to capitalize on cloud computing as the private sector, government spending on cloud technologies is beginning to jump.

Researchers at IDC estimate federal private cloud spending will reach $1.7 billion in 2014, and $7.7 billion by 2017. In other industries, software services are considered the leading cloud technology, but in the government that honor goes to

infrastructure services, which IDC expects to reach $5.4 billion in 2017.

In addition to its $600 million deal with the CIA, Amazon Web Services also does business with NASA, the Food and Drug Administration and the Centers for Disease Control and Prevention. Most recently, the Obama Administration tapped AWS to host portions of HealthCare.gov.

CIA could be planning billions in cloud contracts, open bidding beyond just AWS

Surveillance capitalism and surveillance classic, working together

The Central Intelligence Agency could be planning to solicit tens of billions of dollars of cloud contracts from multiple providers.

Way back in 2013, the CIA signed a huge 10-year deal with Amazon Web Services for cloud services.

Beyond the immediate financial benefits of the deal, at a time when the cloud was far smaller, AWS was also able to point to the contract as proof that its services were secure - after all, if one of the world's largest foreign intelligence services trusted AWS, then your regional enterprise could probably trust it, too. The deal was also used by AWS to help cement its place as the foremost provider for the US government, and parlayed into numerous other federal contracts.

Now, the opportunity could be there for other cloud providers to get a piece of the pie.

CIA

About Data Center Resources

In the data center market, things are quick to change.

Yesterday's solution may not solve today's problem. Our leadership team has a deep industry knowledge, we constantly identify trends and shifts in the data center environment to incorporate into our superior solutions.

Leadership

Justin Tocco

PRINCIPAL

As a co-founding partner of DCR, Justin provides project management services within the data center and network operation center environments. Justin has extensive knowledge of data center construction, cooling and power. In addition to data centers, Justin has a passion for the Network Operations Center/Process Control Console environment. He has supported the design of thousands of large and small control rooms for private and federal government projects. Justin has helped to establish DCR's global reach in countries such as the United States, Canada, Europe, South America, China and more.

Rick Berendes

PRINCIPAL

Rick has been providing power, cooling and rack-based solutions to the data center market since 1995. In 2002, Rick a co-founded Data Center Resources, LLC. Initially, the company started as a value-added-reseller of technical infrastructure solutions

such as power distribution, server cabinets, remote access solutions and network operations furniture. In the following years, Rick and his co-founder Justin noticed gaps in what our customers were looking for and what was available. To serve these needs they started to develop new products. These proprietary product lines have grown to be the mainstay of the company.

Kevin Macomber

SALES ENGINEERING MANAGER

Kevin is a Boston, MA resident and US Navy veteran who held numerous qualifications and was a battle engineer. After graduating from the College of Charleston, he used his background and worked extensively in power distribution product development (Cooper Industries; Cyberex/United Power; and Eaton) and then subsequently moved into sales engineering and project management. With over 25 years of industry experience, he has expertise in supplying mission critical power, containment and other solutions for data center design and construction. Kevin has provided solutions for some of the largest data center and colocation facilities in North America.

Kim Forbis

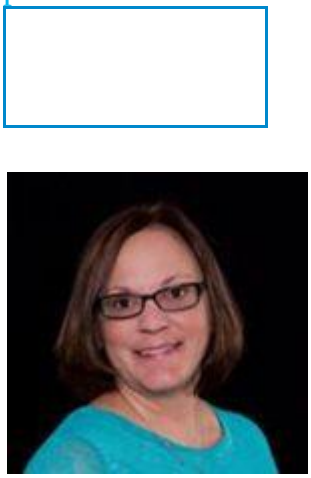
ACCOUNTING & OPERATIONS MANAGER

Kim overseas the critical ‘behind the scenes’ support to include Accounting, Finance and our Buffalo manufacturing facility. Her decades of work experience ties together all of the facets a customer’s experience and production execution. It is not uncommon for her to be involved with larger customer projects to ensure we meet or exceed the complex expectations.

Joe Gutzler

RESELLER CHANNEL

Joe, who was formerly with Liebert, overseas our channel partner program in North America. Joe has a long history of managing and working alongside channel partners to grow their business. He coordinates the resale of both our Inracks command control room console product line as well as our extensive Cool Shield aisle containment portfolio.



Jamie Peake

PROJECT MANAGER – MIDWEST REGION

Jamie, formally with Geist Manufacturing, one of the largest enclosure power distribution companies in the US, has over 10 year's industry experience. With extensive application engineering and product implementation in both data center enclosure power distribution and data center cooling solutions, she is an integral part of our field applications, trouble shooting and project management team. Jamie has significant experience working with OEM's and major accounts, giving her a broad understanding of the complexities of large data center projects. Jamie is based in Lincoln Nebraska.

About Data Center Resources

Our Project Managers are knowledgeable on current trends AND what really works.

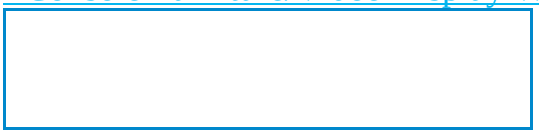
Data Center Resources strives to surpass our customer's expectations in all areas and prides itself on creating innovative solutions. Responsiveness, after sale service, ease of purchasing and breadth of product offerings are other important factors. We are a satisfaction-driven company who is dedicated to bringing a measure of simplicity to the hectic demands our customers face. Our principals, project managers and engineers each have years of experience in providing technical environment solutions. Utilize our unique solutions and experience to assure that your next project is a success.

The data center market is a fast-paced industry that is constantly changing.

Data Center Resources ensures that you have the best available solution with high reliability, efficiency and overall satisfaction.

Our project specialists will visit your site and recommend solutions to improve power and efficiency. We offer the following products and services:

• Power Management
• Aisle Containment/Cooling Optimization
• Server Rack Infrastructure
• Data Center Monitoring
• Data Center Flooring
• Console Furniture/Video Display Walls/Services



Cities, States, and Areas Where We Provide Data Center Aisle Containment Products and Installation Services:

Alabama

- Birmingham, AL
- Mobile, AL
- Huntsville, AL
- Montgomery, AL

Arizona

- Phoenix, AZ
- Tuscon, AZ
- Scottsdale, AZ
- Flagstaff, AZ

Arkansas

- Little Rock, AR
- Fayetteville, AR
- Hot Springs, AR
- Fort Smith, AR

California

- Los Angeles, CA

- San Francisco, CA
- San Diego, CA
- Sacramento, CA
- Silicon Valley

Colorado

- Denver, CO
- Colorado Springs, CO
- For Collins, CO
- Summit County

Connecticut

- Hartford, CT
- New Haven, CT
- Stamford, CT
- Bridgeport, CT

Florida

- Miami/Ft. Lauderdale, FL
- Orlando, FL
- Jacksonville, FL
- Tampa/St. Petersburg, FL
- Naples/Ft. Meyers, FL

(OFFICE LOCATION)

Georgia

- Atlanta, GA

(OFFICE LOCATION)

- Savannah, GA
- Augusta, GA
- Athens, GA
- Macon, GA
- Social Circle, GA

Hawaii

- Honolulu, HI
- Kailua-Kona, HI
- Military Bases

Idaho

- Boise, ID
- Idaho Falls, ID
- Twin Falls, ID
- Meridian, ID

Illinois

- Chicago, IL
- Rockford, IL
- Champaign, IL

Iowa

- Des Moines, IA
- Cedar Rapids, IA
- Iowa City, IA
- Ames, IA

Kansas

- Wichita, KS
- Topeka, KS
- Kansas City, KS
- Lawrence, KS

Kentucky

- Lousville, KY
- Lexington, KY
- Bowling Green, KY
- Frankfort, KY

Louisiana

- New Orleans, LA
- Baton Rouge, LA
- Lafayette, LA
- Shreveport, LA
- Lake Charles, LA
- Alexandria, LA
- Metairie, LA
- Military Bases

Maine

- Portland, ME
- Bangor, ME
- Augusta, ME

Maryland

- Baltimore, MD
- Annapolis, MD
- Frederick, MD
- Rockville, MD

Massachusetts

- Boston, MA
- Cambridge, MA
- Salem, MA

Michigan

- Detroit, MI
- Grand Rapids, MI
- Lansing, MI
- Flint, MI
- Ann Arbor, MI
- Kalamazoo, MI

Minnesota

- Minneapolis, MN
- Saint Paul, MN
- Duluth, MN
- St. Cloud, MN
- Bloomington, MN

Mississippi

- Jackson, MS
- Hattiesburg, MS
- Biloxi, MS
- Gulfport, MS
- Military Facilities

Missouri

- St. Louis, MO
- Kansas City, MO
- Springfield, MO
- Columbia, MO
- Branson, MO

Montana

- Billings, MT
- Bozeman, MT
- Helena, MT
- Military Facilities

Nebraska

- Omaha, NE
- Lincoln, NE
- Grand Island, NE
- Kearney, NE

Nevada

- Las Vegas, NV
- Reno, NV
- Henderson, NV
- Carson City, NV
- Military Facilities

New Hampshire

- Manchester, NH
- Nashua, NH
- Portsmouth, NH
- Concord, NH

New Jersey

- Newark, NJ
- Jersey City, NJ
- Trenton, NJ
- Atlantic City, NJ

New Mexico

- Albuquerque, NM
- Santa Fe, NM
- Las Cruces, NM
- Roswell, NM
- Military Facilities

New York

- New York, NY
- Buffalo, NY
  (OFFICE LOCATION)
- Rochester, NY
- Albany, NY
- Syracuse, NY
- Military Facilities

North Carolina

- Charlotte, NC
- Raleigh, NC
  (OFFICE LOCATION)
- Greensboro, NC
- Winston-Salem, NC

North Dakota

- Fargo, ND
- Bismarck, ND
- Minot, ND
- Grand Forks, ND

Ohio

- Columbus, OH
- Cleveland, OH
- Cincinnati, OH
- Dayton, OH
- Akron, OH

Oklahoma

- Oklahoma City, OK
- Tusla, OK

- Norman, OK
- Edmond, OK
- Lawton, OK
- Military Facilities

Oregon

- Portland, OR
- Salem, OR
- Eugene, OR
- Bend, OR

Pennsylvania

- Philadelphia, PA
- Pittsburgh, PA
- Harrisburg, PA
- Erie, PA

Rhode Island

- Providence, RI
- Newport, RI
- Warwick, RI
- Cranston, RI

South Carolina

- Charleston, SC
- Columbia, SC
- Greenville, SC
- Spartanburg, SC
- Military Facilities

South Dakota

- Sioux Falls, SD
- Rapid City, SD
- Pierre, SD
- Aberdeen, SD

Tennessee

- Nashville, TN
- Memphis, TN
- Knoxville, TN

- Chattanooga, TN

Texas

- Houston, TX
- Dallas, TX
- Austin, TX
- San Antonio, TX
- Fort Worth, TX
- Military Facilities

Utah

- Salt Lake City, UT
- St. George, UT
- Ogden, UT
- Provo, UT
- Military Facilities

Vermont

- Burlington, VT
- Stowe, VT
- Montpelier, VT

Virginia

- Richmond, VA
- Virginia Beach, VA
- Norfolk, VA
- Alexandria, VA

Washington

- Seattle, WA
- Spokane, WA
- Vancouver, WA
- Tacoma, WA

West Virginia

- Charleston, WV
- Morgantown, WV
- Military Facilities

Wisconsin

- Madison, WI
- Milwaukee, WI
- Green Bay, WI

Wyoming

- Cheyenne, WY
- Casper, WY
- Jackson, WY
- Laramie, WY

Alaska

- Anchorage, AK
- Fairbanks, AK
- Juneua, AK

Canada

- Toronto
- Montreal
- Vancouver
- Calgary
- Edmonton
- Ottowa-Ontario
- Military Facilities

Mexico

- Mexico City
- Guadalajara
- San Juan

Carribbean

- San Juan
- Santo Domingo
- Havana
- Nassau
- Kingston
- Port-Au-Prince

Europe

- United Kingdom
- Spain
- Switzerland
- France
- Germany
- Italy

Power Management A spectrum of solutions provide constant power to mission-critical hardware.

Rack Power Distribution

Reliable rack power distribution units designed for critical applications with features tailored to your application and budget requirements.

Data Center Resources has the experience and expertise to assist you with load estimates, site assessments and product selection. Since we are not beholden to a specific manufacturer, you can be assured that our recommendations are the best available for your specific application. Because of this approach, our customers receive the best technology at very competitive pricing.

METERED RACK PDU'S MONITORED RACK PDU'S OUTLET MONITORED PDU'S SWITCHED RACK PDU'S AUTOMATIC TRANSFER SWITCHES

Row Distribution

Space-saving power distribution options for your data center power. Easily deliver power from your UPS power supply to wherever you need it.

1831

Supply UPS power to all cabinets in the row. Slim 9.5” enclosures provide A&B 225 or 400amp power feeds to server cabinet rows. Each panel has two separate 42-pole breaker positions. With the ever increasing demands for power in a small footprint, the HD400 provides an unprecedented 230kW of power in about 14 cubic feet. Configured into a 9.5” deep enclosure, position the HD400 at the end of an of a server cabinet aisle. EXPLORE OUR ROW DISTRIBUTION PRODUCTS

UPS Systems

UPS systems provide reliable back-up power for a few minutes to keep servers up during a power surge. Additionally, they provide back-up power for more than an hour during a major outage.

The correct UPS solution depends on several factors including level of criticalness of connected systems, purpose of the UPS, and other back-up infrastructure. Contact Data Center Resources. We help determine the right solution for you. EXPLORE OUR UPS SYSTEM PRODUCTS

1833

Modular Conceptpower 120 UPS

Designed for small to medium-sized data centers, Conceptpower 120 fulfills power requirements from 20 kW to 600 kW and provides maximum flexibility, redundancy and reliability to protect critical hardware.

> Explore Modular Conceptpower 120 UPS

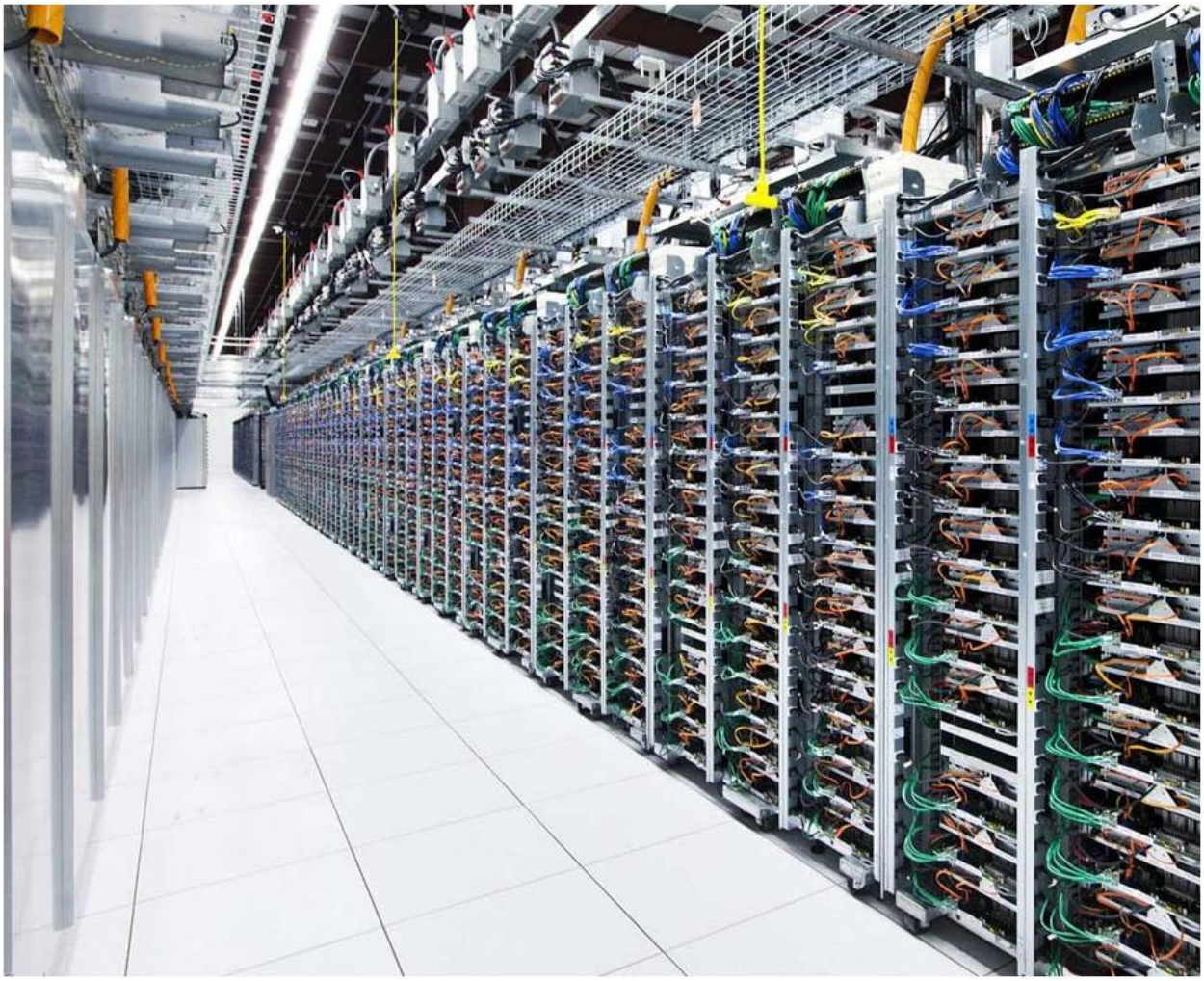
Modular Conceptpower 500 UPS

Designed for medium-sized and large data centers, Conceptpower 500 fulfills power requirements from 100 kW to 3 MW and provides maximum availability for those who aim for zero downtime as well as a low cost of ownership.

> Explore Modular Conceptpower 500 UPS

Cords and Cabling

Cords, cables and other accessories to help efficiently distribute power throughout the data center.



Efficient cords and cables provide the obvious benefit of reducing clutter, shorter power cords facilitate a number of datacenter best-practices. Furthermore, shorter power cords improve the airflow of the rear of the rackmount enclosure, likewise they also improve the efficiency of the servers. EXPLORE OUR CORDS AND CABLES

Server Power Cords

Simplify cable management and improve air flow with shorter server cables.

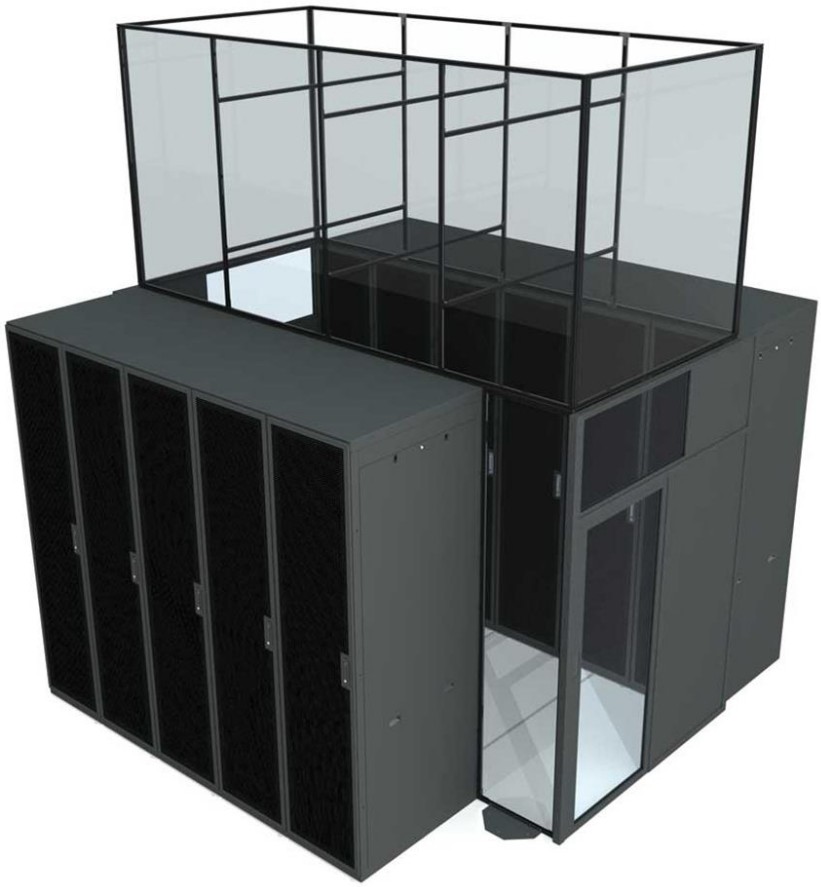
Power Distribution Cables

Bring power from the distribution panel to rack power distribution units or UPS.

Cooling Optimization Maximize capacity, efficiency and performance in existing cooling infrastructures.

Maximize Cooling Infrastructure

Data Center Resources helps customers to get the most capacity, efficiency and performance out of their existing data center cooling infrastructures. Data Center Resources has several proprietary solutions to maximize cooling efficiency. In addition, we have relationships with several leading manufacturers to offer the best products on the market.TOOLS TO MAXIMIZE COOLING INFRASTRUCTURE



Aisle Containment

The Cool Shield aisle containment system maximizes cooling efficiency in new and existing data centers.

EC Fan Replacement

Replace belt driven CRAC/CRAH fans with energy efficient variable speed EC fans. Compatible with all leading manufacturers.

1842

Crac Hoods

Increase efficiency and capacity of cooling units by channeling warm exhaust air directly to the top return.

1843

Floor Grommets

Sealing the openings in the raised floor prevents cool supply air from escaping where it is not needed.

Maximize Airflow

Managing the airflow around server cabinets is critical to prevent hardware from overheating. Solutions to manage airflow include blanking panels, expanding foam tape and high-flow floor tiles.PRODUCTS THAT HELP TO MAXIMIZE AIRFLOW

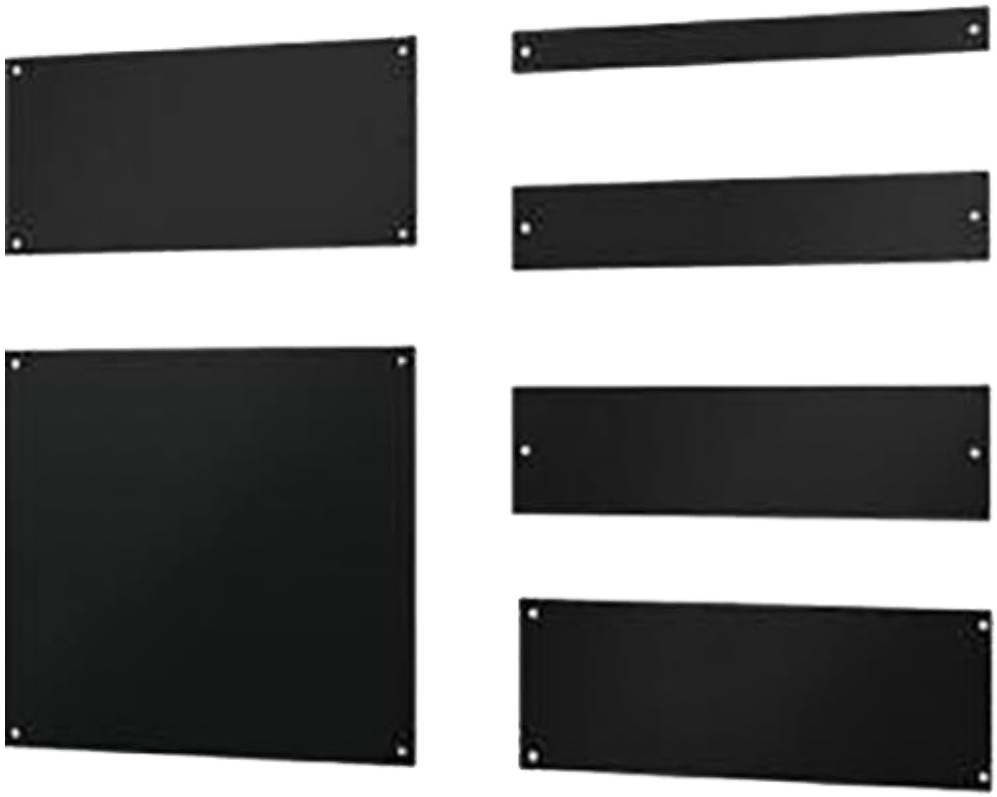
Ceiling Grid Fan Tray

Designed to aggressively vent air from equipment rooms or closets.

1846

High Flow Floor Tiles

55% to 66% perforated high flow floor tiles help cool high density racks & hot spots.



Blanking Panels

Fill 1U to 10U Empty Spaces in Server Cabinet to Minimize Bypass Airflow and Increase Cooling Efficiency.

Cool Shield Air Strip – Foam Gap Seal

Expanding Foam Gap Fill Seal Strips Prevent Mixing Supply And Exhaust Air.

Supplemental Cooling Optimization

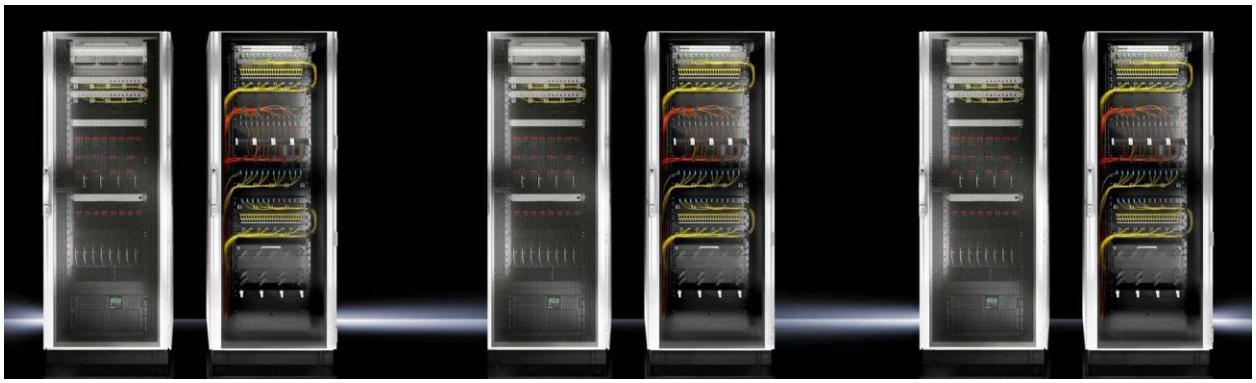
As server loads grow, cooling capacity must increase beyond existing room capabilities. Data Center Resources offers several options for providing additional cooling optimization without the expense and disruption of a new CRAC unit.SUPPLEMENTAL COOLING PRODUCTS

Air Conditioned Enclosures (ACE)

Controls internal temperature in locations without room/constant air-conditioning.

DCR In-Row Cooling

Point of use cooling to remove 20kW of heat load within a row of data center racks.



Server Rack Infrastructure Fully integrated solutions provide an optimized environment for critical hardware.

Server Cabinets

Our data center server rack cabinets combine strength, airflow management and customizability. Eighty percent door perforation and restrictions on front to back airflow make our cabinets ideal for contained aisles. Numerous power mounting and cable management options make cabinets extremely configurable. EXPLORE SERVER CABINET PRODUCTS

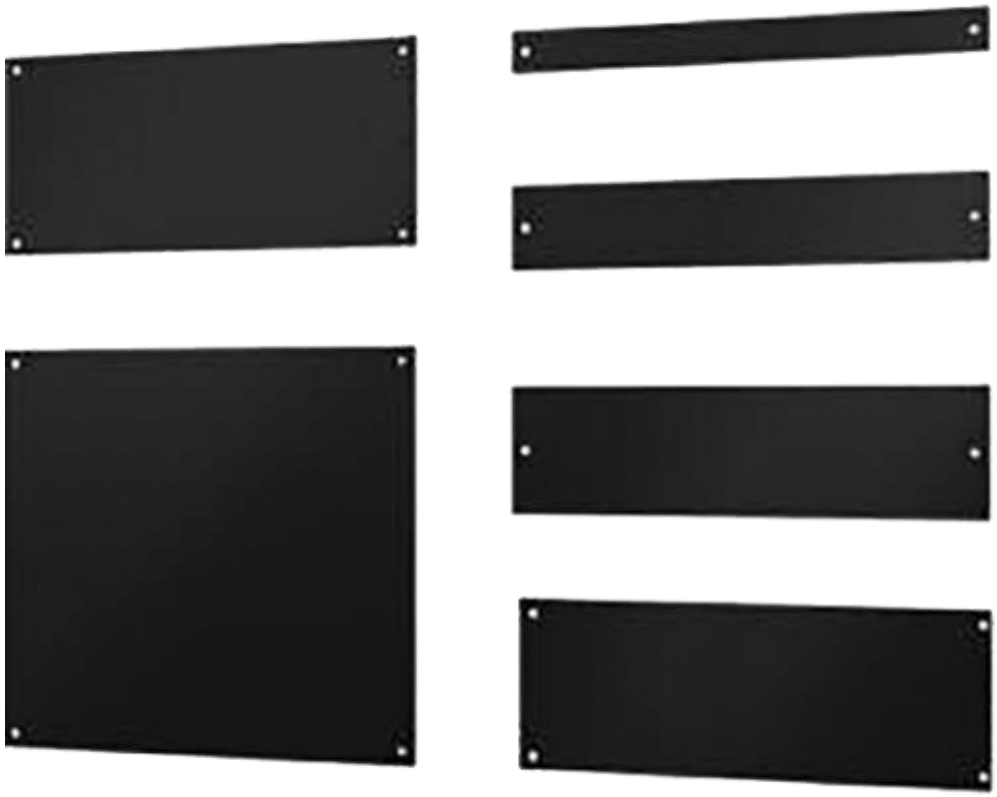
ESG Series 2000

A highly configurable colocation server cabinet available in multiple heights.

DirectCool Cabinet System

Closed loop 25 kW water-cooled module cools one or two cabinets

1855



Cool Shield Blanking Panels

Fill 1U to 47U Empty Spaces in Server Cabinet to Minimize Bypass Airflow and Increase Cooling Efficiency.

1856

Containment Pods

Out integrated Containment Pods infrastructure solution features components for servers to thrive. Server cabinets, rack power, remote monitoring, airflow containment and cooling (optional) combine for a highly efficient and effective approach to housing critical hardware. EXPLORE CONTAINMENT PODS

Modular Data Center

Certain buildings have room or appropriate conditions to house data center equipment. Our standalone Modular Data Center solution buildings are built for indoor or outdoor deployment. The base building includes a raised floor, fire suppression, and lighting and light electrical. Best-of-breed infrastructure including cabinets, power and bundled cooling. EXPLORE MODULAR DATA CENTER

Data Center Cages

Our Edge Cage is designed for the data center environment. Eight-sided posts provide unmatched configurability. Replaceable panels feature a combination wire mesh, polycarbonate or solid metal inserts to manage airflow or privacy concerns. Easily create custom size rooms with modular component partitions. Hinged and sliding doors available. EXPLORE DATA CENTER CAGES

Data Center Monitoring Gather and organize data to prevent downtime and maximize efficiency.

Battery Data Center Monitoring

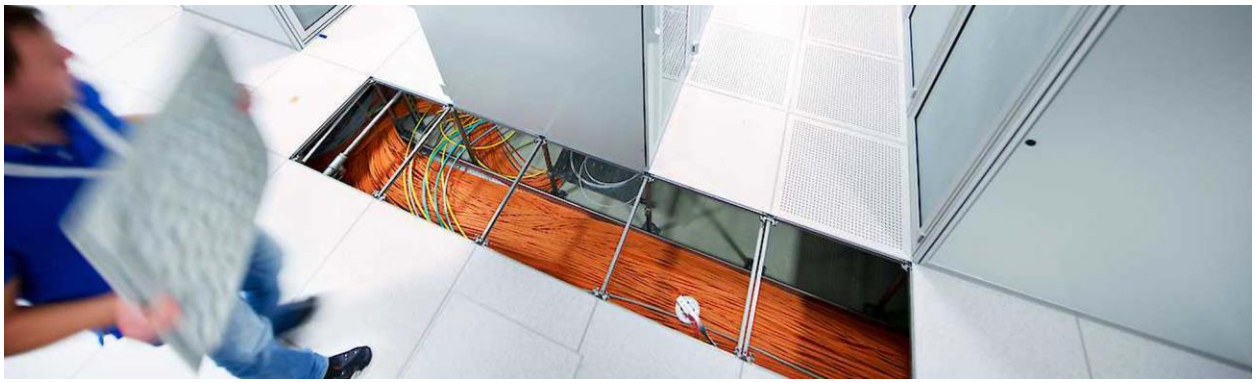
Our patented Battery Data Center Monitoring System continuously captures and securely transmits battery performance data. Suited for both VRLA and wet cell batteries, the system is cost effective and can be easily installed on UPS, generator, DC plant and switchgear battery systems. It is the most comprehensive, cost effective and easy to install system on the market today. EXPLORE BATTERY DATA CENTER MONITORING PRODUCTS

Branch Circuit Data Center Monitoring

This software solution is a simple but effective tool for aggregating and visualizing critical environmental and power information to safely achieve efficiency gains. They allow you to visualize power and cooling conditions in real-time by linking rack power and temperature sensors to your cabinet floor plans. Alert thresholds monitoring the effects of incremental changes to maximize efficiency programs. EXPLORE BRANCH CIRCUIT MONITORING PRODUCTS

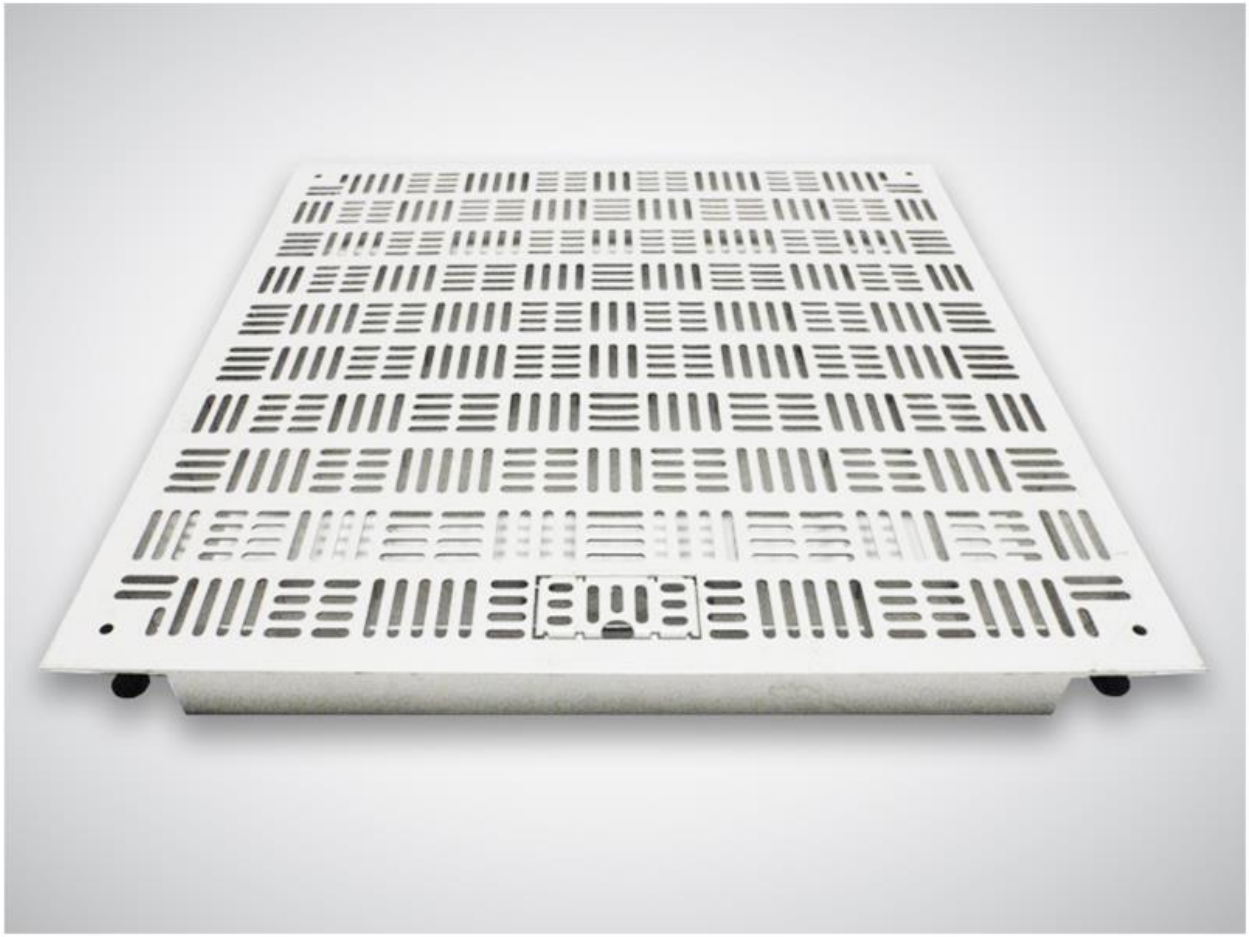
DCIM Software

Data Center Infrastructure Monitoring (DCIM) solutions monitor and manage critical assets in one integrated package. Our enterprise solution allows multi-site management for power, cooling, racks and servers, enabling you to instantly see power and cabling chain for each server. Asset tracking provides location, service, warranty and performance information on each data center component. EXPLORE DCIM SOFTWARE PRODUCTS



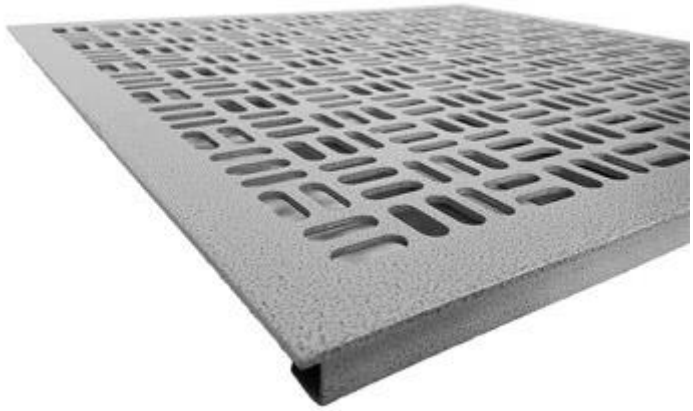
Data Center Flooring High performance access floor tiles for new and existing raised floor systems.

High-Flow Access Floor Tiles



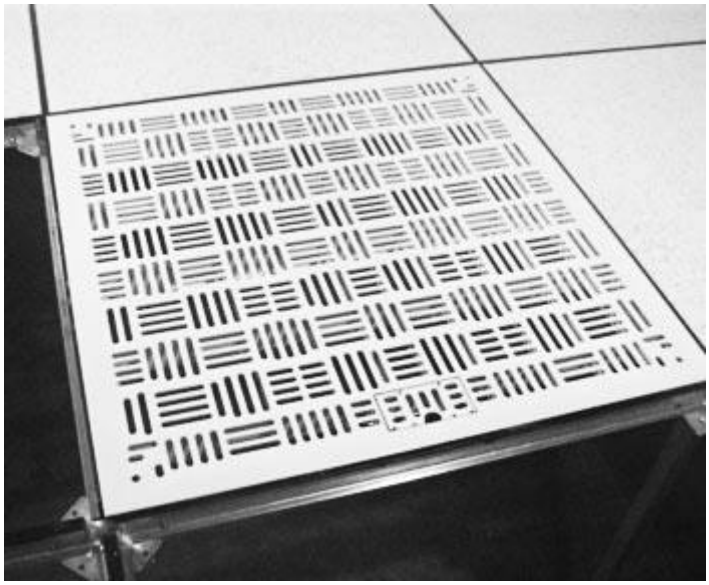
1863

High-Flow raised access floor tiles range from 32% to 66% perforation models, delivering more than twice as much cool air in your cold aisle. The result is cold air delivered exactly where your servers need it most. EXPLORE HIGH-FLOW RAISED ACCESS FLOOR TILES



32% High Flow Floor Tile

Air-Flow Panel provides both strength and performance in an attractive package.

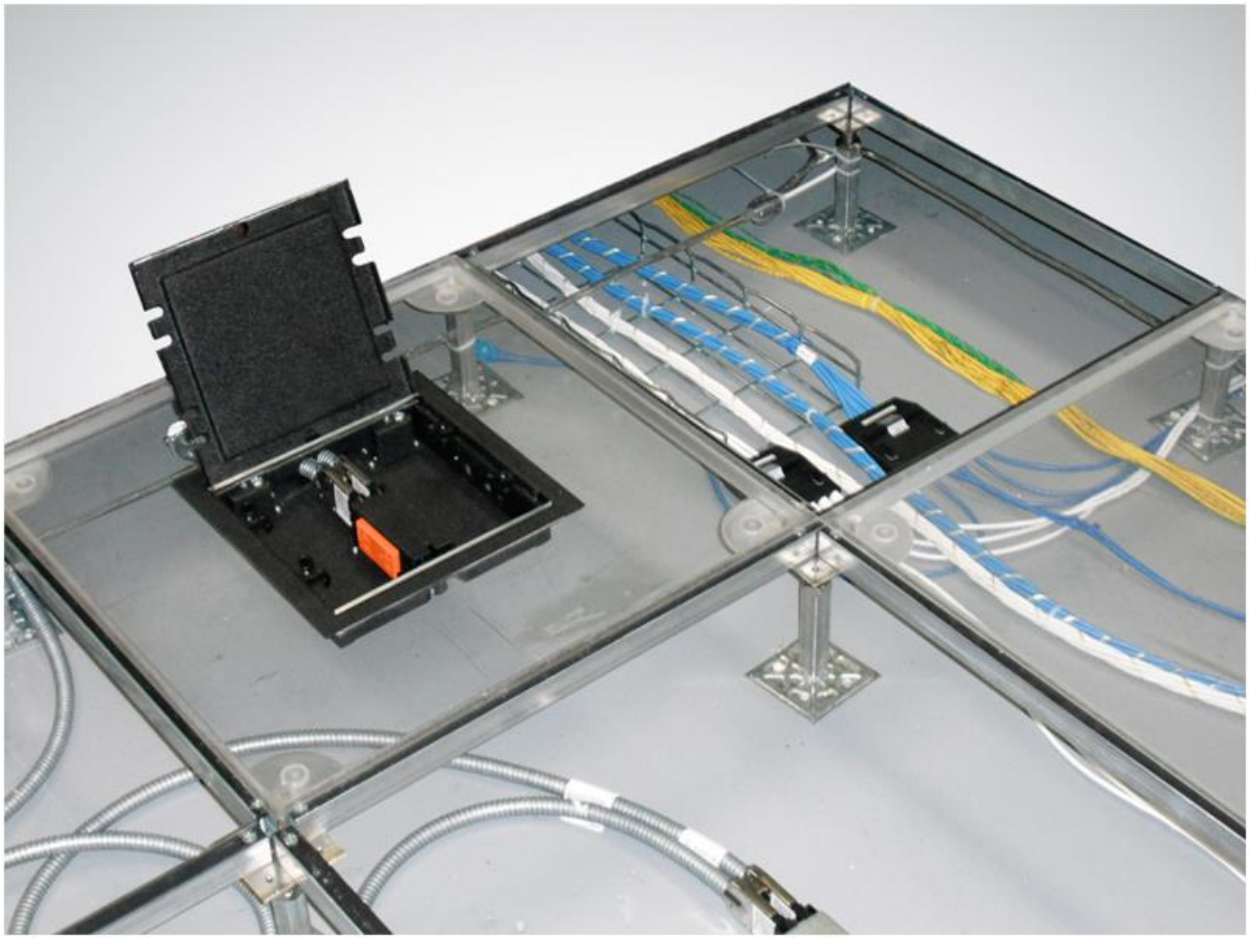


Infinity 66% High Flow Floor Tile

1864

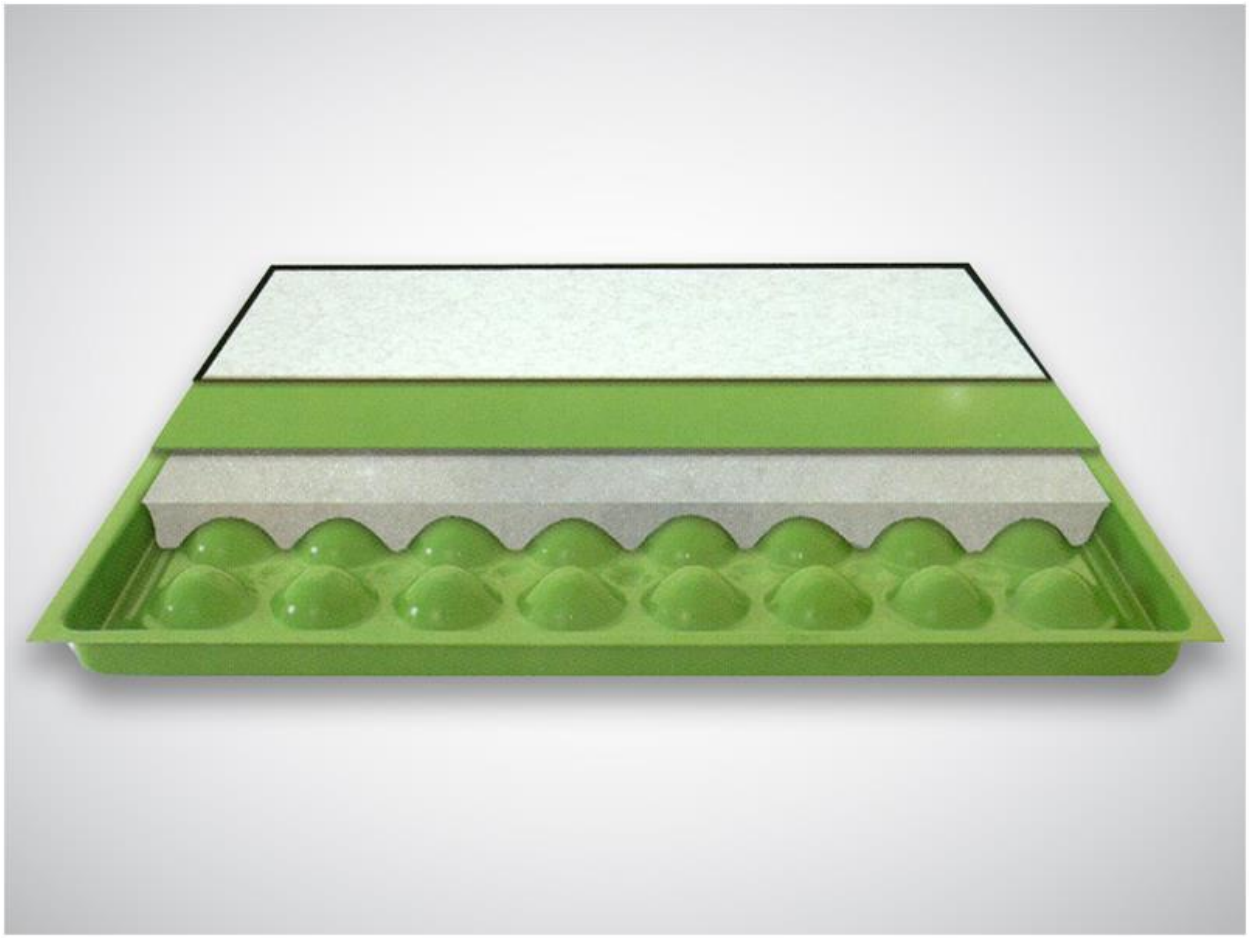
66% perforated high flow floor tile to help cool high density racks & hot spots.

Clear Access Floor Tiles



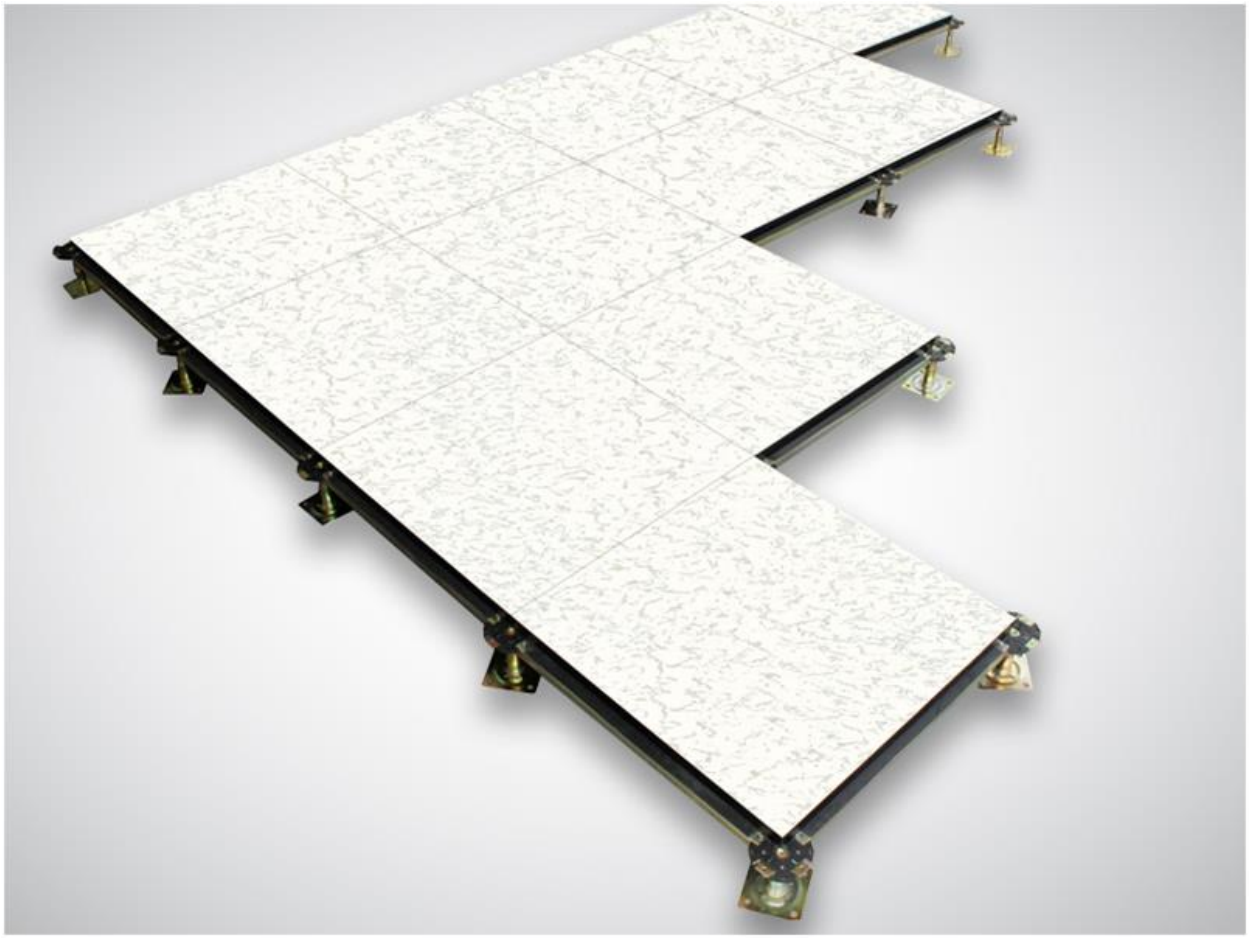
Designed for easy viewing of valves, meters and underfloor cavity, clear floor tiles provide a quick glimpse of the general condition of the plenum space. EXPLORE CLEAR FLOOR TILES

New Access Floor Tiles



High strength raised access floor systems designed for the critical data center. Access floor systems available to support loads ranging from 1,250 to 10,000 lbs per sq. foot. We offer woodcore raised access floor tiles, concrete floor tiles, steel floor tiles, structural stringer footer supports and data center installation services. EXPLORE NEW ACCESS FLOOR TILES

Replacement Access Floor Tiles



Raised access floor tile systems have existed since the 1960's. Raised access floor technology is constantly evolving. Our experts can match and source replacements for your existing raised floor tile sytem. We carry new old stock, used floor tiles and a variety of universal tiles. Contact us to replace or upgrade your current worn or damaged current tiles. EXPLORE REPLACEMENT ACCESS FLOOR TILES

Access Floor Tile Accessories

Solutions to enhance and support your raised access floor infrastructure.

1867

Floor tile pulls allow quick movement of tiles for servicing in the plenum area. Grommets and covers for floor tile cutouts to provide access underneath floor tiles while preventing unwanted air leakage. Frostbyte cooler panels mount to your existing high flow tiles and use an internal fan to deliver spot cooling. EXPLORE ACCESS FLOOR TILE ACCESSORIES

KoldLok Wave

Wave shaped thermoplastic elastomer floor grommet is the best sealing available

1869

Frost-Byte Cooler Panel

High Efficiency Rack Cooler. Delivers cooling air where and when it is needed.

Control Room Furniture Consoles 24x7 control room furniture for NOC’s, EOC's, process control and monitoring.

Command Watch

Command Watch is a modular and versatile furniture-grade NOC control room furniture console solution. Most noteworthy, the Command Watch control room furniture console system is designed for the rigorous demands of the critical operations, EOC’s and monitoring environments. Furthermore, Command Watch units features a versatile 3-sided slatwall and under work surface cabinet for CPU storage. Additionally, integrated power, cooling lighting and monitor arms are

seamlessly designed into the system. EXPLORE NOC CONTROL ROOM FURNITURE & CONSOLE SOLUTIONS

Command Flex

An economical yet robustly designed technical NOC control room console furniture solution that features integrated power, CPU holder and cable management. Moreover, Command Flex units feature an open-air configuration and our highly functional proprietary slatwall. EXPLORE COMMAND FLEX NOC CONTROL ROOM CONSOLE FURNITURE

Command Flex

Economical and robustly designed open air slatwall furniture solution with power and cable management.

1873

Command Flex Adjustable Height

A sit/stand console solution intended for the emergency response & control room environments.

1874

Custom Millwork

There is often a requirement for other specialized furniture items to further enhance productivity. In fact, we manufacture custom enhancements attached to the control room console furniture itself including cases for hardware, locking storage units or shelving for tops of slatwalls. EXPLORE CUSTOM MILLWORK FOR CONTROL ROOMS

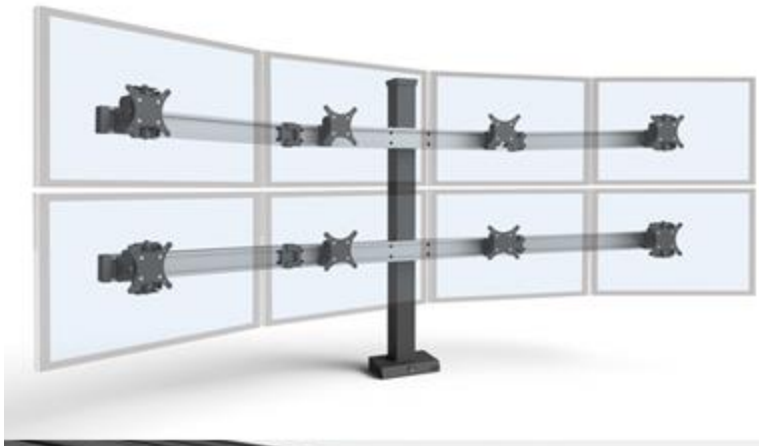
7×24 Operator Chairs

Our ergonomic console seating and chairs are designed specifically for the 7×24 environment. Industrial strength fabrics and materials combine to make our chairs incredibly resistant to spills while remaining comfortable. Perfect for the Police, Fire, EMS, government and private applications.

- Variety of seating materials
- Heavy Duty
- Spill Resistant
- Comfortable Foam
- Mold and Mildew Resistant
- Highly Adjustable
- Sit Stand Compatible

7×24 OPERATOR CHAIRS

Control Room Furniture Accessories



Our control room furniture accessories enhance and complete your network operations center console furniture solution. As shown below, accessories improve productivity and operator comfort.

- Monitor Arms
- Slatwall Accessories
- Task Lighting
- 24×7 Operator Seating
- Large Display Mounts
- Pop-Up Power /Data/USB

EXPLORE NOC FURNITURE CONSOLE ACCESSORIES

Control Room Design

Our expert engineers design your control room project for maximum comfort and productivity for your critical operators. Efficient control room design is our business, DCR designs and builds your control room based on your budgetary needs and specifications.

- Control Room Findings & Audit Studies
- Design for Specialty Applications (SCADA & DCS)
- Efficiency Consulting
- Training & Resources
- Control Room Design Objectives
- Fatigue Mitigation Rooms
- Permitting/Lockout Target Rooms

EXPLORE CONTROL ROOM DESIGN

Control Room Furniture

Modular furniture-grade NOC consoles designed for critical operations, EOC's, Process Control Plants and 24/7 monitoring environments.

Our Enterprise Control Room Furniture Console units feature a versatile 3-sided slatwall and integrated under work surface cabinet for CPU storage.

- 12 gauge steel welded frame
- 1 1/4″ thick HPL work surface with polyurethane edge nosing
- 24″, 36″, 48″, 60″ & 72″ straight modules

- 15, 22.5, 45, & 90 degree standard & reverse corner units
- Standard 30″ & 36″ depth
- 6″h x 3″d Slatwall Cube allows mounting monitors to top, back or top
- Multi-monitor and large display (up to 70″) slatwall mounts
- Console furniture desk accessories include binder shelf, phone shelf, task light and pop-up power/data ports

EXPLORE NOC CONSOLE FURNITURE SOLUTIONS

Central Intelligence Amazon

The CIA's HQ in Langley– CIA / Wikimedia Commons

Bloomberg reports that the CIA outlined its Commercial Cloud Enterprise (C2E) initiative to potential vendors in late March. Unlike the controversial single-award $10bn JEDI cloud contract pursued by the US military, the CIA aims for C2E to use multiple cloud companies to handle both unclassified and secret information.

In documents seen by the publication, the agency indicated that it intends to award one or more contracts lasting at least five years, with options for extension. After draft contract requirements are revealed, a final request for proposals is set for May 2020, with awards granted the following year.

The CIA is still soliciting feedback on the scope and scale of the project, and has not completely settled on the plan to use multiple companies.

AWS would still be a strong contender for the contracts, but others have rapidly expanded their services, and gained necessary certifications. Microsoft, for example, last year was allowed to offer its Azure

Government services to seventeen US agencies, including the CIA. IBM, meanwhile, built a private cloud data center for the US Army as part of the Army Private Cloud Enterprise (APCE) program, and provides cloud services to the Army's logistics division.

The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)

Digital Manipulation

1883

The spring air in the small, sand-dusted town has a soft haze to it, and clumps of green-gray sagebrush rustle in the breeze. Bluffdale sits in a bowl-shaped valley in the shadow of Utah's Wasatch Range to the east and the Oquirrh Mountains to the west. It's the heart of Mormon country, where religious pioneers first arrived more than 160 years ago. They came to escape the rest of the world, to understand the mysterious words sent down from their god as revealed on buried golden plates, and to practice what has become known as "the principle," marriage to multiple wives.

Today Bluffdale is home to one of the nation's largest sects of polygamists, the Apostolic United Brethren, with upwards of 9,000 members. The brethren's complex includes a chapel, a school, a sports field, and an archive. Membership has doubled since 1978—and the number of plural marriages has tripled—so the sect has recently been looking for ways to purchase more land and expand throughout the town. But new pioneers have quietly begun moving into the area, secretive outsiders who say little and keep to themselves. Like the pious polygamists, they are focused on deciphering cryptic messages that only they have the power to understand. Just off Beef Hollow Road, less than a mile from brethren headquarters, thousands of hard-hatted construction workers in sweat-soaked T-shirts are laying the groundwork for the newcomers' own temple and archive, a massive complex so large that it necessitated expanding the town's boundaries. Once built, it will be more than five times the size of the US Capitol.

Rather than Bibles, prophets, and worshippers, this temple will be filled with servers, computer intelligence experts, and

armed guards. And instead of listening for words flowing down from heaven, these newcomers will be secretly capturing, storing, and analyzing vast quantities of words and images hurtling through the world's telecommunications networks. In the little town of Bluffdale, Big Love and Big Brother have become uneasy neighbors.

>The NSA has become the largest, most covert, and potentially most intrusive intelligence agency ever.

Under construction by contractors with top-secret clearances, the blandly named Utah Data Center is being built for the National Security Agency. A project of immense secrecy, it is the final piece in a complex puzzle assembled over the past decade. Its purpose: to intercept, decipher, analyze, and store vast swaths of the world's communications as they zap down from satellites and zip through the underground and undersea cables of international, foreign, and domestic networks. The heavily fortified $2 billion center should be up and running in September 2013. Flowing through its servers and routers and stored in near-bottomless databases will be all forms of communication, including the complete contents of private emails, cell phone calls, and Google searches, as well as all sorts of personal data trails—parking receipts, travel itineraries, bookstore purchases, and other digital "pocket litter." It is, in some measure, the realization of the "total information awareness" program created during the first term of the Bush administration—an effort that was killed by Congress in 2003 after it caused an outcry over its potential for invading Americans' privacy.

But "this is more than just a data center," says one senior intelligence official who until recently was involved with the program. The mammoth Bluffdale center will have another important and far more secret role that until now has gone unrevealed. It is also critical, he says, for breaking codes. And code-breaking is crucial, because much of the data that the center will handle—financial information, stock transactions, business deals, foreign military and diplomatic secrets, legal documents, confidential personal communications—will be heavily encrypted. According to another top official also involved with the program, the NSA made an enormous breakthrough several years ago in its ability to cryptanalyze, or break, unfathomably complex encryption systems employed by not only governments around the world but also many average computer users in the US. The upshot, according to this official: "Everybody's a target; everybody with communication is a target."

For the NSA, overflowing with tens of billions of dollars in post-9/11 budget awards, the cryptanalysis breakthrough came at a time of explosive growth, in size as well as in power. Established as an arm of the Department of Defense following Pearl Harbor, with the primary purpose of preventing another surprise assault, the NSA suffered a series of humiliations in the post-Cold War years. Caught offguard by an escalating series of terrorist attacks—the first World Trade Center bombing, the blowing up of US embassies in East Africa, the attack on the USS *Cole* in Yemen, and finally the devastation of 9/11—some began questioning the agency's very reason for being. In response, the NSA has quietly been reborn. And while there is little indication that its actual effectiveness has improved—after

all, despite numerous pieces of evidence and intelligence-gathering opportunities, it missed the near-disastrous attempted attacks by the underwear bomber on a flight to Detroit in 2009 and by the car bomber in Times Square in 2010—there is no doubt that it has transformed itself into the largest, most covert, and potentially most intrusive intelligence agency ever created.

In the process—and for the first time since Watergate and the other scandals of the Nixon administration—the NSA has turned its surveillance apparatus on the US and its citizens. It has established listening posts throughout the nation to collect and sift through billions of email messages and phone calls, whether they originate within the country or overseas. It has created a supercomputer of almost unimaginable speed to look for patterns and unscramble codes. Finally, the agency has begun building a place to store all the trillions of words and thoughts and whispers captured in its electronic net. And, of course, it's all being done in secret. To those on the inside, the old adage that NSA stands for Never Say Anything applies more than ever.

A former CIA agent wants tighter controls around how tech giants use personal data

Yael Eisenstat said during her recent visit to Sydney that it's time for governments to step up the level of regulation.

She might consider herself as someone who doesn't easily open up to others, but Yael Eisenstat, a former Central Intelligence Agency (CIA) officer and special advisor on national security to former Vice President Joe Biden, has argued that governments have a responsibility to lift regulations surrounding how tech giants use data to fuel the growth of their companies.

"I don't want them regulated as a publisher. What they're doing is more dangerous. They're curating our content. They're deciding what we do or do not see. They're deciding which rabbit holes to take us down and what filter bubbles to put us in, and if they're curating our content then they need to be classified as digital curators and regulated that way," she said.

"You can't tell me they bear no responsibility for the content they put out there if they are curating your content."

During her visit to Sydney as part of D61+ Live last month, Eisenstat, who also worked briefly at Facebook to head up the company's social media integrity operations after being caught in election meddling issues, spoke bluntly about her concerns around the ferocious approach tech giants have taken to grab the attention of users.
"Advertising in itself is not necessarily the problem, it's the fact these [technology] companies have taken all the advertising away from what used to be a regulated market -- think TV and newspapers -- and really brought it into their platforms," she said.

"But they may not be selling our data to advertising companies, but they are selling to advertisers this idea that they have so much data on us they know exactly how to target us with exactly what we want to see.

"Why is that problem? Because in order to do so, their number one metric they need to care about is user engagement. They have to be able to keep you on their platform … the business model tries to ensure you're addicted to their technology so badly they can target you."

She continued describing how the engines of these companies are "bringing us down more and more extreme paths, and it's wrecking our ability to see that we're even being manipulated".

While engineers may be the designers behind these "addictive" platforms, Eisenstat doesn't necessarily blame them.

"I don't think it's the engineers who sat there and thought let's show somebody who is trying to look up the NASA moon landing … a conspiracy video …The algorithm has figured out if I show you slightly more extreme content than the last thing you saw, you will click and watch."

Eisenstat's concerns are not unique. She joins a growing list of other individuals and organisations who are equally disturbed by the dominance of these technology giants.

Just last week, Facebook and Google were slammed by Amnesty International for having "surveillance-based" business models that "poses a systemic threat to human rights".

In its 60-page Surveillance Giant report [PDF], the human rights organisation called for governments to overhaul regulation to ensure the likes of Google and Facebook are prevented from making access to their service conditional on individuals "consenting" to the collection, processing, or sharing of their personal data for marketing or advertising.
"The internet is vital for people to enjoy many of their rights, yet billions of people have no meaningful choice but to access this public space on terms dictated by Facebook and Google," said Kumi Naidoo, Secretary General of Amnesty International, who described the control that these companies have as insidious.
"To make it worse, this isn't the internet people signed up for when these platforms started out. Google and Facebook chipped away at our privacy over time. We are now trapped.

"Either we must submit to this pervasive surveillance machinery -- where our data is easily weaponised to manipulate and influence us -- or forego the benefits of the digital world. This can never be a legitimate choice. We must reclaim this essential public square, so we can participate without having our rights abused."

The scathing analysis by the organisation and outcry from Eisenstat follows on from when Australia, the United States, and the United Kingdom jointly requesting for Facebook to stop end-to-end encryption.
In a letter to Facebook last month, the three countries said that the company "should not deliberately design their systems to preclude any form of access to content" as the governments believe encryption will put citizens and societies at risk of child sexual exploitation and abuse, terrorism, and foreign interference.

They also called for Facebook to allow law enforcement to obtain lawful access to content in a readable and usable format; engage in consultation with governments to facilitate lawful access and allow this to influence Facebook's design decisions; and to not implement any encryption changes.

Read more: Facebook data privacy scandal: A cheat sheet (TechRepublic)
In July, Facebook agreed to pay a record $5 billion fine and agreed to conduct an overhaul of its consumer privacy practices as part of a settlement with the US Federal Trade Commission (FTC).
The FTC launched an investigation, following the events of the Cambridge Analytica scandal, alleging that Facebook

repeatedly used "deceptive disclosures and settings to undermine users' privacy preferences" in violation of its 2012 agreement with the FTC. The FTC also alleged that Facebook was inadequate in dealing with apps that it knew were violating its platform policies.

Going forward, Facebook said it would be required to conduct a privacy review of every new product, service, or practice it develops before it's implemented, as well as establish an independent privacy committee to strip Zuckerberg of his "unfettered control" over user privacy decisions.

As a result of these reviews, Facebook announced in September that it suspended tens of thousands of apps from about 400 developers.
Facebook also pledged that it would impose stricter rules on political advertisers ahead of the 2020 election as part of its ongoing efforts to bolster ad transparency.
As of June 2019, the social media giant touted it had more than 2.1 billion people using Facebook, Instagram, WhatsApp, or Messenger every day every day on average, and 1.59 billion daily active Facebook users.

In Australia alone, the social network raked in AU$125.5 million dollars for the year ended 31 December 2018, and of that AU$125 million was from its online advertising sales.

Security Think Tank: CIA at heart of infosec-data architect partnership

How infosec professionals and data can architects work together to support business goals and achieve a good level of cyber security?

The role of the information security professional is to secure information and, by extension, systems. In times before computers, information was often protected using a variety of methods, such as a stamped seal on paper, to maintain the confidentiality, integrity and availability (CIA) of the information for the intended recipient.

Today, although the demands made on the information security function are far more complex, the principles remain the same.

The amount of data and information in an organisation increases significantly hour by hour, day by day, and year by year. This is where the data architect comes in.

A data architect is responsible for conceptualising and designing the framework that the enterprise uses for managing these huge volumes of data, to deliver available and reliable information to the enterprise.

This role has the expertise to support the handling of a range of data sources from wide and varied locations, effectively bringing order to what can often be a chaotic data environment.

The information lifecycle has five stages – create, process, store, transmit, destroy – and the creation and processing of data is where much of the increase in volume comes from. Compliance demands on organisations across the globe mean that knowing what data your organisation is responsible for, where it is and how it is maintained, is now a priority for enterprises.

A natural extension of the role of the data architect in designing the enterprise data framework is to support information security professionals in protecting the CIA of the data.

Of course, not all data is equal. Data needs to be classified so it can be protected appropriately. Working alongside the information security function, the data architect will enable understanding of the data that needs to be protected and where it resides.

The information security function will develop security policies for data across the organisation, and some of that data will require extra protection, such

as encryption (Ovum's software market forecast for security predicts a compound annual growth rate of 8.6% for encryption over the next five years).

The data architect will work with the information security function to ensure the appropriate data is encrypted, while still maintaining the availability of information.

The information security function needs the data architect, and vice versa, to protect the CIA of data and information across the organisation.

CiA® test center

The CiA test center is an official CiA testing facility providing CANopen compliancy assessment services such as CANopen device conformance testing and CiA interoperability testing.
CANopen device conformance testing is used to determine whether CANopen devices comply with the requirements of CiA 301 "CANopen application layer and communication profile" specification. CiA engineers carry out the device conformance testing using the automated test bench CANopen conformance test tool, developed by the CiA consortium. The test tool is implemented according to CiA 310-1 "CANopen application layer conformance test plan" and is free of charge for our members.
CiA members may use this test tool to test their devices even in early stages of implementation for compliance to CANopen, i.e. to verify if their implementation is correct. Moreover, CiA members often pre-test their devices using the test tool before sending the devices to the CiA testing center for conformance testing.

CANopen interoperability testing is a CANopen network compatibility testing to determine whether CANopen devices can be integrated into a CANopen network and perform their function without causing any communication problems.

CiA engineers test the interoperability using a pre-configured CANopen network with CANopen compliant and interoperable devices controlled by PLC Modicon (Schneider Electric). All these devices, manufactured by CiA member companies, are mounted into a portable interoperability

testing bench, that is also used as a interoperability demonstrator at the exhibitions as Embedded World and SPS in Nuremberg.

The test procedure is implemented according to the "CANopen application layer interoperability test plan".
Additionally, CiA organizes joint interoperability testing called CiA plugfests for its members. The plugfest makes sense in case many companies would test interoperability of their devices exceeding the testing capacity of the interoperability testing bench at CiA or CiA interoperability testing bench does not offer the functionality to test devices specially designed for certain applications like CANopen lift (CiA 417) or SIIS-level 2 devices (CiA 443). In case of such applications CiA member companies provide the control devices and end devices to tests their interoperability. Participants connect their devices to a joint network and test the interoperability of their devices with assistance of CiA engineers. Since 2014, the CiA test center also offers CAN FD plugfests to promote this new CAN standard.
These CiA services are not mandatory but give you an independent CANopen product compliancy evaluation especially if your customers require it. CANopen device certification and CANopen interoperability testing can be done independently. However, CiA recommends to perform the CANopen device testing using CANopen conformance test tool or CANopen conformance testing at CiA test center for application specific devices before the CANopen interoperability testing. CANopen interoperability testing is an optional service but it is recommended because the CANopen conformance testing alone doesn’t guarantee device interoperability.

Is the CIA's new cloud procurement a signal to DoD to update JEDI?

The CIA created quite a stir in the federal IT community as word spread over the last week that it's ready to upgrade its commercial cloud offering called Commercial Cloud Services (C2S).

As the industry day documents spread like wildfire across industry and the media, the question we have to ask is the CIA, the intelligence community more generally, trying to give the Defense Department some top cover for its controversial and protest entangled Joint Enterprise Defense Initiative (JEDI) cloud procurement?

When you review the CIA's market research survey as well as its industry day presentation, everything about it seems to be saying "Hey DoD, we have seen the light and multi-cloud, multi-vendor is the only way to go."

The intel agency said in its market research survey that it "will acquire foundational cloud services, as defined in the scope section below, from multiple vendors."

In industry day documents, the CIA said that the Commercial Cloud Enterprise's (C2E) program objective is to "acquire cloud computing services directly from commercial cloud service providers…"

The CIA said it plans to award one or more indefinite delivery, indefinite quantity type contracts.

Industry experts said the message couldn't be any clearer to DoD and it's plans for JEDI.

Trey Hodgkins, president and CEO of Hodgkins Consulting, said the CIA's C2E puts the conversation around DoD's JEDI on a different trajectory.

"C2E puts the conversation on a different trajectory. It puts out there that the IC has identified new needs so the prudent person would go back and ask the question, 'if they need hybrid, on premise and commercial cloud, does that change the thinking at DoD?'" said Trey Hodgkins, president and CEO of Hodgkins Consulting. "I don't think there is any visibility into DoD's thought process, but you'd have to think they are asking the same question at the department."

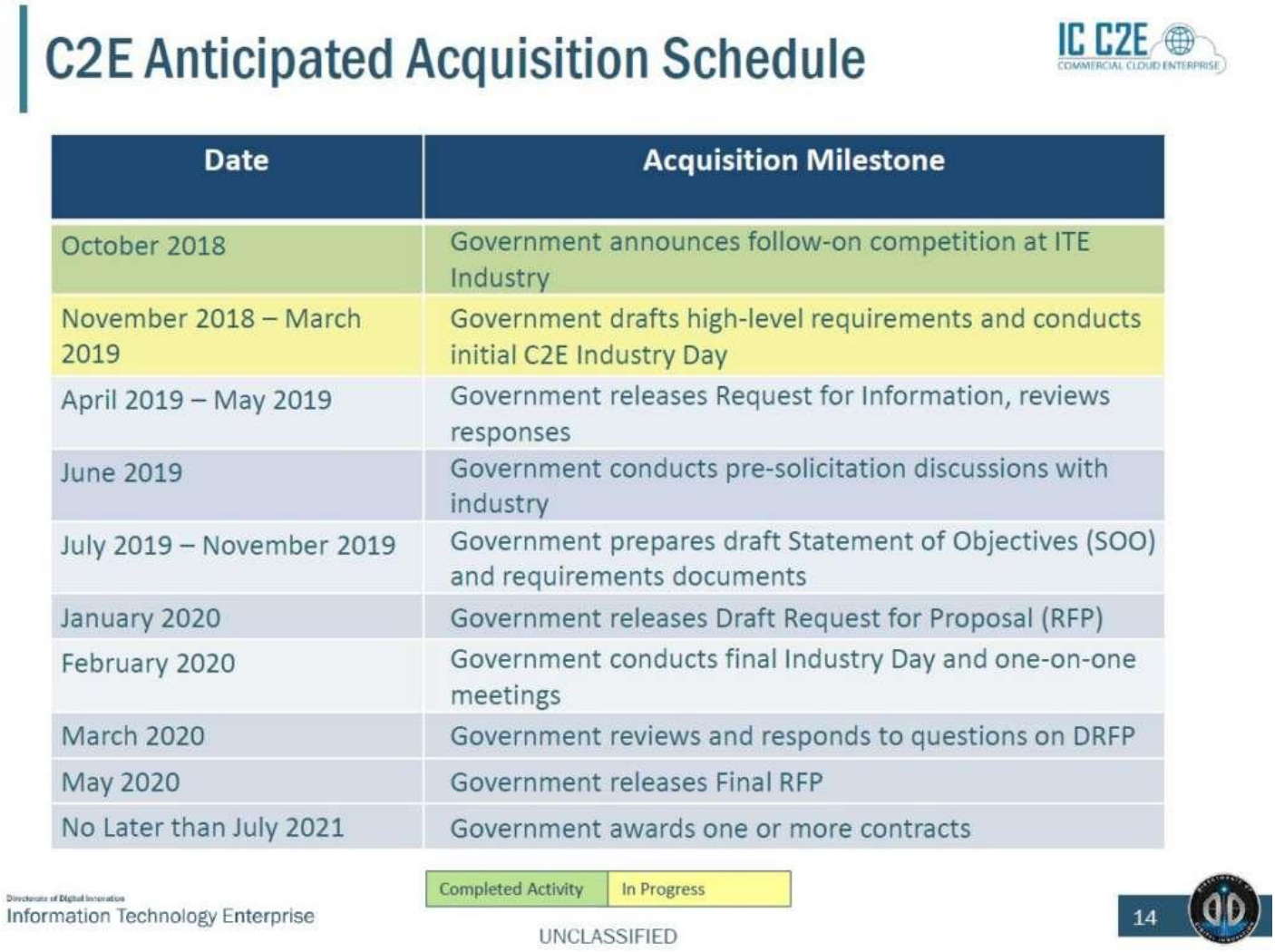
DoD currently is conducting an internal review of JEDI after a bid protest from Oracle highlighted a potential conflict of interest. Additionally, DoD and JEDI are facing a potential FBI investigation.

Sam Gordy, the general manager of IBM federal, said the CIA strategy with C2E should not only inform DoD, but influence the Pentagon's plans going forward.

"These [C2E and JEDI] are diametrically opposed approaches. Clearly the CIA has five-to-six years of experience in a single cloud environment and they are making a strategic decision to wholeheartedly move into multi cloud world. It's a critical next step for the evolution of IT support for the IC," Gordy said in an interview with Federal News Network. "DoD should take advantage of those five-to-six years of experience in the IC and the national security community to inform what they are doing going forward."

Gordy said the CIA is taking the approach that the private sector has moved to over the last few years. He added that unlike JEDI, the CIA is making it clear why the multi-cloud approach is necessary because they are saying in the industry day documents and the market survey what they want to use the cloud for today and in the future.

Under phase 1, the CIA said it wants vendors to provide infrastructure-, platform- and software-as-a-service capabilities as well as support services.



Source: CIA industry day presentation from March 22, 2019.

"Knowing they have an enterprisewide cloud contract already and that they are using that capability, this tells me they need hybrid, on-premise and commercial solution and this creates a mechanism to do that," Hodgkin said. "I didn't see anything shocking or that caught me off guard. The CIA has clearly spelled out to the industrial base what they need, and one of them is to deliver some or all of the three types of cloud, and when they put their data into those clouds, it must be portable so they can move it to another cloud or somewhere else. Those are the two elements that are

different than what they have now, and ones that you haven't seen it called out in previous acquisitions, at least not at this level."

CIA needs cloud diversity, data portability

John Weiler, the executive director of the IT Acquisition Advisory Council and an outspoken critic of JEDI, said the CIA's approach for C2E is a recognition that the current C2S contract isn't working like they expected.

"If it had worked they would've just resigned up with Amazon Web Services," Weiler said. "One cloud can't solve all your problems. When you look at workloads on Oracle or legacy Microsoft platforms, it makes no sense to move them to Amazon or Google or IBM. Those clouds are not designed for those environments. These strategies to be effective have to acknowledge that there are certain platforms that are legacy can move to a specific cloud and not just to any cloud."

One industry source, who requested anonymity in order to talk about inner working of the IC, said there have been varying degrees of unhappiness with the Amazon contract, including at least two IC agencies rejecting the C2S cloud and building their own.

Another industry source said in many ways C2S was a long-term pilot and now the CIA and others in the IC recognize they weren't happy with the price they were getting for cloud services, interoperability was more difficult than first imagined especially between C2S and existing data centers, and they were limited in the ability to add new features in a timely manner.

"They've had time to see what works and what doesn't, and they've realized cloud providers are becoming specialized. It's easier to

move workloads from on-premise to the cloud with the same vendor. They realized migrations can be expensive," the source said. "The CIA realized that cloud diversity and price competition help bring down costs. The industry and the CIA weren't in a position to do that six years ago, but now they are, which is good."

The first industry source added the IC had real concerns about vendor lock-in and how hard it was to move data between cloud infrastructures.

"I've heard that a lot that people didn't expect going into Amazon to have the level of lock-in that they have. Once they migrated data to Amazon, it became much more difficult to lift and shift to say a Microsoft cloud because the systems was configured in way that was only good for the Amazon cloud," the source said.

Implementation of cloud services is key

A third industry source was even more blunt about the C2S contract:

"AWS has relentlessly leveraged C2S since its inception, proclaiming to federal agencies that there was only one cloud service provider good enough for the CIA, so they needn't look further. But like a handsy, insecure boyfriend, it seems like AWS held the CIA a little too close, proudly boasting about their exclusive relationship while competing suitors flexed their innovation muscles," source said. "Not surprisingly, since the relationship first began, the CIA has noticed it has options and doesn't need to commit. So while it's understandable AWS wants to put a ring on it, the agency would clearly rather stay friends and play the field."

An AWS spokesman said they are excited about C2E and the CIA's intent to build on the existing C2S efforts.

“As a customer obsessed organization, we’re focused on driving innovation that supports the mission and spurs solutions that allow for missions to be performed better, faster, and in a more secure manner,” the spokesman said.

Weiler said no matter the strategy that the CIA or DoD chooses, the key is the implementation. He said nearly every agency needs to address legacy systems and the consistent challenge of cloud migration.

IBM’s Gordy said C2S shouldn’t be considered a failure by any means as it greatly helped inform the CIA’s current strategy.

“This does sync up with a recompete on C2S, but I don’t think C2E is in anyway a replacement for C2S,” he said. “The CIA will probably continue to have the need for a broad business application cloud which is what C2S is being used for today. And then they will need to have a mission oriented cloud, which is the reason they are going to C2E, which seems to be for the optimization of those mission workloads.”

WHAT IS THE CIA TRIAD?

Understanding the significance of the three foundational information security principles.

In the information security (InfoSec) community, “CIA” has nothing to do with a certain well-recognized US intelligence agency. These three letters stand for *confidentiality*, *integrity*, and *availability*, otherwise known as the *CIA Triad*.

Together, these three principles form the cornerstone of any organization’s security infrastructure; in fact, they (should) function as goals and objectives for every security program. The CIA Triad is so foundational to information security that anytime data is leaked, a system is attacked, a user takes phishing bait, an account is hijacked, a website is maliciously taken down, or any number of other security incidents occurs, you can be certain that one or more of these principles have been violated.

Security professionals evaluate threats and vulnerabilities based on the potential impact they have on the confidentiality, integrity, and availability of an organization's assets—namely, its data, applications, and critical systems. Based on that evaluation, the security team implements a set of security controls to reduce risk within their environment. In the next section, we'll provide precise and detailed explanations of these principles in the context of InfoSec, and then look at real-world applications of these principles.

Confidentiality

Confidentiality refers to an organization's efforts to keep their data private or secret. In practice, it's about controlling access to data to prevent unauthorized disclosure. Typically, this involves ensuring that *only those who are authorized have access to specific assets* and that *those who are unauthorized are actively prevented from*

obtaining access. As an example, only authorized Payroll employees should have access to the employee payroll database. Furthermore, within a group of authorized users, there may be additional, more stringent limitations on precisely which information those authorized users are allowed to access. Another example: it's reasonable for ecommerce customers to expect that the personal information they provide to an organization (such as credit card, contact, shipping, or other personal information) will be protected in a way that prevents unauthorized access or exposure.

Confidentiality can be violated in many ways, for example, through direct attacks designed to gain unauthorized access to systems, applications, and databases in order to steal or tamper with data. Network reconnaissance and other types of scans, electronic eavesdropping (via a man-in-the-middle attack), and escalation of system privileges by an attacker are just a few examples. But confidentiality can also be violated unintentionally through human error, carelessness, or inadequate security controls. Examples include failure (by users or IT security) to adequately protect passwords; sharing of user accounts; physical eavesdropping (also known as shoulder surfing); failure to encrypt data (in process, in transit, and when stored); poor, weak, or nonexistent authentication systems; and theft of physical equipment and storage devices.

Countermeasures to protect confidentiality include data classification and labeling; strong access controls and authentication mechanisms; encryption of data in process, in transit, and in storage; steganography; remote wipe capabilities; and adequate education and training for all individuals with access to data.

Integrity

In everyday usage, integrity refers to the quality of something being whole or complete. In InfoSec, integrity is about *ensuring that data has not been tampered with and, therefore, can be trusted. It is correct, authentic, and reliable*. Ecommerce customers, for example, expect product and pricing information to be accurate, and that quantity, pricing, availability, and other information will not be altered after they place an order. Banking customers need to be able to trust that their banking information and account balances have not been tampered with. Ensuring integrity involves protecting data in use, in transit (such as when sending an email or uploading or downloading a file), and when it is stored, whether on a laptop, a portable storage device, in the data center, or in the cloud.

As is the case with confidentiality, integrity can be compromised directly via an attack vector (such as tampering with intrusion detection systems, modifying configuration files, or changing system logs to evade detection) or unintentionally, through human error, lack of care, coding errors, or inadequate policies, procedures, and protection mechanisms.

Countermeasures that protect data integrity include encryption, hashing, digital signatures, digital certificates, intrusion detection systems, auditing, version control, and strong authentication mechanisms and access controls.

Note that integrity goes hand in hand with the concept of *non-repudiation: the inability to deny something*. By using digital signatures in email, for example, a sender cannot deny having sent a message, and the recipient cannot claim the message received was

different from the one sent. Non-repudiation assists in ensuring integrity.

Availability

Systems, applications, and data are of little value to an organization and its customers if they are not accessible when authorized users need them. Quite simply, availability means that *networks, systems, and applications are up and running*. It ensures that *authorized users have timely, reliable access to resources when they are needed*.

Many things can jeopardize availability, including hardware or software failure, power failure, natural disasters, and human error. Perhaps the most well-known attack that threatens availability is the denial-of-service attack, in which the performance of a system, website, web-based application, or web-based service is intentionally and maliciously degraded, or the system becomes completely unreachable.

Countermeasures to help ensure availability include redundancy (in servers, networks, applications, and services), hardware fault tolerance (for servers and storage), regular software patching and system upgrades, backups, comprehensive disaster recovery plans, and denial-of-service protection solutions.

Applying the Principles

Depending on an organization's security goals, the industry, the nature of the business, and any applicable regulatory requirements, one of these three principles might take precedence over another. For example, confidentiality is vital within certain government agencies (such as intelligence services); integrity takes priority in the financial sector where the difference between $1.00 and

$1,000,000.00 could be catastrophic; and availability is critical in both the ecommerce sector (where downtime can cost companies millions of dollars), and the healthcare sector (where human life could be lost if critical systems are unavailable).

A key concept to understand about the CIA Triad is that prioritizing one or more principles can mean the tradeoff of others. For example, a system that requires high confidentiality and integrity might sacrifice lightning-speed performance that other systems (such as ecommerce) might value more highly. This tradeoff is not necessarily a bad thing; it is a conscious choice. Each organization must decide how to apply these principles given their unique requirements, balanced with their desire to provide a seamless and safe user experience.

Leaks of NSA, CIA Tools Have Leveled Nation-State Cybercriminal Capabilities

The wide availability of tools leaked by the Shadow Brokers and WikiLeaks in 2016 and 2017 has given emerging cyber powers a way to catch up, DarkOwl says.

The public leaks of classified NSA and CIA hacking tools in 2016 and 2017 appear to have leveled the playing field for nation-state cybercriminals to some extent, new research shows.

Threat intelligence firm DarkOwl recently analyzed Dark Web data gathered from public and proprietary sources and found the leaked cyber weapons have strengthened the ability of emerging nation-state actors to attack rivals and project attribution to others.

The NSA and CIA data — released publicly by a group called the Shadow Brokers and WikiLeaks, respectively — included an NSA espionage and mass-surveillance system called UNITEDRAKE, a multiplatform CIA malware suite called HIVE, and numerous documents describing sophisticated false-flag and other cyber-offense tactics.

The leaked cyber weapons have given adversaries new ways to capture text, video, and images from target systems, including the Internet of Things (IoT) and smart TVs; attack smart vehicles; hide implants in Windows and other operating systems; and conduct a range of other surreptitious actions. Significantly, the leaks also made widely available capabilities that let attackers conceal the origins of an attack or to make it appear as if an attack originated from somewhere else entirely.

Details on the NSA and CIA tools and processes have been extensively studied on the Dark Web and are now part of the arsenal of everyone from nation-state actors to ordinary cybercriminals, DarkOwl says. "The wide dissemination of cyber

weapons from the NSA and CIA has changed the international cyber battle space considerably," says Andrew Lewman, vice president at DarkOwl. "Sophisticated, weapons-grade cyber tools are available on the Dark Web and [are creating] numerous challenges in determining who could be behind various cyber campaigns."

The US, Russia, and China continue to be cyber superpowers in terms of skills, influence, money, and manpower. But other less powerful nations have acquired formidable strength because of their access to these previously unattainable tools. "At this time, we do not have enough intelligence to support a statement on what country has benefited the most from the leaks of these tools," Lewman says. But a generalized leveling of the playing field since the NSA and CIA leaks is clear, he notes.

In DarkOwl's assessment, Israel, Germany, and the UK rank behind the top three nations in their cyber capabilities, followed by Ukraine, France, Iran, and India. But it is Iran and North Korea that present a major threat to US interests in cyberspace, especially given their ongoing cooperation and collaboration in military and technology development, Lewman says.

Cyber proxies, specifically as contracted by the Kingdom of Saudi Arabia, are another increasing concern because previously the Kingdom displayed little to no cyber capabilities. "Financial resources and international influence is of concern for them and their role in international conflicts," Lewman says.

Utah Data Center

Background

The Utah Data Center, code-named Bumblehive, is the first Intelligence Community Comprehensive National Cyber-security Initiative (IC CNCI) data center designed to support the Intelligence Community's efforts to monitor, strengthen and protect the nation. Our Utah "massive data repository" is designed to cope with the vast increases in digital data that have accompanied the rise of the global network.

1913

NSA is the executive agent for the Office of the Director of National Intelligence (ODNI) and is the lead agency at the center.

The 1.5 billion-dollar one million square-foot Bluffdale / Camp Williams LEED Silver facility houses a 100,000 sq-ft mission critical Tier III data center. The remaining 900,000 SF is used for technical support and administrative space. Our massive twenty building complex also includes water treatment facilities, chiller plants, electric substation, fire pump house, warehouse, vehicle inspection facility, visitor control center, and sixty diesel-fueled emergency standby generators and fuel facility for a 3-day 100% power backup capability.

Utah Data Center Administration Building

Learn More About Our Secret Surveillance Activities

In recent months, numerous *Top Secret* documents have been leaked to the media relating to surveillance activities carried out by our Intelligence Community. In an effort to increase transparency, a new website called "IC OFF THE RECORD" was created to provide the American People immediate, ongoing and direct access to these unauthorized leaks.

Click on the Photo Below to View Utah Data Center Up Close at High Resolution

June 2013

Utah Data Center Technical Specifications

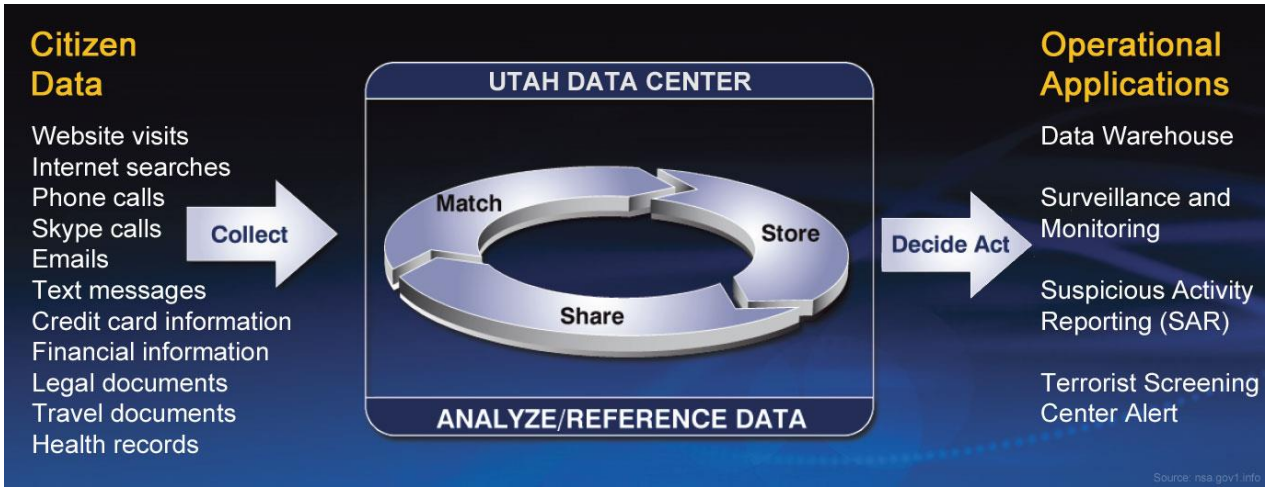
Data Storage Capacity

In February 2012, Utah Governor Gary R. Herbert revealed that the Utah Data Center would be the "first facility in the world expected to gather and house a yottabyte". Since then, conflicting media reports have also estimated our storage capacity in terms of zettabytes and exabytes. While the actual capacity is classified for NATIONAL SECURITY REASONS, we can say this: The Utah Data Center was built with future expansion in mind and the ultimate capacity will definitely be "alottabytes"!

The steady rise in available computer power and the development of novel computer platforms will enable us to easily turn the huge volume of incoming data into an asset to be exploited, for the good of the nation.

Learn more about the domestic surveillance data we process and store in the Utah Data Center. Also, view our strategy for using the PRISM data collection program, nationwide intercept stations, and the "Boundless Informant" mapping tool to gather and track

this data.



Code-Breaking Supercomputer Platform

The Utah Data Center is powered by the massively parallel Cray XC30 supercomputer which is capable of scaling high performance computing (HPC) workloads of more than 100 petaflops or 100,000 trillion calculations each second.

Code-named "Cascade", this behemoth was developed in conjunction with the Defense Advanced Research Projects Agency (DARPA) to meet the demanding needs of the Intelligence Community.

Our Ultimate Target: 256-bit AES

The Advanced Encryption Standard (AES) algorithm is used worldwide to encrypt electronic data on hard drives, email systems, and web browsers. Computer experts have estimated it would take longer than the age of the universe to break the code using a trial-and-error brute force attack with today's computing technology.

In 2004, the NSA launched a plan to use the Multiprogram Research Facility in Oak Ridge, Tennessee to build a classified supercomputer designed specifically for cryptanalysis targeting the AES algorithm. Our classified NSA Oak Ridge facility made a stunning breakthrough that is leading us on a path towards building the first exaflop machine (1 quintillion instructions per second) by 2018. Since the capability to break the AES-256 encryption key within an actionable time period may still be decades away, our Utah facility is sized to store all encrypted (and thereby suspicious) data for safekeeping.

Photos of the Utah Data Center

We are pleased to announce that as of 8/31/2014, the Utah Data Center is 100% complete. Here is a nice collection of wintry photos taken in early 2014

Aerial view of the Utah Data Center - 2014

Work at the Utah Data Center

Are you interested in a career in Domestic Surveillance? Check out our Utah Data Center jobs page for exciting employment opportunities.

Operating the Utah Data Center

It takes a large dedicated team to keep a data center up and running. Here are just some of the positions staffed by contractors at the UDC:

Mechanical Engineer

Responsible for mechanical systems, to include chiller plants, cooling distribution, cooling towers, water treatment, air handler units, generator plants, and fuel distribution systems. More details

Network Engineer

Responsible for Real Time Database and SCADA network. More details

1919

High Voltage Electrician

Install or repair electrical equipment such as generators, transformers, switchboards, controllers, circuit breakers, motors, heating units, conduit systems, or other transmission equipment. More details

Mechanical Technician

Responsible for the facility Heating, Ventilation, Air Conditioning (HVAC) systems. More details

Controls Engineer/Electrical Systems Operator

Operate and maintain the Data Center's Electrical Systems providing oversight and incorporating SCADA and EMCS control system changes. More details

Facilities Control Center Watch Officer

Plan and manage operations, maintenance and all other activities for assigned shift at a 24-hour data center facility to ensure safe, compliant, and productive operations. More details

Fuel Distribution Operator

Responsible for the fuel oil receiving, storage, and distribution system. More details

Quality Specialist

Responsible for developing, implementing and administering quality programs and plans affecting compliant and reliable operations. More details

Systems Analyst

Provides technical support for data center asset management system by performing all database administrative duties. More details

Master Plumber

Installs, modifies and repairs utility, supply, and disposal systems, and equipment, such as water; mains, oil and gas distribution systems, air compression equipment, water closets, as well as all type of fire protection equipment. More details

Instrumentation & Electrical Technician

Performs routine maintenance, installation and integration of various types of process control instrumentation equipment such as transmitters, controllers, integrators, recorders, square root extractors and final drive elements. More details

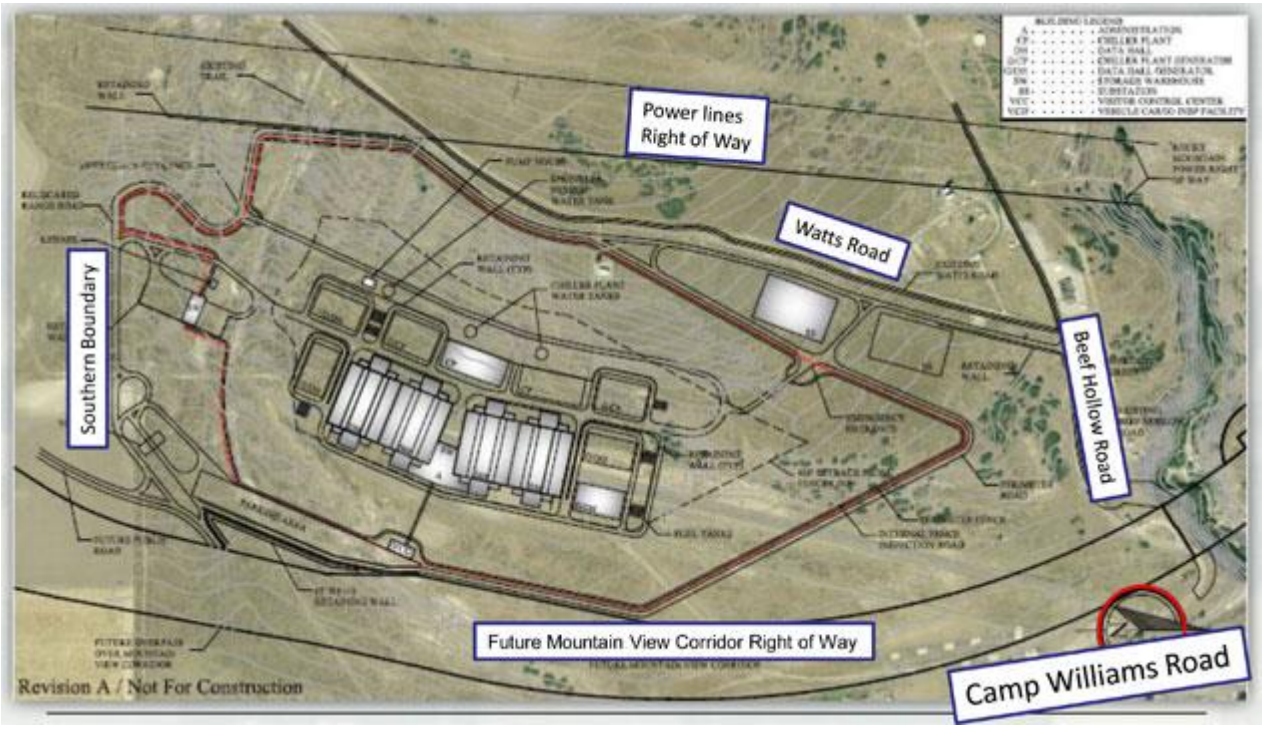
The Utah Data Center keeps watch 24 hours a day

What's Being Said About the Utah Data Center on Twitter

We are Americans first, last, and always. We treasure the U.S. Constitution and understand that a spirited debate is often a necessary precursor to acceptance.

Utah Data Center Site Plan

The site plan below shows the location of the administration building, chiller plant, data halls, generators, fuel storage tanks, storage warehouse, power substations, visitor control center, and vehicle inspection facility.



Utah Data Center Construction Photos: March 2012 - June 2013

Aerial front view of the Utah Data Center under construction - March 2012

October 2013 - Southern end of the Utah Data Center facility

Openness and Transparency: Our Cooperation with Privacy Groups

As proof of our genuine concern for privacy protection, we recently gave permission for several privacy groups to fly their little blimp over our massive data center. We would like to thank these airborne privacy pioneers for the stunning photo below of our impressive facility. By allowing harmless publicity stunts like these, we can have our data and store it too.

Utah Data Center

1924

Aerial view of the UDC construction site - July 2012

1925

Last two data halls still under construction - July 2012

Chiller plant, generator plant, and tank storage under construction - July 2012

Close-up view of completed data hall and Administration building - June 2013

Four completed data halls and the Administration building - June 2013

Completed chiller/generator plants and fuel tanks - June 2013

Close-up view of the cooling units - June 2013

Storage warehouse, canine kennel, and sprinkler system water tank - June 2013

Vehicle cargo inspection facility

October 2013 - Southern end of the Utah Data Center facility

1929

The Utah Data Center - On the Map

View the NSA Utah Data Center on Google Maps and Bing Maps

1595, 3949 s 200 E suite b1, Salt Lake City, UT 84107, United States

1595, 3949 200 E suite b1, Salt Lake City, UT 84107, United States

IBM updates cloud strategy after CIA loss

Two years after losing high-profile government work to Amazon Web Services, IBM has revamped the way it structures enterprise cloud services contracts, thanks in part to its US$2 billion acquisition of cloud services provider SoftLayer.

“The loss of the CIA contract to AWS was obviously a blow to IBM,” said Charles King, head of the IT analyst firm Pund-IT, responding to a query in email. “Whether it will make the company more competitive is hard to say until a similar deal comes along. But I believe IBM’s cloud solutions are considerably stronger with SoftLayer than they were before.”

In 2012, IBM protested the Central Intelligence Agency's award of a $600 million contract vehicle to competing vendor AWS. Reviewing the case, the Government Accountability Office agreed with the CIA that IBM did not show sufficient evidence that it could easily scale up cloud operations to meet increasing demands, and ranked AWS as having more technical skills to offer in the area of cloud services.

That IBM is widely used within the U.S. government and AWS was a relative novice to supplying such services to this market did not reflect well on IBM's then-nascent set of cloud services.

Since the SoftLayer purchase last July, IBM has used the company's assets to power its growing portfolio of hosted services, such as the Bluemix set of hosted services. SoftLayer also provides the underpinning for a growing number of services based on the company's Watson intelligent assistant technology, as well as a set of infrastructure services that was widely used even before the IBM purchase.

On Monday, IBM announced that SoftLayer has snagged 6,000 customers in the past year, including some high-profile names such as the department store Macy's and appliance-maker Whirlpool.

Other enterprise customers IBM name-checked included Heineken, Philips Smart TV, Radio Shack, University of Pittsburgh Medical Center, Honda, Toyota, Pacific Gas & Electric, L'oreal, AT&T, Visa and Continental Airlines.

“We’ve made significant inroads in every industry in the world, whether it is retail, industrial manufacturing, health care, financial, insurance,” said Lance Crosby, CEO of IBM SoftLayer.

The customer wins suggests that much has changed since IBM lost the CIA bid, and the changes reflect IBM’s growing sophistication in offering cloud services. Enterprise IT decision makers will often look to a potential vendor’s customer list as a validation of the vendor’s service.

In the government market, IBM is now building two facilities dedicated for federal government use, in Ashburn, Virginia, (outside of Washington) and in Dallas. These facilities will meet the government’s requirements for security and data privacy.

Also, over the past several months, IBM Global Business Services sales people have been trained to sell cloud to the federal agencies, Crosby said.

Perhaps most importantly, IBM’s billing mechanisms have changed to better reflect a customer’s usage in the cloud. Billing was one of the chief issues with IBM’s original proposal to the CIA.

“Prior to SoftLayer, the previous billing to the cloud was not by the hour, like SoftLayer is,” Crosby said. “These two new facilities will have everything available by the hour, like other cloud providers.”

The work seems to be paying off: IBM has recently won contracts from U.S. federal government agencies such as the U.S. Department of Interior, the General Services Administration, Housing and Urban Development, the Defense Advanced Research Projects Agency, National Archives and Records Administration, the Department of Veterans Affairs and the National Weather Service.

IBM has invested a lot of energy in bulking up its cloud services since last year.

The company is investing $1.2 billion in building out data centers, working to bring the total to 40 around the globe by 2015. IBM is also investing an additional $1 billion to round out its Bluemix portfolio of services, and has purchased three companies—Cloudant, Silverpop and Aspera—just to embed their cloud management technologies into SoftLayer.

“We’re still in early days for cloud in terms of both cloud offerings and customer adoption. While I expect cloud services and providers will continue to evolve, IBM is a strong foundation to build on,” King wrote.

About half of the 6,000 SoftLayer customers are existing IBM customers looking to move at least some of their operations to the cloud, Crosby said.

This is not surprising. Nearly half of all large enterprises will move at least some of their operations to hosted environments, IT analyst firm Gartner has estimated.

One big driver for new business has been mobile computing, Crosby said. Banks, for instance, have been scrambling to offer mobile services, even if their back-end operations aren’t prepared for this new type of workload.

“Five years ago, people checked their bank account balance maybe once or twice a month. Now they are checking it on their smartphone 10 times a day. Mainframes weren’t built for Web levels of interaction, so we’re doing that type of interaction out of the cloud,” Crosby said.

Macy’s, which operates 840 department stores in the U.S., obtained IBM hosted services, and a private connection to IBM’s

data centers, in order to offer more computerized interactions with its customers, as well as to set up a disaster recovery system that would resume operations should Macy's primary systems fail.

"Macy's is really doing big data analytics to understand their customers better," Crosby said. The company wants to go past the routine of addressing potential customers by marketing segments, and is looking at ways to customize offers and other material for each person.

In addition to the customer wins, IBM is also announcing a number of new cloud services.

The company launched Watson Engagement Advisor, a service that can help businesses in marketing, sales and customer service.

A high-speed transfer technology, called Aspera, provides a way to move large amounts of data across SoftLayer data centers. A new storage service, called Elastic Storage on Cloud, provides hosted storage.

The company has also issued new services for managing IBM services as well as an adapter, called JumpGate, for moving OpenStack workloads across different cloud providers.

Our Data Centers

AWS pioneered cloud computing in 2006, creating cloud infrastructure that allows you to securely build and innovate faster. We are continuously innovating the design and systems of our data centers to protect them from man-made and natural risks. Then we implement controls, build automated systems, and undergo third-party audits to confirm security and compliance. As a result, the most highly-regulated organizations in the world trust AWS every day. Take a virtual tour of one of our data centers to learn about our security approach to protect the data of millions of active monthly customers.

PERIMETER LAYER

EXPLORE »

PERIMETER LAYER

AWS data center physical security begins at the Perimeter Layer. This layer includes a number of security features depending on the location, such as security guards, fencing, security feeds, intrusion detection technology, and other security measures.

DATA LAYER

The Data Layer is the most critical point of protection because it is the only area that holds customer data. Protection begins by restricting access and maintaining a separation of privilege for each layer. In addition, we deploy threat detection devices and system protocols, further safeguarding this layer.

INFRASTRUCTURE LAYER

EXPLORE »

INFRASTRUCTURE

The Infrastructure Layer is the data center building and the equipment and systems that keep it running. Components like back-up power equipment, the HVAC system, and fire suppression equipment are all part of the Infrastructure Layer.

ENVIRONMENTAL LAYER

The Environmental Layer is dedicated to environmental considerations from site selection and construction to operations and sustainability. AWS carefully chooses our data center locations to mitigate environmental risk, such as flooding, extreme weather, and seismic activity.

Perimeter Layer

AWS data center physical security begins at the Perimeter Layer. This Layer includes a number of security features depending on the location, such as security guards, fencing, security feeds, intrusion detection technology, and other security measures. Scroll down to learn more about the types of security measures we have in place within the Perimeter Layer of the data centers we operate around the world.

PERIMETER LAYER

ACCESS IS SCRUTINIZED

AWS restricts physical access to people who need to be at a location for a justified business reason. Employees and vendors who have a need to be present at a data center must first apply for access and provide a valid business justification. The request is reviewed by specially designated personnel, including an area access manager. If access is granted, it is revoked once necessary work is completed.

ENTRY IS CONTROLLED AND MONITORED

Entering the Perimeter Layer is a controlled process. We staff our entry gates with security officers and employ supervisors who monitor officers and visitors via security cameras. When approved individuals are on site, they are given a badge that requires multi-factor authentication and limits access to pre-approved areas.

AWS DATA CENTER WORKERS ARE SCRUTINIZED, TOO

AWS employees who routinely need access to a data center are given permissions to relevant areas of the facility based on job function. But their access is regularly scrutinized, too. Staff lists are routinely reviewed by an area access manager to ensure each employee's authorization is still necessary. If an employee doesn't have an ongoing business need to be at a data center, they have to go through the visitor process.

MONITORING FOR UNAUTHORIZED ENTRY

We are continuously watching for unauthorized entry on our property, using video surveillance, intrusion detection, and access log monitoring systems. Entrances are secured with devices that sound alarms if a door is forced or held open.

AWS SECURITY OPERATIONS CENTERS MONITORS GLOBAL SECURITY

AWS Security Operations Centers are located around the world and are responsible for monitoring, triaging, and executing security programs for our data centers. They oversee physical access management and intrusion detection response while also providing global, 24/7 support to the on-site data center security teams. In short, they support our security with continuous monitoring activities such as tracking access activities, revoking access permissions, and being available to respond to and analyze a potential security incident.

Data Layer

The Data Layer is the most critical point of protection because it is the only area that holds customer data. Protection begins by restricting access and maintaining a separation of privilege for each layer. In addition, we deploy threat detection devices, video surveillance and system protocols, further safeguarding this layer. Scroll down to learn more about some of the security measures we have in place within the Data Layer.

DATA LAYER

TECHNOLOGY AND PEOPLE WORK TOGETHER FOR ADDED SECURITY

There are mandatory procedures to obtain authorization to enter the Data Layer. This includes review and approval of a person's access application by authorized individuals. Meanwhile, threat and electronic intrusion detection systems monitor and automatically trigger alerts of identified threats or suspicious activity. For example, if a door is held or forced open an alarm is triggered. We deploy security cameras and retain footage in alignment with legal and compliance requirements.

PREVENTING PHYSICAL AND TECHNOLOGICAL INTRUSION

Access points to server rooms are fortified with electronic control devices that require multi-factor authorization. We're also prepared to prevent technological intrusion. AWS servers can warn employees of any attempts to remove data. In the unlikely event of a breach, the server is automatically disabled.

SERVERS AND MEDIA RECEIVE EXACTING ATTENTION

Media storage devices used to store customer data are classified by AWS as Critical and treated accordingly, as high impact, throughout their life-cycle. We have exacting standards on how to install, service, and eventually destroy the devices when they are no longer useful. When a storage device has reached the end of its useful life, AWS decommissions media using techniques detailed in NIST 800-88. Media that stored customer data is not removed from AWS control until it has been securely decommissioned.

THIRD-PARTY AUDITORS VERIFY OUR PROCEDURES AND SYSTEMS

AWS is audited by external auditors on more than 2,600 requirements throughout the year. When third-party auditors inspect our data centers they do a deep dive to confirm we're following established rules needed to obtain our security certifications. Depending on the compliance program and its requirements, external auditors may interview AWS employees about how they handle and dispose of media. Auditors may also watch security camera feeds and observe entrances and hallways throughout a data center. And they often examine equipment such as our electronic access control devices and security cameras.

Infrastructure Layer

The Infrastructure Layer is the data center building and the equipment and systems that keep it running. Components like back-up power equipment, the HVAC system, and fire suppression equipment are all part of the Infrastructure Layer. These devices and systems help protect servers and ultimately your data. Scroll down to learn more about the types of security measures we deploy in the Infrastructure Layer of our data centers.

INFRASTRUCTURE LAYER

LAYER-BY-LAYER ACCESS REVIEW

Like other layers, access to the Infrastructure Layer is restricted based on business need. By implementing a layer-by-layer access review, the right to enter every layer is not granted by default. Access to any particular layer is only granted if there is a specific need to access that specific layer.

MAINTAINING EQUIPMENT IS A PART OF REGULAR OPERATIONS

AWS teams run diagnostics on machines, networks, and backup equipment to ensure they're in working order now and in an emergency. Routine maintenance checks on data center equipment and utilities are part of our regular operations.

EMERGENCY-READY BACKUP EQUIPMENT

Water, power, telecommunications, and internet connectivity are designed with redundancy, so we can maintain continuous operations in an emergency. Electrical power systems are designed to be fully redundant so that in the event of a disruption, uninterruptible power supply units can be engaged for certain functions, while generators can provide backup power for the entire facility. People and systems monitor and control the temperature and humidity to prevent overheating, further reducing possible service outages.

Environmental Layer

The Environmental Layer is dedicated to environmental considerations from site selection and construction to operations and sustainability. AWS carefully chooses our data center locations to mitigate environmental risk, such as flooding, extreme weather, and seismic activity. Scroll down to learn more about the types of security measures we have in place within the Environmental Layer.

ENVIRONMENTAL LAYER

PREPARED FOR THE UNEXPECTED

AWS proactively prepares for potential environmental threats, like natural disasters and fire. Installing automatic sensors and responsive equipment are two ways we safeguard our data centers. Water-detecting devices can alert employees to problems as automatic pumps work to remove liquid and prevent damage. Similarly, automatic fire detection and suppression equipment reduces risk and can notify AWS employees and firefighters of a problem.

HIGH AVAILABILITY THROUGH MULTIPLE AVAILABILITY ZONES

Unlike virtually every other technology infrastructure provider, each AWS Region has multiple Availability Zones. Each Availability Zone consists of one or more data centers, are physically separated from one another, and have redundant power and networking. Customers who care about high availability and performance of their applications can deploy applications across multiple Availability Zones in the same region for fault tolerance and low latency. Availability Zones are connected to each other with fast, private fiber-optic networking, enabling you to easily architect applications that automatically fail-over between Availability Zones without interruption.

SIMULATING DISRUPTIONS & MEASURING OUR RESPONSE

The AWS Business Continuity Plan is an operations process guide outlining how to avoid and lessen disruptions due to natural disasters with detailed steps to take before, during, and after an event. To mitigate and prepare for the unexpected, AWS tests the Business Continuity Plan regularly with drills that simulate different scenarios. We document how our people and processes perform, then debrief on lessons learned and any corrective actions that may be needed to improve our response rate. We are trained and ready to rebound from

disruptions quickly, which includes a methodical recovery process to minimize further downtime due to errors.

GREENER IN THE AWS CLOUD

In addition to addressing environmental risks, we also incorporate sustainability considerations into our data center design. AWS has a long-term commitment to use 100% renewable energy. When companies move to the AWS Cloud from on-premises infrastructure, they typically reduce carbon emissions by 88% because our data centers can offer environmental economies of scale. Organizations generally use 77% fewer servers, 84% less power, and tap into a 28% cleaner mix of solar and wind power in the AWS Cloud versus their own data centers.

Google is investing $3.3B to build clean data centers in Europe

Google announced today that it was investing €3 billion (approximately US$3.3 billion) to expand its data center presence in Europe. What's more, the company pledged the data centers would be environmentally friendly.

This new investment is in addition to the $7 billion the company has invested since 2007 in the EU, but today's announcement was focused on **Google's** commitment to building data centers

running on clean energy as much as the data centers themselves.

In a blog post announcing the new investment, CEO Sundar Pichai made it clear that the company was focusing on running these data centers on carbon-free fuels, pointing out that he was in Finland today to discuss with prime minister Antti Rinne building sustainable economic development in conjunction with a carbon-free future.

Of the €3 billion the company plans to spend, it will invest €600 million to expand its presence in Hamina, Finland, which he wrote "serves as a model of sustainability and energy efficiency for all of our data centers." Further, the company already announced 18 new renewable energy deals earlier this week, which encompass a total of 1,600-megawatts in the U.S., South America and Europe.

In the blog post, Pichai outlined how the new data center projects in Europe would include some of these previously announced projects:

Today I'm announcing that nearly half of the megawatts produced will be here in Europe, through the launch of 10 renewable energy projects. These agreements will spur the construction of more than 1 billion euros in new energy infrastructure in the EU, ranging from a new offshore wind project in Belgium,

to five solar energy projects in Denmark, and two wind energy projects in Sweden. In Finland, we are committing to two new wind energy projects that will more than double our renewable energy capacity in the country, and ensure we continue to match almost all of the electricity consumption at our Finnish data center with local carbon-free sources, even as we grow our operations.

The company is also helping by investing in new skills training, so people can have the tools to be able to handle the new types of jobs these data centers and other high-tech jobs will require. The company claims it has previously trained 5 million people in Europe for free in crucial digital skills, and recently opened a Google skills hub in Helsinki.

It's obviously not a coincidence that the company is making an announcement related to clean energy on Global Climate Strike Day, a day when people from around the world are walking out of schools and off their jobs to encourage world leaders and businesses to take action on the climate crisis. Google is attempting to answer the call with these announcements.

Chapter 13:

How to Build a Big Data Center?

8 Steps to Building a Modern Datacenter

Mozilla packs 'em tight. *Photo: Ariel Zambelich/Wired*

Today's enterprises are beginning to hit a wall with their old-school datacenters. As noted in a recent article by IO's Patrick Flynn on why the datacenter is broken, datacenters have become too big and too slow when they really need to be more cost-effective, efficient and responsive. As it is, enterprise IT architects are struggling to keep pace with accelerating business demands for more storage and compute resources, and are unable to take full advantage of new technologies designed to improve infrastructure performance, scale and economics. No more building bigger and more expensive silos of proprietary hardware. What is needed is a complete rethinking of how datacenters are designed and managed.

Here are eight fundamental steps to creating a more efficient, manageable and scalable datacenter that evolves with your organization's needs:

1. Be Modular

Datacenter infrastructure gets more complex each year as new technologies get added, creating a mishmash of incompatible frameworks and consoles across network, server and storage silos. Switching to a modular design can afford enterprises far more simplicity and flexibility, allowing enterprise IT architects to add or remove building blocks as needed.

Over the years, "modularization" has evolved from 40-foot shipping containers filled with racks of equipment to much smaller and compact single rack solutions. For example, Virtual Computing Environment's (VCE) vBlock is a pre-engineered, fully cabled rack containing servers, network switches and storage devices. But for many companies, those devices are too pricey at $500,000 or more. They also incorporate fixed, vendor-defined ratios for computing resources and storage capacity, and are built with legacy components from multiple vendors that make overall management unnecessarily complex.

However, when building blocks can be quickly added to or removed from an infrastructure so you can have resources on-demand and avoid over-provisioning, you get true modularization. An increasingly popular approach is to use a single appliance that consolidates the compute and storage tiers. The modules are not only scalable on demand, but they're interoperable and streamline overall datacenter management with a single console, greatly reducing the headaches for overworked datacenter admins.

2. Converge When Possible

Enterprise IT managers have been moving to converged datacenter infrastructure because it uses fewer dedicated resources and is, therefore, more economical and more efficient. Storage convergence started more than a decade ago with hard disk drives migrating from servers to centralized shared storage arrays, connected via high-speed networks. More recently, flash memory has been added to enterprise storage devices to create hybrid storage solutions with up to 100 times faster than legacy architectures.

Rather than having specialized devices for computing and storage, the functions can be combined into one appliance. The datacenter is then built with a single resource tier containing all of the server and storage resources needed to power any application or workload. This improves scalability without the need to spend more on additional hardware or high-speed, dedicated networking equipment.

3. Let Software Drive

The days of expensive, specialized hardware in datacenters are ending. They aren't flexible or portable, and many are powered by field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs) that don't support new software capabilities that datacenters or cloud infrastructures today demand. Separating policy intelligence and runtime logic from the underlying hardware and abstracting it to a distributed software layer allows it to be automated and centrally controlled. This enables datacenter admins to provision new services without adding hardware, which saves on cost and offers more agility. And distributed applications can improve uptime, global scalability and service continuity during site failures.

4. Embrace Commodity Hardware

Google grew its Web search and other cloud services on the back of low-cost commodity hardware running distributed software. This innovative approach allowed it to scale fast with minimum investment. Traditional enterprises have been caught in an expensive cycle of upgrading datacenter hardware every three to five years, replacing it with newer, more expensive equipment. Today, they can reap the same benefits from commodity hardware that cloud providers do. A distributed software layer abstracts all resources across clusters of commodity nodes, delivering aggregate capacity that surpasses even the most powerful monolithic approaches. The value is in the software that powers low-cost hardware.

5. Empower End Users

Datacenters today need to be more resilient and reliable than ever. They must continue to handle traditional enterprise data needs, but also meet the growing demands from applications ranging from virtual desktop infrastructures (VDI) to employees toting handheld devices with them everywhere. To deal with the "consumerization" of IT, admins are moving to end-user computing models in which desktops, applications and data are centralized within the datacenter and accessed by employees via any device from anywhere. Modernizing the datacenter will enable datacenter managers to better address the wide range of workload demands brought on by the new "consumerization," as well as deal with compute-intensive VDI systems, storage-intensive enterprise data services (like Dropbox) or existing virtualized enterprise applications.

6. Break down Silos

The increasing complexity and functionality of datacenters has led to the formation of technology silos, with each managed by a team of specialists. For example, one team might handle the data management and information archive in the storage silo, while other teams oversee the networking, server and virtualization silos. Using combined appliances means you don't need separate teams of specialists for each technology. Integrating the technologies into a single scalable unit, or datacenter building block, reduces the need for highly specialized staff.

7. Go Hybrid

Many enterprises want to be able to use the public cloud for some things but still keep business-critical applications involving confidential data safe within the confines of the private datacenter. To meet these dual needs, corporations are using hybrid cloud environments. Public clouds offered by Amazon Web Services and others offer on-demand provisioning and resource-sharing across multiple tenants. Private clouds can do that too, but the difference is they remain under the management of the datacenter team and allow more control over of security, performance and service level agreements. Hybrid environments offer the best of both worlds.

8. Focus on Service Continuity

Enterprise disaster recovery strategies tend to be reactive. Consumerization, however, has radically altered user expectations. If there are interruptions or latency problems, users will go around enterprise IT and use unauthorized cloud-based services. To provide near 100 percent availability, admins have to be more proactive and focus on service continuity rather than disaster recovery. This means re-architecting datacenters to be highly

available, which means having a lot of bandwidth and low round-trip times. Also, enterprises should re-architect their applications to be distributed. By distributing applications architectures across multiple sites, regions or datacenters they can better scale globally, perform well and increase uptime. Facebook, Amazon and Google have seen great success with this model.

Enterprises are learning that to remain competitive they have to adapt to the rapidly changing business environment. They need to be able to increase data computing and storage capacity quickly and add new capabilities, but without spending a lot of extra money. With datacenters, it’s not about building out anymore; it’s about building smart.

Building a big data center of excellence

In today's business world big data can be a vital competitive differentiator for organizations. Traditionally, businesses needed to limit the scope of the data that they could use to make the critical decisions for driving successful business outcomes. Big data solutions have eased many of those limitations, so organizations can look at far more data and therefore make better decisions. Big data solutions and technologies such as Apache Hadoop and Apache Spark enable this fundamental shift.

I often ask customers what inhibits big data initiatives in their organization. Frequent answers include: no compelling business need, or difficulty identifying use cases; lack of data science skills; not enough staff to support them; and the complexity of collecting and managing the data. The concept of a center of excellence (CoE) for big data, which I attempt to demystify here, helps ensure these responses are not inhibitors in any organization.

The key to a data-driven business is in bringing data and insight to all workflows in the business and integrating it into the decision making at every step. This approach enables organizations to take advantage of the longitudinal analytics available with new technology

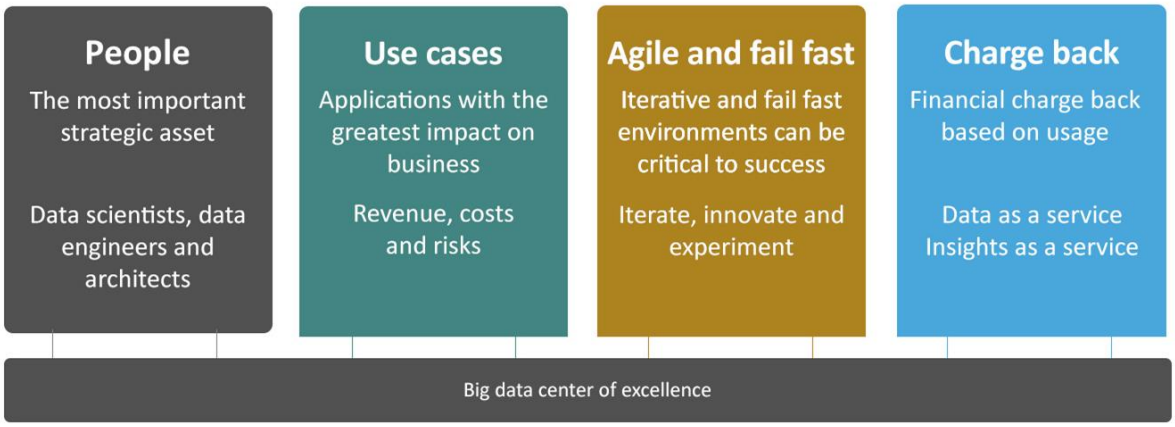
advances such as Hadoop and Spark as well as machine learning for past-, present- and future-looking analytics simultaneously.

Defining big data centers of excellence

A big data CoE is a framework that takes an organization from zero knowledge to having a fully functional practice of Hadoop, Spark and emerging open source technologies to deliver robust business results. A CoE is where organizations identify new technologies, learn new skills and develop appropriate processes that are then deployed into the business to accelerate adoption.

A centralized big data CoE can be the bedrock for establishing a data-driven company that treats data as a strategic asset. The big data CoE can partner with the business to identify data that is invaluable, explore use cases that differentiate its products and services in the market and help jump-start the business with insights that can yield real-time client value. Data’s strategic importance is the value it represents for the business, but success with big data is not just about data. The people

and the organization also play a vital role in that success:



As a strategic asset, the quality of data scientists, data engineers and data architects is paramount for creating a successful big data CoE. These data professionals need to be experienced practitioners who have a history of working within Hadoop and Spark ecosystems and possess vertical knowledge. Their goal is to prepare the business to build, run and operate production-quality big data applications on their own by maximizing their ability to leverage data. Data engineers, in particular, have to be committed to helping ensure the data—from acquisition to advocacy—functions as the organization's highly strategic asset.

And the other vital element for successful business outcomes is identifying strong leaders to lead the transformation. Strong leadership includes a business sponsor for the overall big data mission along with business stakeholders for individual use cases.

In many cases, the business comes up with the use cases, but the CoE has the responsibility of facilitating this work. The CoE needs to assume a leadership role in understanding which applications and use cases can be driven with available sets of data sources.

Sometimes businesses can be more proactive by bringing use cases to the CoE because the list of use cases can be overwhelming and put a strain on available resources. A transparent process for prioritizing these use cases is important and should be adopted. The CoE needs to prioritize use cases based on parameters such as ease of data availability, data quality, business revenue–based value and impact, costs and risks.

Applying agile methodology—the fail-fast approach

Agility and the ability to fail fast are essential to reaching the potential of big data. A lightweight agile process provides tools to deliver outcomes quickly and transparently, typically within two- to three-week sprints. The ability to fail fast is a key big data opportunity; business and technical roadmaps for delivering value need to change more often than in a traditional waterfall environment.

Data itself is also highly agile when it is collected in native form and transformed potentially many times to meet the needs of different use cases. Using the basic ideas of agile development methodology, a CoE can provide the leadership across the organization to ensure business users can quickly gain value from the data.

Developing financial models

At the heart of a big data CoE is creative financial models that support the innovation. The charge-back strategy can be a function of data as a service, insights as a service or analytics as a service.

As is often the case with shared services, a charge-back model is necessary to properly handle the maintenance and growth of the emerging technologies, which in this case can be Hadoop and Spark clusters. An organization needs to develop a charge-back model for the business units that will be engaging with the CoE for project, personnel, and infrastructure and application resources. Some important questions need to be considered when determining the charge-back model for business units:

- How many users will access the application and cluster?
- How much data will be ingested initially?
- How much data growth is expected over time?
- What is the data retention policy?

Business leaders and decision makers acknowledge that creating a data-driven organization requires a change of culture. Big data CoEs can be the key to this culture change. An important recommendation for building a CoE framework is starting with a small, secure data lake—a Hadoop- or Spark-based service—that can store and process data from various internal groups to support multiple use cases. When building a data lake, organizations learn and employ operational best practices for a number of processes:

- Cluster build out
- Data exploration
- Data ingestion and processing
- Disaster recovery
- General operations and maintenance
- Hadoop and Spark development
- Infrastructure integration
- Model building and testing
- Multitenancy and security
- Third-party software evaluation and integration
- Use-case evaluation

A leading telecommunications firm, for example, began by developing a CoE that asked each business division to come up with business use cases that would generate powerful insights through analytics. It then established regular training boot camps in which business users learned how to use data with self-service tools, and it created a community of data scientists and data engineers to support line-of-business managers in their analyses and to validate findings. As a result, this CoE enabled big data as a shared service that opened up the conversation for creative financial models that involve charge backs and show backs.

Leveraging big data centers of excellence

I foresee creative CoE adaptations such as the one just described helping businesses move beyond the hope of becoming a data-driven organization enabled by big data to the reality of an organization using a data-ingrained business model. If you're a working data scientist, data

engineer or data application developer, register to attend the IBM DataFirst Launch Event that takes place 27 September 2016 in New York, New York. You'll have the opportunity to engage with open source community leaders and practitioners and learn how to leverage your data analytics CoE to accelerate your transformation to a cognitive business.

CyrusOne data center in Carrollton, Texas

Think You Want to Build Your Own Data Center?

The economies of scale achieved by specialist data center providers make it hard to justify a dedicated enterprise build.

The construction of multi-megawatt hyperscale data centers by colocation companies has been accelerating globally. A shift in the construction techniques they use has allowed them to shorten their time to market and reduce construction costs. In a nutshell, the shift

has been from stick-built construction, where everything is assembled mostly or entirely on-site, to semi-prefabricated construction, where major pre-assembled modules are delivered to the construction site.

Time to market is paramount for colocation companies. Many are organized as Real Estate Investment Trusts (REITs) and make no money until tenants start deploying equipment racks to new data centers. And results of the recent push to shrink time to market have been impressive. CyrusOne, for example, claims to have set construction speed records at their Sterling II data center in Northern Virginia. The company said it completed a 30MW buildout in 180 days – from breaking ground to commissioning. It credited the accomplishment to its use of "massively modular" power distribution and chiller units.

This column's author, Hector Diaz, senior partner of iDiaz Advisors and a Data Center Institute board member, will be co-hosting a workshop on colocation strategies and techniques at Data Center World in Phoenix on Tuesday, March 19.

Modular power and cooling units are typically assembled in factories run by system integrators. The modules can be assembled simultaneously with construction of the concrete data center shell

and delivered on-site when needed. This concurrent fabrication results in:

- Shorter time to market through parallel construction and fabrication paths
- Increased reliability of prefabricated modules, as quality control can be tighter in a manufacturing environment than in a field construction environment
- Optimized deployment of capital, as capacity can be added to match projected data center occupancy

Capacity is typically added in well-defined increments. Today, those tend to be somewhere in the 2MW to 2.5MW range. This is all driven by economies of scale. If you look at the dollars-per-kW ratio for a generator set, the best “bang for the buck” is somewhere in the stated range. The dollar amount per kW will be higher if you deploy in smaller chunks. If you want to save space in your equipment yard and have fewer but larger generators, the dollars-per-kW goes up as well, plus, you are now dealing with special-order items with longer lead times. Economies of scale apply to all the other major components for the MEP (mechanical, electrical, and plumbing) fit-out of a new data center.

I have talked with top-tier colocation companies that claim their construction cost for a concurrently maintainable multi-megawatt data center is $6.5 million to $7 million per 1MW. There are large

data center providers that have not achieved these record-breaking efficiencies, citing per-MW construction costs in the $7 million to $10 million range – even while using pre-fabricated power and cooling modules. That's still much lower than the $12 million to $17 million a company pays for a typical under-1MW enterprise "stick build" data center.

As a result, I generally advise enterprise IT groups against building data centers unless they are looking at IT loads greater than 2MW. Surprisingly, many IT groups don't understand the true cost of building a data center, which includes eventual depreciation of all the capital sunk in construction, utilities, and facilities maintenance – costs typically paid by the real estate and facilities group, not by the CIO's organization.

Before You Start Building a New Data Center, Consider These Three Things

Getting a clear picture of what already exists helps determine if building a new data center is indeed the right option.

Ninety percent of the world's data was created in the last two years, and more than 2.5 exabytes of new data are created every day. This rapid growth underscores the need for more and more data centers to house and manage all this information. Yet, choosing to build a data center is not a simple task for an organization – these are long-term and costly projects. When an enterprise begins to think about the need for a new data center facility, much of the work takes place before ever breaking ground for construction. Without the proper planning, organizations risk creating a new data center that is too big, not optimized for efficiency, incurs higher capital costs than necessary, or is not adequate to meet both existing and future needs.

There are three key items to consider when planning to build a new data center to ensure the project meets the needs of the organization and constituencies it's designed to serve: existing facilities, the organization's business needs, and the data center and technology environment. Having answers to specific questions in these three areas will put your organization in a position to make the best choice for design, build, installation, and start-up of your new data center.

Existing Facilities

Key to understanding how to plan and build a new data center is knowing the details of your existing data center setup and capacity, whether that is your own facility, a colocation provider, or a cloud service. Getting a clear picture of what already exists helps determine if building a new data center is indeed the right option. It's possible that colocation, outsourcing, or a cloud or hybrid model could best meet your organization's data center needs.

Questions to ask at this stage:

- How old is the current data center?
- Does the current data center support our current needs and forecasted future needs?
- Is it possible to expand the existing facilities?

- What is the risk of expanding our existing data center and will outages be required?
- Would outsourcing, colocation, or a cloud/hybrid model give us the capacity we need now and in the foreseeable future?
- What are the costs of these alternative solutions?

Business Needs

As part of this evaluation process, it's also critical to understand the goals of the business, and how the data center meets the needs of the company. A key part of this is understanding the role of IT in the organization, as well as how the data center and its operations align with the business. That requires looking closely at how much data the business deals with, how it is gathered, stored, and accessed, who uses the data and how, and the relevant industry standards the data center must comply with.

Questions to ask while evaluating business needs:

- What are our corporate goals, and how does our IT infrastructure and data center support those goals?
- How have our IT or data center needs changed in the past three to five years?
- How do we anticipate our business to change in the next five to ten years, and what impact will that have on our IT and data center needs?
- Are we, or clients who will be using the data center, subject to regulatory compliance such as PCI data security standard to protect information of credit card owners or the Health Insurance Portability and Accountability Act (HIPAA) to safeguard personal medical information?

Data Center and Technology Environment

After getting a good sense of organizational goals and how the existing facilities are serving the business, it's also critical to understand the company's technology and data center physical environments.

Questions to ask about the physical environments:

- What computing, networking, and storage equipment do we currently have and what do we anticipate needing? A forecast of the future will provide physical space for future equipment and hardware refresh.
- How do we cool our existing data center, how do we plan to cool the new data center, and what equipment and utility access does that require?
- What are our existing power requirements for the computing and storage technology, as well as the cooling solution?
- What are our current fire suppression and security systems, and how will that need to change for the future?
- What is the physical size of our current data center facilities, and how do we anticipate that changing for future facilities?
- Do we have adequate data center monitoring that allows us to properly manage the infrastructure and plan for IT hardware installation in the future?

Data center projects are complex undertakings. When it's done right, a brand new data center will support your business objectives and meet current and anticipated future needs. The way to set up your organization for success with this project is to understand

existing data center facilities, know the business needs of the organization, and have knowledge of the necessary data center and technology environment.

The information gained from this assessment and evaluation process helps you determine the specific requirements for your new data center, get a clear understanding of total costs for the project and time it will take to complete, evaluate appropriate sites, and execute a construction plan. With this information, your organization will be able to create a data-driven and efficient plan to meet your data center needs today and tomorrow.

Step-by-Step: Build a Data Center

Overview of key issues facing IT leaders as they consider build-outs or upgrades of their data centers, including free tools and research notes to support the decision-making process

– 42% of IT leaders will face a data center build or upgrade in 2010 - one of the most expensive and complex IT projects, it can cost up to $1.6 M for a 1000 square foot data center.
– Each data center project is unique and should have its own detailed budget. Planning up front and establishing a clear project scope will minimize expensive changes in later project stages.
– Industry benchmarks are not always accurate – the absence of a budget risks cost overruns by 22%. Hidden costs for landscaping, permits, and security can increase capital investments. Ongoing operating expenses must also be budgeted, as they represent 65% of the total costs associated with building a data center.
– Accurate estimates of facility requirements are a critical input to the project's budget.
– Power and cooling supply requires careful

estimation to accommodate future growth yet avoid unwarranted costs.
– Fire protection is mandated by law and comes with real consequences as it can impact life safety.
– Site selection should identify any potential issues that may lead to additional construction costs, especially in an existing building.
– Hire a specialized general contractor and avoid those who build office spaces only. Data centers have more stringent, unique code requirements. Use specialized talent to avoid costly retrofit later.

Many IT organizations face a data center build but fail to recognize the design and engineering complexities.

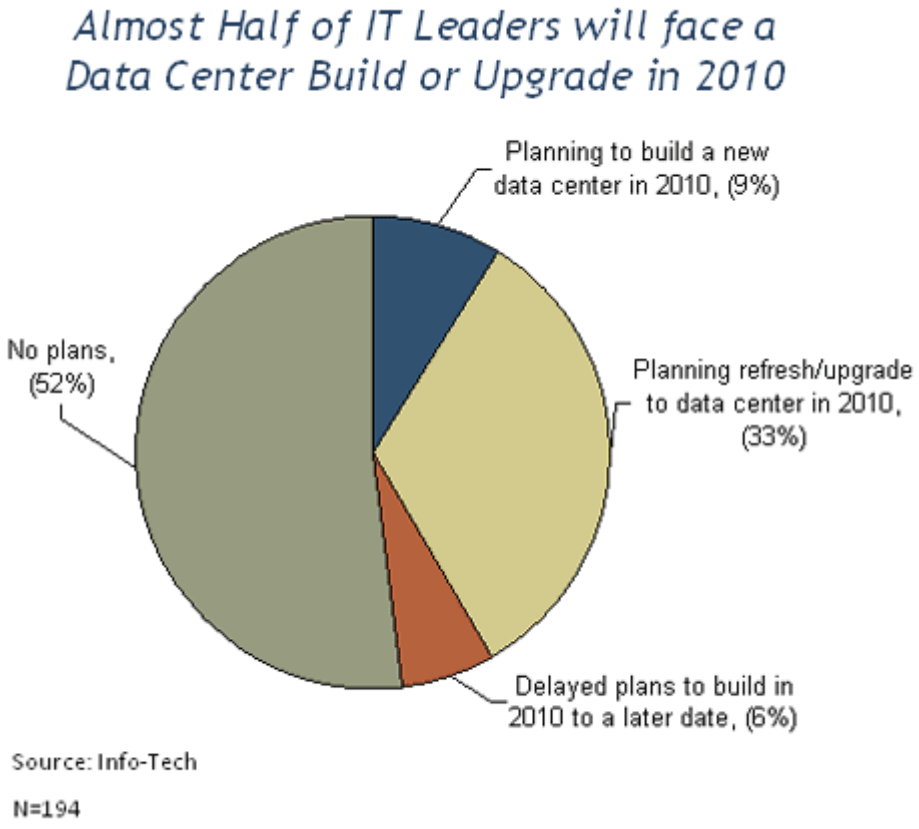
Data Center facilities require industrial design and engineering

– Data centers are getting bigger and more complex.
– Despite the availability of consolidation

technologies, organizations Info-Tech surveyed were two times more likely to increase the size of their data center than decrease it.

– Internal expertise is most often lacking to undertake the build.

– Data center facilities require specific industrial design and engineering requirements to meet the needs for fire-protection, power provisioning, stand-by power, cooling, physical security, and layout.

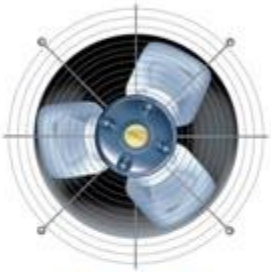


Careful analysis of five key design elements done internally now will yield better
cost and build outcomes later.

Analyze These 5 Now

Power

Stand by Power



Cooling

Fire-Protection

Site Selection

1982

To Yield Better Cost & Outcomes Later

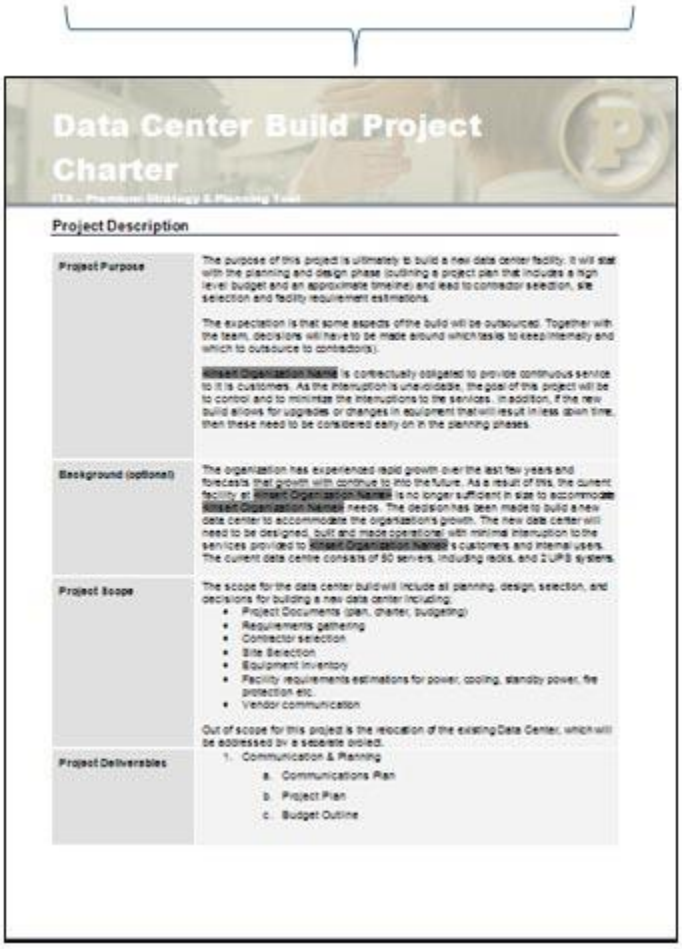
– Analyze these five factors before engaging outside consultants and contractors to correctly align the facility costs and capacity with enterprise needs.
– Data Center construction projects executed without these five design activities, before awarding contracts, demonstrate a higher likelihood of cost overruns, overcapacity, and problems.

Establish clear project scope to minimize expensive changes in later project stages.

Complete this project charter before the project commences to identify a clear and defined scope, decision rights, potential risks, and receive executive sponsorship.

The "**Data Center Build Project Charter Template**" will help you:

– Understand just how big a project you are building.
– Provide input into a build vs. buy decision.
– Provide a project approval for business leaders to sign off on.
– Reduce external costs of the project overall.



Data Center Build Project Charter

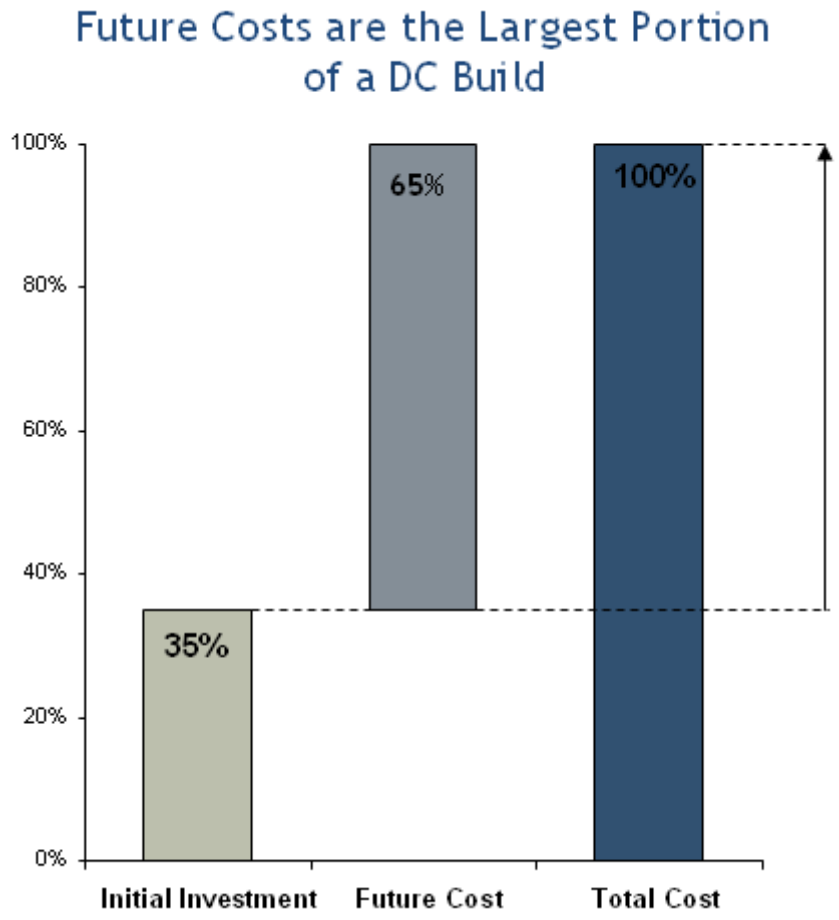
Project Description

Project Purpose	The purpose of this project is ultimately to build a new data center facility. It will start with the planning and design phase (outlining a project plan that includes a high level budget and an approximate timeline) and lead to contractor selection, site selection and facility requirement estimations. The expectation is that some aspects of the build will be outsourced. Together with the team, decisions will have to be made around which tasks to keep internally and which to outsource to contractor(s). [Insert Organization Name] is contractualy obligated to provide continuous service to it is customers. As the interruption is unavoidable, the goal of this project will be to control and to minimize the interuptions to the services. In addition, if the new build allows for upgrades or changes in equipment that will result in less down time, then these need to be considered early on in the planning phases.
Background (optional)	The organization has experienced rapid growth over the last few years and forecasts that growth with continue to into the future. As a result of this, the current facility at [Insert Organization Name] is no longer sufficient in size to accommodate [Insert Organization Name] needs. The decision has been made to build a new data center to accommodate the organization's growth. The new data center will need to be designed, built and made operational with minimal interruption to the services provided to [Insert Organization Name]'s customers and internal users. The current data centre consists of 50 servers, including racks, and 2 UPS systems.
Project Scope	The scope for the data center build will include all planning, design, selection, and decisions for building a new data center including: • Project Documents (plan, charter, budgeting) • Requirements gathering • Contractor selection • Site Selection • Equipment inventory • Facility requirements estimations for power, cooling, standby power, fire protection etc. • Vendor communication Out of scope for this project is the relocation of the existing Data Center, which will be addressed by a seperate project.
Project Deliverables	1. Communication & Planning a. Communications Plan b. Project Plan c. Budget Outline

65% of costs occur after the data center is built — get a handle on all costs before the build.

1985

- – On average, a 1000 square foot data center costs $1.6 M.
- – Each project is unique and should have its own detailed budget; create a detailed list of expected expenses for an accurate budget.
- – Consider ongoing costs in your initial budget calculations, as they make up the bulk of the total costs associated with building a data center.
- – Some major ongoing operating expenses include:
- – Building leases
- – Fire suppression systems
- – Mechanical systems for electrical distribution, and in-room cooling



Info-Tech research shows that data centers cost between $400 and $700 per square foot.

1987

Lower than industry conventions.

- – Using conventional industry standards can lead to over-estimating the build budget; Info-Tech's research shows actual costs are 22% less.
- – Organizations that rely on the industry standard of $600 to $900 per square foot when creating their budgets run the risk of over budgeting and over spending.
- – As a result, general benchmarks should not be used for a data center build.
- – Organizations should complete their own detailed investment and expenses to accurately budget for their builds.

For a more accurate and inclusive cost estimate
use Info-Tech's budgeting tool.

- – The ITA Premium, “Data Center Facility Budget Tool” enables the organization to account for all costs in order to arrive at an accurate estimate.
- – Accounts for hidden costs, such as janitorial expenses, landscaping, permits, and physical security.
- – Provides a complete example for a 1000 sq ft facility that you can tailor to suit your situation.
- – Factors all necessary renovations, repairs, and upgrades related to purchasing an existing data center into the final cost.
- – Even if you are planning on renovating an existing data center, you can still use the tool to determine ongoing operational costs.

Data Center Facility Budget Tool

Enter the organization's data into the WHITE cells below to estimate the facility budget for an 8 year period.
Be sure to enter in the square footage of the facility to calculate the total cost per square foot.

Enter Facility Size (in Square Feet)	1000
Total Cost per Square Foot	$585.93

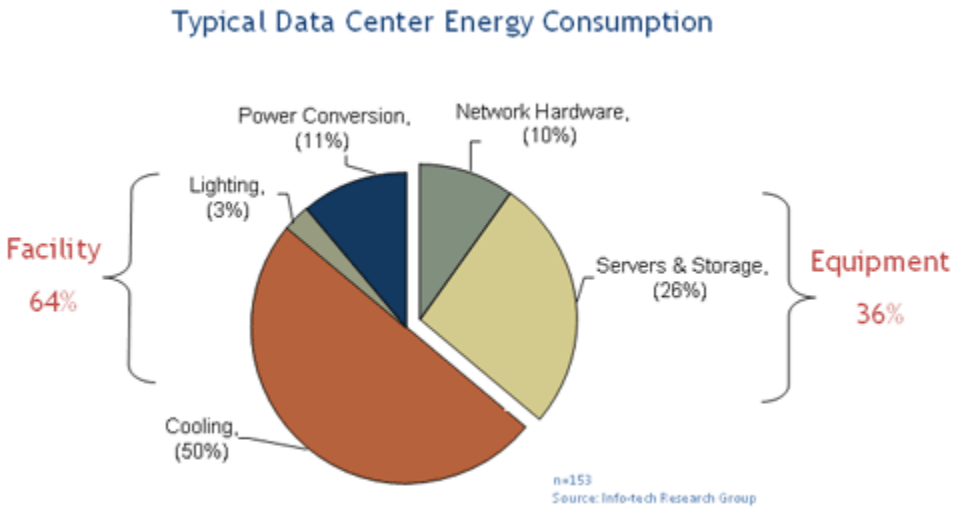
	Initial Investment (CapEx)	Year 1	Year 2	Year 3	Ong
Building Costs					
Land Costs	N/A	$ -	$ -	$ -	$
Building Lease	$12,230.00	$20,000.00	$20,000.00	$20,000.00	
Facilities site management fees	$ -	$3,600.00	$3,600.00	$3,600.00	
Janitorial and landscaping	$ -	$2,200.00	$2,200.00	$2,200.00	
Property Taxes	*N/A - Included in Lease costs*	$ -	$ -	$ -	$
Reimbursable expenses (leasehold improvement allowances, government incentives, etc.)	-$6,000.00	$ -	$ -	$ -	$
Sub-total Building Costs	$6,230.00	$25,800.00	$25,800.00	$25,800.00	
Site Infrastructure Costs					
Architectural and engineering fees	$27,000.00	$ -	$ -	$ -	$
Permits and Variances	$ 4,440.00	$ -	$ -	$ -	$
General Contractor fees	$ 30,200.00	$ -	$ -	$ -	$
General Construction costs	$ 67,510.00	$ -	$ -	$ -	$
Access Floor System	$ 6,210.00	$ -	$ -	$ -	$
Wall, Floor and Ceiling finishes	$ 6,570.00	$ -	$ -	$ -	$
Fire detection, alarms, and EPO	$ 8,123.00	$ 1,708.00	$ 1,708.00	$ 1,708.00	$
Fire suppression system - pre-action sprinkler, piping, inert gas installation, etc	$ 35,770.00	$ 10,015.60	$ 10,015.60	$ 10,015.60	$
Mechanical systems - non-critical electrical distribution(transformers, panels, lights, convenience plugs, etc.)	$ 22,800.00	$ -	$ -	$ -	$
Mechanical systems - critical electrical distribution (automatic transfer switch, maintenance bypass, PDU, etc.)	$ 66,760.00	$ 27,179.00	$ 27,179.00	$ 27,179.00	$
Stand-by power - Generators	$ 75,632.00	$ 21,000.00	$ 21,000.00	$ 21,000.00	$
Stand-by power - UPS	$ 52,310.00	$ 7,500.00	$ -	$ 14,000.00	$
Mechanical systems - heat rejection (Exchangers, Chillers, Towers, etc.)	$ 46,472.00	$ 6,506.00	$ 6,506.00	$ 6,506.00	$
Mechanical systems - in room cooling (CRAC, Hot aisle containment, rack containment, etc.)	$ 45,610.00	$ 12,770.00	$ 12,770.00	$ 12,770.00	$
Physical Security - Access control systems, monitoring, security guards, video surveillance	$ 12,730.00	$ 3,564.00	$ 3,564.00	$ 3,564.00	$
Sub-total Site Infrastructure Costs	$508,137.00	$90,242.60	$82,742.60	$96,742.60	
IT Infrastructure Costs					
Point of Presence / Carrier Access	$ 28,610.00	$ 19,200.00	$ 19,200.00	$ 19,200.00	$
Structured cabling - voice and data	$ 25,980.00	$ -	$ -	$ -	$
Racking costs	$ 10,650.00	$ -	$ -	$ -	$
Furniture and cabinets	$ 6,320.00	$ -	$ -	$ -	$
Sub-total IT Infrastructure Costs	$ 71,560.00	$ 19,200.00	$ 19,200.00	$ 19,200.00	$
Total Data Center Costs	$585,927.00	$135,242.60	$127,742.60	$141,742.60	$
Total Cost per sq. ft.	$585.93				

Power estimates should accommodate

1991

facility and equipment requirements.

Do not focus solely on the IT equipment housed in the data center.



Leverage the "Data Center Power and Cooling Requirements Calculator" to ensure budget accuracy.

- – Use the “Data Center Power and Cooling Requirements Calculator” to estimate the facility requirements for power, standby power, and cooling for your facility.
- – Accurate estimates of facility requirements are a critical input to the project’s budget.
- – Use this tool to estimate existing & future:
- – IT equipment power load requirements.
- – Cooling load requirements.
- – Standby power load requirements.
- – For more detailed guidance on requirements gathering, refer to the research note, “Data Center Facility Requirements Estimations At-a-Glance.”

Server Room Power and Load Requirements Estimates

Next Step >>>

The following table summarizes the power load requirements for the organization's data center environment based on the inputs gathered in Tabs 1-3.

Item	Description	Subtotal kW
Power Requirement - Electrical		
Existing or actual equipment load	Ratings for servers and other equipment (calculated in Tab 3 and 4).	**92.450**
Estimated future loads	Based on the average server load and anticipated number of future servers to be added.	**45.283**
Peak power draw from variation in critical loads	The total steady state load. Based on current and future loads.	**144.620**
UPS inefficiency and battery charging	Actual and future loads (kW).	**44.075**
Lighting	Based on the floor area in the server room.	**2.400**
Total power required to support electrical demands	Power drawn from servers, equipment, UPS, and lighting.	**191.094**
Power Requirement - Cooling		
Power to support chiller system	Based on the standard assumption that power required is 70% of total electrical demand.	**N/A**
Power to support DX system	Based on the standard assumption that power required is 100% of total electrical demand.	**191.094**
Total power required to support cooling demands	Power drawn from cooling devices.	**191.094**
Total Power Requirement		
Total amount of power required to support both electrical and cooling demands	Total amount of power required that includes current and future needs for power and cooling.	**382.188**
Size of Electrical Service Estimate		
Regulator (i.e.NEC) requirements	Total electrical and cooling requirements multiplied by 1.25 to meet regulator's requirements.	**477.735**
Three phase AC voltage	Existing AC voltage from building (can be found at service entrance).	**240.000**
Electrical service required from utility company (calculated in Amps)	Utility company estimate.	**1150.615**
Size of Standby Generator Estimate (if Applicable)		
Critical loads that require generator backup	Based on total power to support electrical demand.	**248.422**
Cooling loads that require generator backup	Based on total power to support cooling.	**286.641**
Estimated size of generator required		**535.064**

Total electrical service required from utility company
1150.615

Size of generator needed (estimated in kV)
535.064

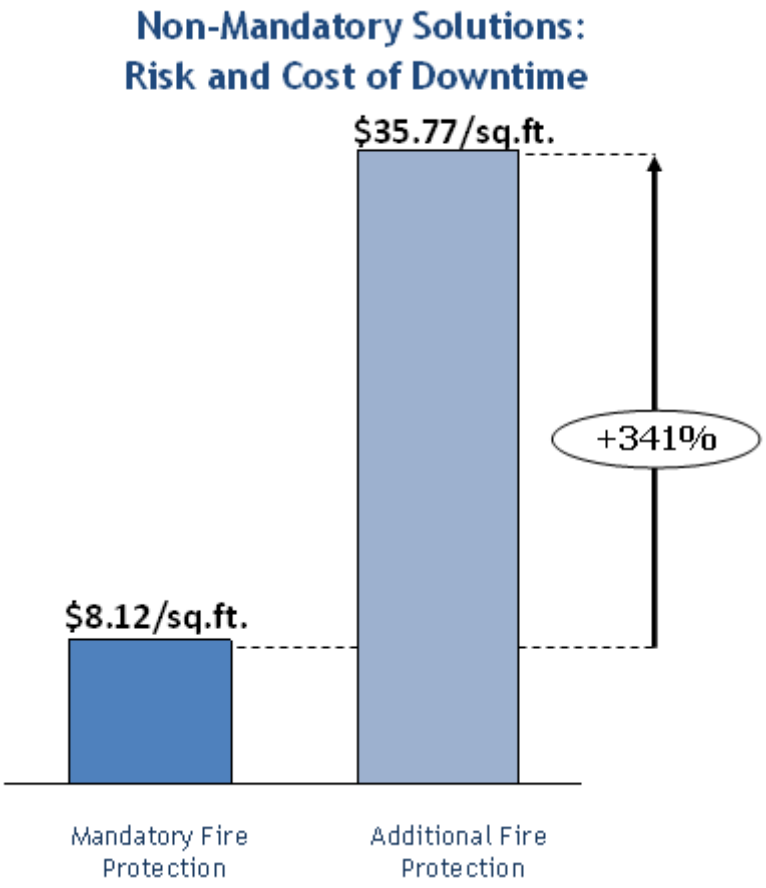
Electrical service required from utility company in Amps per sq. ft
0.959

Fire protection is not solely about IT assets;

1994

additional protection must be evaluated to avoid unnecessary costs.

- – The number one consideration when evaluating fire protection options is life safety in the data center. It is mandated by law and non-negotiable.
- – While clean agents can be very effective in suppressing fire and protecting property, they are also expensive to install and maintain and, may be toxic.
- – Refer to the ITA Premium research note, "**Fire Protection: Only YOU Can Prevent Data Center Fires**" for more detailed guidance.
- – Under NFPA 75 a data center must have a sprinkler system, detection and alarm systems, emergency power-off, and portable fire extinguishers.



Existing building sites often lead to additional construction costs. Site evaluation

1996

should identify any potential issues.

The selected site must be evaluated before the data center budget can be finalized.

- – Use the free "Data Center Site Selection Checklist" to target issues that will lead to cost overruns of safety/security concerns.
- – External environment, including physical hazards.
- – Internal checkpoints (power, cooling, etc.) in existing facilities.
- – Physical building attributes such as loading docks and water lines.
- – Security checkpoints for the protection of systems and staff.
- – Access control systems to prevent intruders.
- – RFacility layout must be planned before construction starts.
- – Data centers house more than just servers. Provide space for all equipment and to serve all functions.
- – Refer to the research note, "Planning for Data Center Physical Layout and Support Center Space" for more detailed guidance.

Data Center Site Selection Checklist

ITA – Premium: Selection Advice

Table 1. Site Details

Site Address:	
Location Type: (i.e. Urban, Suburban, Industrial complex)	
Is it in an existing building or a new building?	
Floor Number	
Leased or owned?	
Floor space (square feet):	

Table 2. Building Site Selection Checklist

Item #	Item	Yes	PAR	NO	N/A	Comments
1	Are there low risks from environmental threats such as:					
	Aircraft traffic					
1.1	Earthquakes?					
1.2	Floods?					
1.3	Hurricanes or tornados?					
1.4	Heavy ice or snow?					
1.5	Nearby hazardous industry?					
1.6	Does the location have a high crime rate?					
2	Are there secure utility service entries?					
3	Does the site have the ability to limit or restrict physical access to the data center?					
4	Is hard water a potential problem?					
5	Are multiple-grid commercial power services available?					
6	Are there data/voice communications 'stick' nearby and capable of supporting requirements?					
7	Is there public transportation nearby available?					

A common scorecard makes the contractor selection process transparent.

- – Each contractor should be evaluated based on a standard scoring method to ensure accuracy and fairness.
- – Use the "Data Center Contractor Scripted Interview Scorecard" to show senior management why a particular contractor is the best fit and ensure you account for the following important aspects:
- – Data center build experience of similar size and scale
- – Project management practices
- – Site selection processes
- – End user requirements processes
- – System execution and management practices
- – Support during and after the build

Data Center Contractor Scripted Interview Scorecard

ITA – Premium: Selection Advice

Contractor Company Name: _____

Contractor Interviewees Present: _____

Organization Interviewers Present: _____

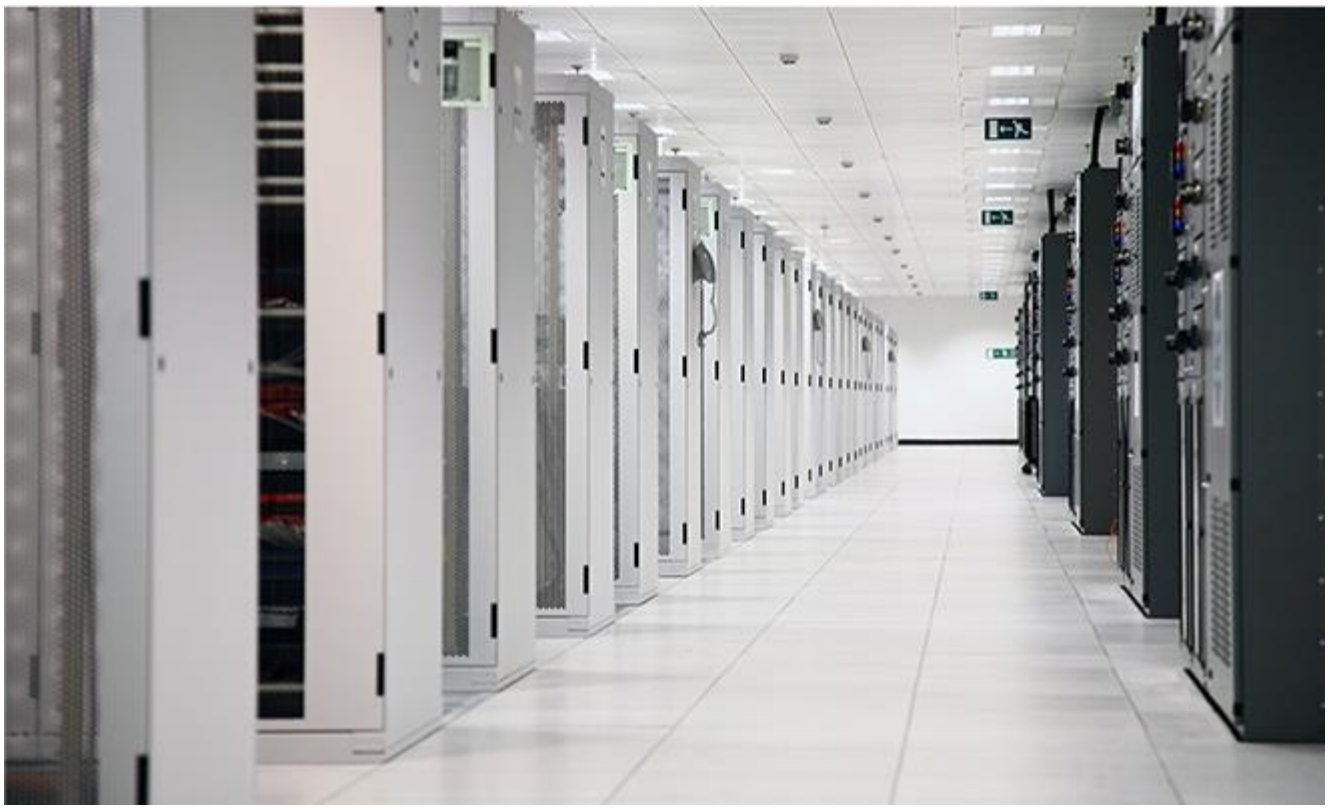
Interview Date: _____

Scorer's Name: _____

Contractor's Final Score: _____

Section #	Section Name	Weight	Key Components	Score (1 to 6)	Additional Comments
2.1	Introduction		Number of data center build projects and the types of businesses constructed for		
2.2	Service Overview		Key engagement practices Project management practices		
2.3.1	Site Selection		Site selection criteria Data Center end user		
2.3.2	Power Requirements		Critical loads for server and storage equipment. Loads of other equipment		
2.3.3	Cooling Requirements		Server and storage device heat output. Other IT equipment's heat output in the data center.		
2.3.4	Standby Power Requirements		Risk tolerance of the enterprise. Power requirements for critical servers and equipment.		
2.3.5	Fire Protection		Fire codes and standards Smoke and heat detection systems		
2.3.6	Physical Layout & Design		Security Accessibility		

How to Build a Company Data Center? Part One



2002

Have you decided to build a brand new data center to meet your company's needs? Then you have surely though about what type of a data center you're going to build, how much technology is going to be inside it and at what temperature will these machines operate. However before all that, an absolutely key factor is just choosing the right place. The companies that chose to put all their data in Karlin-Prague – a district of Prague devastated by 2002 flooding – would now surely agree.

A key question: Where?

There are two possibilities that hinge on your budget. The first one is to build a data center from the ground up, the second one is a so-called retrofit – this means that the data center will be placed into an already existing building. Both approaches have advantages and disadvantages. A completely new building allows you to build it so that it precisely fits the data center's needs, which in turn means you will be able to reach better energy parameters and higher levels of certification. The location should allow for future growth and should be in a seismically stable region outside flooding areas. The TIA 942 standard that contains proposals and recommendations for data center design talks about the ability to withstand even a 500-year flood.

It is generally best to choose a location with a high-quality access to both electrical energy and internet connectivity. Ideally, there should be two independent optical fibre tracks laid down already. When there's only one fibre connection available, laying down a second one is an item that will strain your budget and building times immensely. Building a fibre connection usually entails a long process of acquiring the land or the rights to use it and so on.

Going the retrofit way usually brings an added bonus of lower costs. On the other hand, it brings limitations imposed by the existing building and its condition. One thing you should be sure to check is the structural strength of floors and their load-bearing capacity. This usually amounts to around 200 kg per square meter in older buildings. However, a full server rack can hold three

or four times as much! And you shouldn't forget about the load-bearing capacity of your roof, as that will hold some of the air conditioning equipment and other components as well.

What is the Tier System?

The Tier System separates data centers into categories based on their availability. The higher the tier, the better the theoretical availability of the data center.

- Tier I basic This data center has both uninterruptible power supplies (UPS) and generators, but they don't have redundant backups. This can mean outages when the technicians are doing maintenance. They have guaranteed availability of 99,671 % (i.e. outages can amount up to 28 hours per year).
- Tier II redundant components Means the data center has N+1 component redundancy (UPS, routers, AC), but only one canal of distribution. Equals to guarantee availability of 99,741 % (i.e. up to 22 hours per year of outages).
- Tier III concurrently maintainable Any planned maintenance can be done without an outage. Such data center has enough capacity of backup lines through different routes, though a fault in infrastructure can still bring down a part of the data center. Guaranteed availability of 99,982 % means that the overall time of outages can be only 95 minutes per year.
- Tier IV fault tolerantThe infrastructure can withstand at least one of its parts failing, usually all of them are redundant (2x N+1). Guaranteed availability goes up to 99,995 %, so the total amount of outages per year will be less than 26 minutes.

Power is a key factor

Uninterruptible access to energy is the alpha and omega of any data center. That is why data centers are built and why they should be ideally powered by two

independent routes of energy coming through their own transformers. The energy goes through a substation through automatic transfer switch (APS) – with a diesel generator connected – and into the UPS units with their own battery modules. Smaller companies can make do with only one substation, but every machine should be dual powered through two independent routes. Larger data centers who expect to use more than 2 megawatts can meanwhile eschew the UPS with batteries and go for unique battery-less power solutions, the so-called dynamic or rotary UPS. These combine a transformer, a UPS and a motor-generator in one package. It works a bit like a flywheel in a vacuum case that is spinning endlessly and can convert this stored energy through a generator into electric energy when needed.

We should always place UPS units to a separate location, because they need a lower operational temperature. And you should not underestimate the importance of buying a high-quality UPS unit. Some of the cheaper ones have batteries with a limited life-span that have cost some companies a fortune in maintenance fees. And you won't go wrong with a modular UPS that allows you to just add more modules when you want to expand the data center.

The cooling is a matter of cold and hot aisles

A necessary part of every data center is its cooling. A de facto standard is a double flooring and a system of cold and hot aisles. The cold aisles should be roofed over to prevent the mixing of hot and cold air. The server racks need to be placed in a specific way – they should "inhale" the cold air from their front sides and "exhale" it hot out their backs (the so-called front-to-back system). The double flooring mentioned allows the cold air to be led into the correct place and also hides metallic cables. Optical fibres should probably be led up over the racks in special plastic troughs.

Today, the preferred method of cooling is the so-called free cooling. It allows data centers to use cold outside air which saves electric energy that would be otherwise spent cooling the air. This technique works best when the outside is cold but manages to save money on cooling even when the difference between inside and outside temperatures is only a few degrees. This can be helped with a

glycol-based circuits with an outside radiator that takes the heat out of the data center.

But a smaller company on a tight budget will make do with a normal direct expansion (DX) type of air conditioning. The DX units are cheaper to buy, but are more expensive in the long term. They're best used for cooling lower amounts of heat.

Most modern data centers today are built as so-called high-temperature data centers. The newest computer technology can stand temperatures from 27 to even 35 degrees Celsius in some extreme cases. Even a marginal increase in the operation temperature can save a lot of money spent on energy. For example, Google saves 4 % of all energy expenditures by raising their data centers' temperature by one degree Celsius and Intel manages to save even 7 % of their energy costs by doing the same thing. And if the conditions are right – as in the Czech Republic – it's possible to use free cooling throughout the year and save even more money.

Hints and tips for building your own data center

- Pick a spot outside a flooding area
- Look the building over and check the load-bearing capacities of both floors and roofs
- Check the availability of high-voltage power
- Keep the connectivity in mind, ideally there should already be 2 optical fibre routes ready
- When roofing over the aisles, be very meticulous and take care of even the smallest "leaks"
- Read through all the datasheets and focus on energy loss values and the energy conversion efficiency of all the machines

- Don’t forget to revise everything regularly and to hold safety courses for employees
- You should monitor and analyse all inputs and outputs, including the power circuits

The data needs to be safe from fire as well

Your servers need to be protected not only against energy outages, but also against fire and unauthorized access. Even the smallest server rooms should have at least a fire alarm and fire extinguishers for electric appliances. Medium-sized data centers usually employ automatic fire and heat detection systems and the fires are put out through inert gases, e.g. the ecological FM200 that is safe for both the servers and the environment. There are of course cheaper alternatives using for example demineralized water or sodium.

The physical side of security needs to be taken care of by surveillance cameras, coded doors and security workers. Don’t forget to monitor all the components. You should measure the energy consumption of every single rack to evaluate them and prevent unnecessary losses. Monitoring the temperatures and cooling is just as important. If the systems are forced to work outside their ideal operating range, their energy consumption usually rapidly grows. As this is the thing you want to avoid, it’s usually best to use systems and tools to monitor both these factors. There are many commercially available or you can just code your own solution.

But that’s still not all there is

What should you do when your data center is finally all built and equipped? It’s still not time to rest. To see the long road ahead of us, stay tuned for the next part of this series. The next article will tell you all about the virtualization of hardware and cloud building.

You may ask – “what is a build-to-suit data center? And should my company embark on this path, or lease capacity in an existing colocation data center?”

A build-to-suit data center is a custom facility that is designed and built specifically for your company. Designing and constructing a build-to-suit data center is complex and requires much collaboration, but the end result will be exactly what you need to match your requirements.

There are several reasons why your company might want to consider a build-to-suit facility:

1. **Customization**: There are hundreds of individual factors that determine the attributes of a data center – geography, size, design, resiliency just to name a few. Off the shelf existing facilities may not meet your business needs.

2. **Speed to Market**: It's very important for some companies to have online IT capacity ahead of skyrocketing business growth. A build-to-suit solution may deliver in a shorter timeframe than starting from scratch on your own.
3. **CapEx vs. OpEx**: In lieu of investing hundreds of millions of dollars on a capitalized spend (CAPEX), it makes sense to pass through the lease of a facility as an operational capitalized spend (OPEX)
4. **Expertise**: Unless designing and building data centers is your primary business, data center providers have a core competency of building data center facilities. It frequently makes sense to let the experts deliver the solution.

Advantages of a Build-To-Suit Data Center

A build-to-suit data center gives you the advantages of a private, dedicated data center, with far less

capital and operational expense. In building your facility, the data center provider takes on all the construction costs. Once it is complete, the provider owns the facility – but your company leases and occupies it as if it were your own private data center. In many cases, the data center provider will operate the facility for you, providing on-site security and maintenance teams.

Elements to Consider in Planning a Build-To-Suit Data Center

In designing a build-to-suit data center, you work with the data center provider to establish your requirements for the facility. It’s a complex

collaborative process, but it helps to ensure that the facility is built to serve your business needs.

Here is a list of basic questions you should consider when planning a build-to-suit data center.

Define the Purpose

Are you building this facility to host IT servers for mission-critical operations, backup data storage, public/hybrid cloud services, social media, eCommerce, high-security financial transactions, HPC and big data applications? Your overall purpose will influence all the other decisions you make, everything from IT power needs to security and operations requirements.

Location

Some companies require that a facility must be located within easy driving distance of the business location(s) that will be utilizing the servers housed there. Other companies don't mind having an out-of-state data center, as long as there is a nearby airport that allows personnel to fly into that location and check their servers when necessary. Also, some geographies offer significant sales and/or property tax incentives.

Physical Structure

Companies should look at the physical location and natural disaster risk that may occur in that location when analyzing their structure design. Building size will be driven by total critical power and design density of IT infrastructure.

Electrical Capacity

Determining the size and electrical capacity of your build-to-suit facility involves capacity planning. You should look ahead and estimate (1) the size of your initial IT footprint, (2) the number of IT servers you plan to add to this facility over the next 5-10 years, and (3) how much space, power, and cooling you will need to keep pace with the growth of your IT footprint in this facility.

A typical design density for today's data centers is 150 watts per square foot, meaning your facility will provide 150 watts of power for every square foot of space in your data hall. However, with increasing power demands for high-capacity IT infrastructure, many companies are asking for higher power densities, such as 225 watts per square foot.

The lower your density, the bigger the building footprint will be; the higher your density, the smaller the building footprint.

Cooling

The electrical capacity and facility location will drive the cooling design. The overall goal is to design an energy efficient facility with a low PUE (Power Usage Effectiveness).

Redundancy

Your electrical and mechanical redundancy strategy will depend on how much of your IT footprint you want to remain operational in the event of a power outage. A build-to-suit facility can be constructed with multiple redundancy features, including dual power feeds from separate utilities, dual UPS systems, backup power generators, backup cooling units, etc. The

redundancy of a facility helps with the facility “Uptime” (i.e. servers never fail). The data center industry defines the “Uptime” in Tiers – Tier I, II, III, IV. Tier I = 99.671% Uptime; Tier II = 99.749% Uptime; Tier III = 99.982% Uptime; Tier IV = 99.995% Uptime.

Networking

Define the overall fiber connectivity strategy. Analyze which carriers are in the specific location. Do you need cloud on-ramp access?

Life Safety

Analyze life safety requirements (i.e. fire suppression, etc.)

Security

Physical security is a critical component of data centers. Determine how many concentric rings of

security are required. This means, how many physical barriers (i.e. card readers, etc.) are required prior to entering the data hall.

Certifications

Determine what certifications are required for this facility, including government compliance certifications (i.e. HIPAA, Sarbanes-Oxley and PCI).

Financial

Establish what the overall financial strategy is for the business. What can the business afford to spend to maintain facility Uptime? What are the factors that decide the ongoing operational budget for the data center?

Why T5 should be your Build-to-Suit Partner

In planning and constructing a build-to-suit facility, T5 Data Centers takes a collaborative approach to working with our customers. Our Data Center Construction Group meets with representatives from various departments (i.e. architectural, networking, security) within the clients' organization. Using the questions listed above, we help our clients define the features and elements that they will require in their facility.

Through a series of planning meetings with our clients, our designers and engineers develop a floor plan and a "Basis of Design" document, which describes the architectural and mechanical features of the build-to-suit facility. From there, we develop a more detailed "Design Development" package, which we review and revise several times with the client's

input. Once the designs are refined to the point where the construction team can obtain a permit from local jurisdiction, we begin building the facility.

During the construction phase, our engineers have weekly collaborative meetings with the client's representatives and monthly meetings with client executives at the data center construction site, to talk about the facility's progress. We also work with the client through the commissioning process, and through the transition to set up operations crew and security staff.

Chapter 14:

What Kind of Equipments You Need to Build Your Big Data Center?

2019

5 Types of Equipment Your Data Center Must Have

Running a data center requires a lot of planning to ensure everything runs properly today and long into the future. Whether you are building a small data center for your company's computer equipment, or you manage a cloud data center with hundreds of thousands of square feet of floor space, there are certain items that you absolutely must have. Here are five categories of data center equipment with several examples of items to help get you going.

Environmental Control Equipment

When running a lot of computer equipment in one place, you need to ensure things don't get too warm. In addition, this type of equipment will run optimally when kept at very precise humidity levels. It isn't enough to simply keep the facility air-conditioned; you need to be able to manage the environment as a whole using items such as:

- **Environmental Monitors** – Placing sensors throughout the facility to take temperature and humidity readings is critical for monitoring the environment and allowing your system to react.
- **Airflow Plans** – If you don't plan how the air will flow through your facility, you will end up with 'hot spots' of drastically different temperatures, which could end up damaging equipment. There are many ways to manage airflow using fansand fan trays, air pressure, and much more.
- **Humidifiers** – Many people assume that any humidity in a data center is bad, but that's not actually true. For most equipment, keeping the humidity level between 45% and 55% is ideal. This can be a challenge when using air conditioning units, which is precisely why humidifiers are necessary.
- **Air Cooling** – Of course, powerful air-cooling units are a must. These units can be traditional air conditioners, water coolers, or any number of other options. The important thing is ensuring the air-cooling units can keep the data center at the proper temperature at all times.

Server Racks

All data centers house various types of computer equipment including servers, routers, switches, telephone equipment, and much more. It's not enough to simply place these items on a desk or shelf to operate. This is not only an inefficient use of space, but can also cause problems with cooling, can expose them to physical hazards, and result in reduced

security.

Server racks are the ideal place to house computer equipment in any data center. In fact, server racks are the mental image most people have when they think about data centers. Most people have seen pictures of large data centers with several rows of server racks containing equipment and blinking lights. This is an accurate image of how data centers are able to safely house millions of dollars worth of computer equipment and allow it all to operate efficiently.

Redundant Power Sources

Equipment in a data center relies heavily on electricity. This starts with the servers, routers, switches, and other IT equipment, but it goes well beyond that. Air handling and cooling systems, security cameras, and just about everything else in a facility require electricity to run. Not only does it require a large amount of power, but that power must be very reliable. If there is even a brief interruption of power, all the equipment goes down. This then requires many hours of support to bring everything back online. In addition, power interruptions and spikes can damage the equipment, resulting in extensive downtime and costly repairs.

Most data centers today will have a two-layer system for redundant power. First, there is a large battery backup system (often called ‘uninterrupted power supply’). All the power flows through this battery backup so if the commercial power drops for even a fraction of a second, the batteries will continue to supply everything in the facility with a source of electricity.

These batteries are typically sufficient to power a data center for only a short time (5-20 minutes in most cases). When the commercial power is cut and the batteries begin supplying the facility with power, large generators will engage and begin to supply the power that is needed. In most facilities, this is in the form of large, diesel-powered generators. These generators can supply consistent power for as long as necessary, providing that the diesel fuel tanks can be resupplied.

Cable Management Supplies

Anyone who has worked in a data center without a well-thought-out cable management strategy knows what a mess cables make. Before adding any equipment to a data center, it is critical to have cable management equipment in place. This starts with simple items like zip ties that will hold cables in place. In addition, buying colored cables will help make it easier to identify different types of cables at a glance.

Within server racks, it is necessary to have proper cable management kits that will keep the cables properly organized. Guiding the cables either up

to the top of the server rack, or down through the bottom, will make it easier to test and replace cables if they go bad, and also keeps them safer so there is a lower risk of an outage. A well-executed cable management plan in a data center is like a work of art and also dramatically improves the way the entire facility operates.

Computer & Network Equipment

The purpose of data centers is to house computer and network equipment, so it may seem odd to list this last. The fact is, however, that if you have the luxury of planning out a data center from the start, the computer and network equipment are the last items that need consideration. If you properly plan and install all the above-mentioned things, you will be able to add virtually all types of computers, switches, routers, servers, telco equipment, mainframes, and more without any trouble.

This type of equipment is also replaced more regularly according to the lifecycle management of each item. While air handling and cooling equipment, server racks, and other items can last for decades without a problem, the actual IT components typically get replaced every 3-5 years.

Whether you are helping to plan for a new facility, or you're just looking to ensure an existing facility is set up correctly, it is important to think about every piece of data center equipment that is used. Data center management requires extensive planning to get everything right, but all this planning will be well worth the effort!

What Does Data Center Equipment Cost?

Looking for that shopping list to see if you have the budget? You'll never know what data center equipment really costs you until you acquire it and start using it. However, a few rules of

thumb can help to put things in context and understand how different elements drive expenses.

Globally, costs are determined in two ways.

1. The first, naturally enough, is the total amount of data center equipment, which is related to the size of the data center.
2. The second is the level of availability required of the data center. A Tier I data center with basic non-redundant infrastructure will be less expensive to equip than a Tier IV data center with full fault tolerance.

Typical Equipment Required

We can split equipment into two categories. The first is the end-user IT equipment, including servers, storage units, and network equipment. End-user enterprises may choose to organize this is different ways, according to the performance and availability they require. For instance, “scale up” to build a bigger server or “scale out” to add additional servers are two typical choices.

The second is “all the rest,” meaning the additional equipment required to allow the end-user IT equipment to function usefully and reliably. The notions of “scale up” and “scale out” apply less here, even though it is possible to retrofit an existing data center to take it to a higher tier of availability. Basic items are:

- **Electrical systems** – including main switchgear, power distribution units, intermediary distribution, and data center lighting
- **Critical support equipment**, -- such as diesel backup generators and UPS equipment

- **Internal fittings** -- including the access floor and ceiling systems
- **Mechanical systems** -- including HVAC
- **Fire protection system**
- **Plumbing system**
- **Monitoring and security systems** -- facility management, CCTV, and building access system
- **Core networking installation** -- shared between different end-user IT installations

Other non-negligible items are the construction costs of the data center, including the cost of the real estate (highly variable), building, structural steel, walls, and floors. We simplify things by leaving them out of the discussion here, but somebody still has to pay for them!

Sample Price Tags

Let's start with a modest data center size for housing 10 cabinets (standard 19 inch x 73.5 inch/42U) of end-user IT equipment. Estimates will always vary, but a web simulator based on previous data center project costs might indicate prices for a Tier I data center on the order of:

- $222,000 for power equipment (electrical, backup, lighting)
- $48,000 for HVAC, plumbing, floor and ceiling systems, and fire protection
- $14,000 for monitoring, security and core networking.

If you wanted to equip a Tier IV data center instead, each price would almost double. (Add an extra 80% to 90%.)

Data Center Size and Crossover Points

As the number of cabinets increases, so, of course, do these equipment costs. What is of interest to many companies is the relationship between a DIY approach and a data center rental or colocation solution.

At lower levels, a colocation solution is more economical. The crossover point where it becomes financially feasible to think about building your own data center might be around the 75 cabinet level for a Tier I facility, and 120 cabinets for a Tier IV facility.

However, this analysis only takes into account the data center equipment costs described here.

Additional factors are costs of ownership (maintenance, repair, upgrades) and costs of qualified personnel to run the facility. You can look forward to hours of fun with that financial spreadsheet model before you get a definitive answer!

Data center hardware is changing

You might expect the physical design of data center hardware to be pretty standard by now.

Don’t be so sure…

Servers have been installed in 19in racks since before data centers existed. The technology to cool the air in buildings has been developed to a high standard, and electrical power distribution is a very mature technology, which has made only incremental changes in the last few years.

Given all that, you might be forgiven for thinking that the design of data center hardware is approaching a standard, and future changes will be mere tweaks. However you would be wrong. There are plenty of radical approaches to racks, cooling and power distribution around. Some have been proclaimed for years, others have come from out of the blue.

Not all of them will gain traction.

This feature appeared in our special colo design supplement.
Subscribe for free today.

Rack revolution

– OVH

For someone used to racks in rows, a visit to one of French provider OVH’s cloud data centers is a disorienting experience. OVH wants to be the major European cloud provider - it combines VMware and OpenStack based platform as a service (PaaS) public clouds and hosts corporate private clouds for customers - but goes against the standard fare of the industry. Instead of standing vertically, OVH’s racks lie horizontally.

Close to its French facilities in Roubaix, it has a small factory, which makes its own rack frames. These “Hori-Racks” are the same size as normal 48U racks, but configured completely differently. Inside them, three small 16U racks sit side-by-side. The factory pre-populates these “Hori-Racks” with servers before they are shipped to OVH facilities, mostly in France, but some in farther corners of the world.

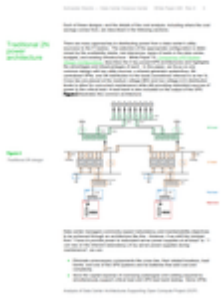
The reason for the horizontal approach seems to be speed of manufacturing and ease of distribution: “It’s quicker to deploy them in the data centers with a forklift, and stack them” said OVH chief industrial officer François Stérin.

Racks are built and tested speedily, with a just-in-time approach that minimizes inventory. Three staff can work side by side to load and test the hardware, and then a forklift, truck or trailer (or ship) can move the rack to its destination, in Gravelines, Strasbourg or Singapore. In the building, up to three racks are stacked on top of each other, giving the same density of servers as conventional racking.

OVH can indulge its passion for exotic hardware because it sells services at the PaaS level - it doesn’t colocate customers’ hardware. Also, it takes a novel approach to cooling (see later) which allows it flexibility with the building itself.

You don’t have to go as far as OVH in changing racks: there are plenty of other people suggesting new ways to build them. The most obvious example is the open source hardware groups like Open Compute Project (OCP), started by Facebook, and Open19, started by LinkedIn.

Both operate as “buyers’ clubs,” sharing custom designs for hardware so that multiple customers can get the benefits of a big order for those tweaks - generally aimed at simplifying the kit, and reducing the amount of wasted material and energy in the end product. Traditional racks, and IT hardware, it turns out, include a lot of wasted material, from unnecessary power equipment down to manufacturers’ brand labels.



Analysis of Data Center Architectures Supporting Open Compute Project (OCP)

Open Compute has had a significant impact on the thinking about data center design. Until now, the focus has been on systems at the rack level, leaving unanswered questions about the power infrastructure upstream of the rack. In this paper, Schneider Electric address critical questions about the implications of Open Compute on the upstream power infrastructure, including redundancy, availability, and flexibility.
07 Jun 2019

OCP was launched by Facebook in 2011 to develop and share standardized OEM designs for racks and other hardware. The group’s raison d’etre was that webscale companies could demand their own customized hardware designs from suppliers, because of their large scale. By sharing these designs more widely it would be possible to spread the benefits to smaller players - while gathering suggestions for improved designs from them.

While the OCP's founders all addressed giant webscale players, there are signs that the ideas have spread further, into colocation spaces. Here, the provider doesn't have ultimate control of the hardware in the space, so it can't deliver the monolithic data center architectures envisaged by OCP, but some customers are picking up on the idea, and the OCP has issued facility guidelines on making a place "OCP ready," meaning that OCP racks and OCP hardware are welcome and supported there.

OCP put forward a new design of racks, which packs more hardware into the same space as a conventional rack. By using more of the space within the rack, it allows equipment that is 21in wide, instead of the usual 19in. It also allows deeper kit, with an OpenU measuring 48mm, compared to a regular rack unit of 44.5mm.

The design also uses DC power, distributed via a bus bar on the back of the rack. This approach appeals to monolithic webscale users such as Facebook, as it allows the data center to do away with the multiple power supplies within the IT kit. Instead of distributing AC power, and rectifying it to DC in each individual box, it's done in one place.

Open Rack version 1 used 12V power, and a 48V power was also allowed in version 2, which also added the option for lithium-ion batteries within the racks, as a kind of distributed UPS system.

That was too radical for some, and in 2016 LinkedIn launched the Open19 group, which proposed mass market simplifications without breaking the 19in paradigm. Open19 racks are divided into cages with a simplified power distribution system, similar to the proprietary blade servers offered

by hardware vendors. The foundation also shared specifications for network switches developed by LinkedIn.

04 Jan 2018

Opening the racks

Rack standards are getting an upgrade, but will Open Rack or Open19 come out on top?

“We looked at the 21 inch Open Rack, and we said no, this needs to fit any 19 inch rack,” said Open19 founder Yuval Bachar,launching the group at DCD Webscale in 2016. ”We wanted to create a cost reduction of 50 percent in common elements like the PDU, power and the rack itself. We actually achieved 65 percent.”

Just as it launched Open19, LinkedIn was bought by Microsoft, which was a major backer of OCP, and a big user of OCP-standard equipment in data centers for its Azure cloud. Microsoft offered webscale technology to OCP, such as in-rack lithium-ion batteries to provide local power continuity for the IT kit, potentially replacing the facility UPS.

While the LinkedIn purchase went through, OCP and Open19 have continued in parallel, with OCP catering for giant data centers, and Open19

aiming at smaller facilities used by smaller companies that nevertheless, like LinkedIn, are running their own data centers. Latterly, Open19 has been focusing on Edge deployments.

In July 2019, however, LinkedIn announced it no longer planned to run its own data centers, and would move all its workloads over to the public cloud - obviously enough using its parent Microsoft's Azure cloud.

Also this year, LinkedIn announced that its Open19 technology specifications will be contributed to OCP. It's possible that OCP and Open19 specifications may merge in future, but too early to say. Even if LinkedIn no longer needs it, the group has more than 25 other members.

For webscale facilities, OCP is pushing ahead with a third version of the OCP Rack, backed by Microsoft and Facebook, and it seems to be driven by increasing power density demanded by AI and machine learning.

"At the component level, we are seeing power densities of a variety of processors and networking chips that will be beyond the ability of air to cool in the near future," said the Facebook blog announcing OCP Rack v3. "At the system level, AI hardware solutions will continue to drive higher power densities."

The new version aims to standardize manifolds for circulating liquid coolant within the racks, and also heat exchangers for cabinet doors, and includes the option of fully immersed systems. It's not clear what the detailed specifications are but they will emerge from OCP's Rack and Power project, and its Advanced Cooling Solutions sub-project.

– OVH

Liquid cooling

For the last couple of decades, liquid cooling has shown huge promise. Liquid has a much higher ability to capture and remove heat than air, but moving liquid through the hardware in a rack is a big change to existing practices. So liquid has remained resolutely on the list of exotic technologies which aren't actually worth the extra cost and headache.

Below 20kW per rack, air does the job cost effectively, and there is no need to bring liquid into the rack. Power densities are still generally well below that figure, so most data centers can easily be built without liquid cooling. However, there are two possibilities pushing liquid cooling to the fore.

Firstly GPUs and other specialized hardware for AI and the like could drive the power density upwards.

Secondly, for those who implement liquid cooling there are other benefits. Once it's implemented, liquid cooling opens up a lot of flexibility for the facility. Air-cooled racks are part of a system which necessarily includes

air conditioning, air handling and containment - up to and including the walls and floors of the whole building.

Liquid cooled racks just need an umbilical connection, and can be stood separately, on a cement floor, in a carpeted space, or in a small cabinet.

This can be difficult to present in a retail colocation space, because it affects the IT equipment. So unless the end customer specifically needs liquid cooling, it’s not going to happen there. But it does play well to the increasing flexibility of data centers, where the provider has control of the hardware, and there isn’t a building-level containment system.

Small Edge facilities are often micro data centers without the resources of a full data center behind them. Other data centers are being built inside repurposed buildings, often in small increments. A liquid cooling system can fit these requirements well.

Early mainframes were water-cooled but in the modern era, liquid cooling has emerged from various sources.

21 Jun 2018

Getting into hot water

Fears over climate change and rising power densities led supercomputing designers to adopt hot water cooling

Some firms, including Asperitas, Submer and GRC; completely immerse the racks in a tub of inert liquid. No energy is required for cooling, but maintenance is complicated because the rack design is completely changed, and servers and switches must be lifted from the tub and drained before any hardware modifications. Iceotope, which now has backing from Schneider, has a system which immerses components in trays within the racks.

Others offer direct circulation, piping liquid through the heatsinks of energy-hungry components. This was pioneered by gamers who wanted to overclock their computers, and developed rigs to remove the extra heat produced. Firms like CoolIT developed circulation systems for business equipment in racks, but they have been niche products aimed particularly at supercomputers. They require changes to the racks, and introduce a circulatory system, which flows cool water into the racks and warm water out.

In Northern France, OVH has its own take on liquid cooling, as well as the racks we saw earlier. It builds data centers in repurposed factories which formerly made tapestry, soft drinks and medical supplies, and liquid cooling allows it to treat these facilities as shells: building in one go, with a raised floor, and building-level air conditioning OVH stacks in racks as needed, connecting them to a liquid cooling system.

“Our model is that we buy existing buildings and retrofit them to use our technology,” OVH chief industrial officer François Stérin explained. “We can do that because we make our own racks with a water cooling system that is quite autonomous - and we also use a heat exchanger door on the back of the racks. This makes our racks pretty agnostic to the rest of the building.”

The flexibility helps with changing markets, said Stérin: “To test the market we don't need to build a giant 100MW mega data center, we can start with like a 1MW data center and see how the market works for us.”

With a lot of identical racks, and a factory at its disposal, OVH has pushed the technology forward, showing *DCD* multiple versions of the concept. OVH technicians demonstrated a maintenance procedure on the current version of the cooling technology.

19 Aug 2019

Future Proofing Colo Design Supplement

If you thought data design was done and dusted, think again. There are innovations that could change the way we build and run facilities, right down to the racks where the IT lives. This supplement will help you design for the future by keeping one step ahead of developments

It seemed a little like surgery. First the tubes carrying the cooling fluid were sealed with surgical clamps, then the board was disconnected from the pipes and removed. Then a hard drive was replaced with an SSD.

Already this design is being superseded by another one, which uses bayonet joints, so the board can be unplugged without the need to clamp the tubes.

Less radical liquid cooling systems are available, including heat exchangers in the cabinet doors, which can be effective at levels where air cooling would still be a viable option.

OVH combines this with its circulation system. The direct liquid cooling removes 70 percent of the heat from its IT equipment, but the other 30 percent still has to go somewhere - and is removed by rear-door heat exchangers. This is a closed loop system, with the heat rejected externally.

Liquid cooling isn’t necessary for a system designed to be installed in a shell. It’s now quite common to see data centers where a contained row has been constructed on its own, on a cement floor, connected to a conventional cooling system. Mainstream vendors such as Vertiv offer modular builds that can be placed on cement floors, while others offer their own take.

One interesting vendor is Giga Data Centers, which claims its WindChill enclosure can achieve a PUE of 1.15, even when it is being implemented in small increments within a shell, such as the Mooresville, North Carolina building where it recently opened a facility.

The approach here is to build an air cooling system alongside the rack, so a large volume of air can be drawn in and circulated.

Don’t be lulled into a false sense of security. The design of hardware is changing just as fast as ever, and those building colocation data centers need to keep track of developments.

Data Center Cooling Solutions

Cooling for Today's Data Center

As the scale and performance demands for today's data centers continue to increase, so does the need for superior data center cooling. Computing creates heat. Data centers, which are filled with servers, generate lots of heat. Excessive heat can cause hot spots that lead to equipment failure and downtime. To remove this heat and avoid downtime, IT teams use a variety of data center cooling methods. It's like air conditioners for computer rooms.

The Path to Cost-efficient Data Center Cooling

Data center cooling is eating up IT budgets. Around 40% (1) of a data center's energy consumption comes from cooling. Fortunately, a new generation of data center cooling technology is keeping energy costs down while enhancing reliability.

There are numerous ways to cool a data center. In fact, our white paper compares 13 methods, and can help you find the best option for cooling your data center. One method is to use precision, in-row cooling units that are placed very close to the source of heat. Recent innovations make it possible to reduce energy consumption to create this cold air. Air economizers bring in cold air from outdoors and send it through HVAC

systems to the computer room. Another promising approach is to use cold water from the ocean. Methods like this are helping data center professionals achieve cooling solutions that meet their IT objectives. Learn how we can help you achieve a healthy data center environment with our flexible, integrated cooling solutions.

(1) Reference: *Science Direct: "Data Center Energy and Cost Saving Evaluation"*

- Data Center Cooling Meets NatureGreen Mountain customer story

Green Mountain, a Norwegian data center company, uses frigid fjord water to cool its data centers year-round. This approach, combined with an intelligent DCIM system, allows the company to cut its energy needs by a third. Plus, Green Mountain's data center produces zero carbon emissions.

Featured Data Center Cooling products

- Netshelter Data Center Containment

- Rack Air Distribution

- Room Air Distribution

-
InRow Direct Expansion

- **Uniflair** Ceiling Mounted Split Room Cooling

Data center Building & General Specification

1. 1. Building & General Specification Data Centers Summarization : Ali Mirfallah 1
2. 2. Contents 1. “ Data Center Design Philosophy ” 2. “ Data Center Design Criteria ” 3. “ Designing a Data Center ” 4. “ Determining Data Center Capacities ” 5. “ Site Selection ” 6. “ Implementing a Raised Floor ” 7. “ Power Distribution ” 8. “ HVAC and Other Environmental Controls ” 9. “ Network Cabling Infrastructure ” 10. “ Shipping, Receiving, and Staging ” 11. “ Avoiding Hazards ” 12. “ Environmental Contaminants ” 2
3. 3. What is a data center? 3 A data center is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and various security devices. Large data centers are industrial scale operations using as much electricity as a small town.
4. 4. sanity 4υ modularity υ scalability υ flexibility υ Simplicity υ1.Data Center Design Philosophy five core values that are the foundation of a data center design philosophy:
5. 5. Simplicity A simple data center design is easier to understand and manage. A basic design makes it simple to do the best work and more difficult to do sloppy work 5
6. 6. Flexibility Nobody knows where technology will be in 3 years, but it is a good guess that there will be some major changes. Making sure that the design is flexible and easily upgradable is critical to a successful long-term design. 6
7. 7. Scalability The design should work equally well for a 2,000, 20,000, or 2,000,000 square foot data center 7
8. 8. Modularity Data centers are highly complex things, and complex things can quickly become unmanageable. Modular

design allows you to create highly complex systems from smaller, more manageable building blocks. 8

9. 9. Sanity Designing and building a data center can be very stressful. There are many things that can, and will, go wrong. Keep your sense of humor. Find ways to enjoy what you're doing 9
10. 10. " Determining the Viability of the Project " 10υ " Insurance and Local Building Codes " υ " System Availability Profiles " υ " Scope , Budget and Criteria " υ2.Data Center Design Criteria
11. 11. Secondary criteria 11υEssential criteria υLocation (or site) υ Criteria : The most important criteria for a data center can be put into the following categories: υ Budget : Designing a data center isn't just about what the company needs or wants, it's what they're willing to pay for something's υ Scope: Most often, it is the project scope that determines the data center design. The scope must be determined based on the company's data center needs (the desired or required capacities of the system and network infrastructure), as well as the amount of money available. υScope , Budget and Criteria
12. 12. Location (or site) It's true that where you choose to locate the data center site (region/building) is important, but this choice is based on many different factors. connectivity through the company's network infrastructure has made it possible for a data center to be located wherever it is practical and affordable 12
13. 13. Bandwidth: Without connectivity, the data center is of little value. The type and amount of bandwidth is device dependent. You must have physical capacity, power, and cooling to even consider connectivity. 13υ Cooling: Without cooling nothing will run for long. This is either on or off, though redundancy increases uptime. You must have physical capacity and power to run HVACs. υ Power:

Without power nothing can run . Power is either on or off. Connections to different parts of the grid and/or utilizing a UPS increases uptime. You must have physical capacity to have room for power and the equipment that needs power. υ Physical capacity: You must have space and weight capacity for equipment, and therefore, the other three criteria. There must be space for the equipment and the floor must be able to support the weight. This is a constant. υEssential criteria All data centers must have the following four elements in whatever capacities are needed or available :

14. 14. A Command Center 14υ Equipment such as forklifts and pallet jacks υ equipment cabinets, etc. υ All of the miscellaneous hardware, security cameras, card readers, door knobs, υ Walls, doors, windows, offices, loading dock υ Fixtures such as plumbing and lighting υSecondary criteria The level of importance of secondary criteria is wholly dependent on the company and project scope. It's conceivable that the budget could be trimmed Examples of secondary criteria are:
15. 15. Total budget 15 Essential criteria Secondary criteria Location Budget
16. 16. Network Redundancies : The amount of network equipment that must be available in the event of failure. The number of connections to your ISP. The number of network feeds needed to multiple ISPs in the event that one has a catastrophic failure. 16υ Cooling Redundancies : The number of extra HVAC units that must be available in the event that one or more units fail. υ Power Redundancies : The number of feeds from different parts of the grid, the number of UPS systems, etc., that must be installed to make sure these systems stay running. υ Device Redundancies :The number of backup devices that must be available in the event of equipment failures. υSystem Availability Profiles

17. 17. Insurance and Local Building Codes Insurance and local building codes will have an effect on many design decisions and should be considered in every aspect of the design process by the entire design team, including all building contractors. 17
18. 18. Inadequate district or too remoteυ Overly expensive location υ Inadequate pool of qualified employees υ Better business decision to use co-location or ISP, if only temporarily υ Retrofit problems such as grounding, cable routing, inadequate floor to ceiling height, no way to set up seismic restraints, etc. υ Inadequate budget υ High history of fires υ Too many weather or seismic problems υ Local building codes, insurance, or fire regulations are too restrictive υ Inadequate ISP service υ Inadequate cooling capacity υ Inadequate power. υDetermining the Viability of the Project There are times when too many compromises must be made to make the data center project viable. It might be something obvious (you can't get enough power from the local power company or there are frequent flooding problems), or it might be a number of small factors that, when looked at collectively, show that the project is a bad risk. Consider the following possible constraints on the project: 18
19. 19. "Planning for Possible Expansion" 19υ "Remote Systems Management" υ "System Monitoring" υ "Physical and Logical Security" υ "Data Center Support Systems" υ "Data Center Structural Layout" υ "Design Process" υ3.Designing a Data Center
20. 20. Design Process The design stages for the data center usually take the skills of Architects, Accountants, Structural, Mechanical, Electrical, HVAC, System, and Network engineers, Project managers, and Procurement personnel. Add also the probability of sales personnel, insurance carriers, and risk management analysts

Overseeing the project is a data center design engineer whose task is to accommodate the requirements of the system and network engineers, and to work with the other members of the team to ensure that the data center requirements (based on the project scope) are met. 20 Design Drawings Designing for Data Center Capacities

21. 21. HVAC design firms :They determine HVAC unit placement and whether they should be 20-ton or 30-ton, determine proper installation of piping that brings chilled fluids to units, and where cooling towers, compressors, and heat exchangers will be located. 21υ Electrical design firms and engineers : They deal with lighting plans, electrical distribution, wire ways under the floor, breaker subpanels, power transformers , wiring for the fire detection system, and smoke alarms. υ Structural engineers :They make sure the building will use materials and construction techniques that will keep the roof from collapsing under the weight of all those cooling towers. υ Interior designers : They create the "look" of the place, sometimes matching company specifications for consistency of styles, from trim to carpet. υ Architectural firms :They might supply actual drawings of the building, showing a wall here, door there, lobby over there, where carpet will be installed, where concrete will be used. This represents the physical building. υDesign Drawings

22. 22. Weight 22υ Space υ Site υ Connectivity υ Cooling υ Power υ Insurance and building code υ District υ Budget υDesigning for Data Center Capacities A major problem in designing a data center is determining how to support equipment of known quantity and capacities, or determining the quantities of equipment of unknown capacities for a data center of known capacities . Some of the major limitations are:

23. 23. Data Center Structural Layout The data center must be designed to accommodate diverse hardware designs and

requirements, and possibly equipment from different manufacturers. Determining RLUs is the best way to decide how the space will be filled with equipment, and with this information, where the equipment will be placed. The following general guidelines should be used in planning the initial layout of the room, keeping future planning in mind. 23

24. 24. Not knowing the structural load rating of raised floors and ramps 24υ Aisles aren't wide enough for newer or bigger machines υ Building in a space with the restrooms built right in the middle υ Building in a space with the restrooms built right in the middle υ Building in the 43rd floor of a high rise building along the San Andreas fault line υ Building in the basement of a building that overlooks a river υ Building in an area with a subfloor to ceiling height of ten feet υStructural Considerations

25. 25. Raised Floor A raised floor is an option with very practical benefits. It provides flexibility in electrical and network cabling, and air conditioning. A raised floor is not the only solution. Power and network poles can be located on the floor and air conditioning can be delivered through ducts in the ceiling. Building a data center without a raised floor can address certain requirements in ISP/CoLo locations. Wire fencing can be installed to create cages that you can rent out. 25

26. 26. Aisles and Other Necessary Open Space Aisle space should allow for unobstructed passage and for the replacement of racks within a row without colliding with other racks. The optimal space would allow for the turn radius required to roll the racks in and out of the row Also, rows should not be continuous. Unbroken rows make passage from aisle to aisle, or from the front of a rack to the back, very time consuming 26

27. 27. Command Center 27 Though an optional consideration, for some companies a separate Command

Center (also called a Command and Control Center) is useful for controlling access to the consoles of critical systems. This is just one of the many security devices used in the data center. In disaster recovery scenarios or other critical times, the Command Center is a key area. In many corporations where computer technology is at the core of their business, this Command Center also serves as a “war room” in times of crisis.

28. 28. Planned redundancies 28 Tire1 Tire2 Tire3 Tire4υ Connectivity to make the devices in the racks available to users υ Cooling to keep the racks from overheating υ Power to run the racks υ Locations on the floor that can support the weight of the racks υData Center Support Systems A data center must provide certain services:

29. 29. “Planning for Equipment Layout” 29υ “Using RLUs to Determine In-Feed Capacities” υ “Creating RLU Definitions” υ “Determining Criteria for RLUs” υ “Data Center Evolution” υ “Purpose of Rack Location Units” υ “Data Center Capacities” υ4.Determining Data Center Capacities

30. 30. Equipment capacities: The various devices (typically equipment in racks) that could populate the data center in various numbers 30υ Data center capacities: Power, cooling, physical space, weight load, bandwidth(or connectivity), and functional capacities υData Center Capacities The design of the data center is dependent on the balance of two sets of capacities:

31. 31. In a data center, most devices are installed in racks. A rack is set up in a specific location on the data center floor, and services such as power, cooling, bandwidth, etc., must be delivered to this location. This location on the floor where services are delivered for each rack is generally called a “rack location.” 31υ The Rack Location Unit (RLU) system is a completely flexible and scalable system that can be used to determine the

equipment needs for a data center of any size, whether 100 or 100,000,000 square feet. υPurpose of Rack Location Units

32. 32. Nowadays , Computers have become a lot faster and a lot smaller. The data center that used to house just one machine now holds tens, hundreds, perhaps thousands of machines and you have different types of servers, storage arrays, and network equipment, typically contained in racks. 32υ In the past, there were mainframes. There was usually only one of them for a company or a data center. The mainframe had a set of criteria: How much power it needed, how much heat it would give off per hour, how large it was, and how much it weighed. These criteria were non-negotiable. If you satisfied these criteria, the machine would run. If you didn't, it wouldn't run. υData Center Evolution

33. 33. 33

34. 34. we should examine the six criteria used in the determinations. These are power, cooling, physical space, network connectivity, rack weight, and logical capacity. The following figure shows in-feeds from power, HVAC, and network connectivity. It also shows that the other three criteria, physical space, weight specifications, and functional capacity, are aspects of the rack 34υDetermining Criteria for RLUs

35. 35. How much power the rack will draw The last item is best described in watts. This information should be part of the manufacturer's specifications. However, if the specifications don't tell you how many watts the device will draw, you can calculate this from the BTUs-per-hour rating of the rack. You will also need to know if the rack has redundant power. 35 BTUs per hour 3.42 = wattsυ Whether it's single phase or three phase υ Its voltage and amperage υ The outlet type υ What is needed to plug in the rack υPower You need to know the following:

36. 36. At minimum, either BTUs per hour or watt usage must be available from the HVAC manufacturer. The requirement is to deliver enough conditioned air to the rack to meet the BTUs per hour requirement 36 Watts X 3.42 = BTUs per hourυ Cooling requirements are specified as BTUs per hour. This should be part of the manufacturer's specifications. If it is not, you can calculate it from the amount of watts the machine uses. υCooling
37. 37. To effectively plan connectivity outside the data center, your ISP service bandwidth should meet or exceed the total capacity of your data center'υ The primary concern with bandwidth (connectivity) is the network and storage cabling within the data center. When determining the bandwidth part of the RLU, the concern will primarily be whether or not there are enough connections for the rack to interface with other devices. υBandwidth 37 s inbound and outbound bandwidth specifications.
38. 38. With this knowledge, and some forethought as to the future weight that will be added, you can decide whether the existing floor can handle the load. Or, if you are building to suit, you can plan for a subfloor and raised floor that will exceed the weight demands. There are a few different floor load capacities to consider: • Total floor load • Total tile load • Point load of tile 38υ It is critical that you know not only the individual weights of each type of rack that will reside in the data center, but the combined weight of all of them. υWeight
39. 39. The free space needed for aisles, row breaks, ramps, and free air circulation (typically 40 to 50 percent of the total square footage). 39υ The cooling dimensions of the rack (the physical dimensions plus the extra space needed for intake and exhaust of air as defined by the rack cooling profile). Cooling dimensions are described in the previous "Cooling" section. υ The width and depth dimensions of the rack. υPhysical Space

40. 40. Functional capacity is required only to determine the quantity and type of RLUs you will need to meet the project scope. 40υFunctional Capacity
41. 41. “Data Center Site Selection” 41υ“Geographic Location” υ5. Site Selection When housing a data center in an existing building, several design issues must be considered to choose the best location. Careful planning is essential to assure that a location will meet not only immediate needs, but future needs as well
42. 42. Vibrationυ Electromagnetic Interference υ Industrial Pollution (Factories , Manufacturing facilities , Sewage treatment plants , Farms) υ Fire 42 Man-Made Hazards υ Temperature Extremes υ High Winds υ Tornados and Hurricanes υ Seismic Activity υ Flooding υGeographic Location Choosing a geographic location for the data center could mean many things. Potential problems in the geographic location might not be obvious. Resource availability and potential problems, whether natural or man-made, are critical issues and uncovering them requires careful research. Natural Hazards
43. 43. Room for Expansion 43υ Environmental Controls υ Risk of Leaks υ Isolation From Contaminants υ Raised Flooring υ Access υ Security υ Retrofitting an Existing Site υData Center Site Selection The location of the center must be based on numerous criteria, including those discussed in the following sections.
44. 44. Is there space for a separate Command Center? 44υ Will the existing subfloor be able to handle the weight load? υ Is the floor-to-ceiling height adequate for a raised floor, ceiling plenum, and equipment height? υ Can a raised floor be added? υ Can walls be added? υ Can walls be removed without creating structural instability? υ Is there room for future expansion? υ Is the area isolated from contaminants and liquid leaks? υ What are the possibilities

for controlling access? υ Where will HVAC units be placed? Inside the area? Outside? υ Is there adequate access for moving in and rotating large equipment? υ What is the proximity of the area to chillers and condenser units? υ Is there enough room for required equipment? υ What is the general layout of the area? υRetrofitting an Existing Site

45. 45. Make sure the design includes fast-acting fire control such as FM200. 45υ In situations where you must share data center space with other companies, an effective means of segregating the space should be considered. υ Make sure the design includes surveillance cameras, motion detectors, and alarms. υ Design the area to limit and control access. υ Avoid sites with windows leading to uncontrolled areas. υ Locate the data center inside the existing building so there are no exterior windows or doors. υ A separate control room and remote access to the systems to minimize the traffic through the data center. υSecurity Not all businesses have a need for high-level security, but most businesses must make sure their data centers are secure from vandalism, industrial espionage, and sabotage.

46. 46. Freightυ Area for a loading dock υAccess Aside from security access considerations, the potential site for the data center should be set up for the loading and unloading of large items such as HVAC units and computer racks. RLU design to ensure open pathways within the data center 46υ Adequate space at corner and column areas υ Adequate turning radius space for racks and vehicles υ Ramps at floor-level height changes υ Wide hallways υ Wide aisles υ Wide doorways υelevators

47. 47. With the reduced space between floor and ceiling, is there enough space 47υ Is a raised floor a viable option for the available space? υ How will wiring, cabling, and outlets be run? υ Are there structural items in place that

might obstruct the free flow of air below the floor? υ Consider the amount of open plenum necessary to channel air for cooling. Too little space will cool inadequately, too much space will cool inefficiently. υ How high can the floor be raised? υRaised Flooring

48. 48. Make sure the exhaust from generators or other sources of exhaust do not enter the intakes of air handlers serving the data center. If the data center must be located near these hazardous locations, adequate filtering systems must be added to the design. Also, maintenance schedules for the filtering system should be more frequent. 48υ Isolate the data center from contaminants or contaminant-producing activities. Avoid locating the center near print rooms, machine shops, wood shops, loading docks, and areas that involve the use of chemicals or generate toxic vapors or dust. υIsolation From Contaminants

49. 49. If you must locate the center where there is a risk of leaks, make sure you design in a way to move water out of the room. Consider troughs under the pipes that are adequate to handle the water from a pipe failure and will carry the water out of the room without overflowing. 49υ Liquids pose another serious hazard to data center equipment. Despite precautions, water pipes and water mains can leak or burst. If you plan to locate the data center at a pre-existing site, make sure you know where all water pipes, valves, pumps, and containments are located. υRisk of Leaks

50. 50. The type of air conditioning system chosen for the data center, and the location of the units, might determine the viability of a location. Chilled water units must be connected to chillers located in the building or an adjoining support facility, and might require cooling towers. 50υEnvironmental Controls

51. 51. “Local Building Code” 51υ “Fire Rating” υ “Air Flow and Pressure” υ “Floor Load Capacity” υ “Anatomy of a Raised Floor” υ6. Implementing a Raised Floor The purpose of a raised floor is to channel cold air from the HVAC units and direct it up where it’s needed to cool equipment, act as an out-of-the-way area to route network and power cables, and act as a framework for equipment grounding. It also provides a sure foundation for data center equipment.
52. 52. Ramps and Lifts 52υ Routing Wires and Cables υ Placement of Wireways and Cable Trays υ Cable Trays υ Wireways and Outlets υ Plenum υ Tile Construction υ Floor Tiles υ Support Grid υ Floor Height υAnatomy of a Raised Floor A raised floor is generally constructed on a grounded framework, with a load surface of two-foot square tiles (also called panels). The space beneath the floor is called the plenum.
53. 53. 53
54. 54. Rolling load 54υ Static load υ Point load υFloor Load Capacity One of the most important issues to be concerned with in the early stages of the data center design is weight. It is important to know how much load will be placed on the raised floor so that a support grid and tiles with an adequate load rating can be ordered.
55. 55. The arrangement of solid and perforated tiles (in different perforation percentages) to deliver air at the correct pressure to each rack.υ How much cooling is needed per rack. (Is it the same for all racks, or do some racks need more cooling than others?) υ The initial pressure (velocity) of the air. υ The initial temperature of the air. υAir Flow and Pressure 55 Calculating air flow and the amount of air pressure needed to cool the data center involves a number of factors:
56. 56. The raised floor system should be in compliance with the specifications laid out in the National Fire Protection

Association document, NFPA 75: Protection of Electronic/ Data Processing Equipment within the USA, or relevant national standards outside the USA. 56υFire Rating

57. 57. Local building code could have something to say about how you implement the raised floor. This might be how many tile pullers (the device used to lift tiles) you need for the size of your raised floor. Or, inspectors could question the power distribution. 57υLocal Building Code

58. 58. “Site Power Analyses” 58υ “Electrostatic Discharge” υ “Electromagnetic Compatibility” υ “Wiring and Cabling” υ “Input Power Quality” υ “Signal Reference Grid” υ “Grounding and Bonding” υ “Power Distribution System Design” υ7. Power Distribution

59. 59. Installation and Placement 59υ Maintenance Bypass υ Sharing Breakers υ Backup Power Generators υ Uninterruptible Power Supply υ Multiple Utility Feeds υ Assessing Power Requirements υPower Distribution System Design well-designed electrical system for the data center ensures adequate and consistent power to the computer hardware and reduces the risk of failures at every point in the system. The system should include dedicated electrical distribution panels and enough redundancy to guarantee constant uptime

60. 60. Equipment Grounding Conductor Impedance 60υ Compliance With the NEC (National Electrical Code) υGrounding and Bonding Grounding is the creation of a path to an electrically conductive body, such as the earth, which maintains a zero potential (not positively or negatively charged) for connecting to an electrical circuit.

61. 61. Signal Reference Grid A Signal Reference Grid (SRG) is a means to reduce high-frequency impedance (also called noise) so that a device or outlet has the lowest impedance path to earth ground. This grid has multiple paths to ground to ensure that grounding loops do not develop. 61

62. 62. Emergency Power Control 62υ Lightning Protection υ Voltage Spikes υ Harmonic Content υ Power Conditioning Technology υInput Power Quality The plans for the data center should include a well-designed power and grounding system to maintain appropriate conditions and avoid unplanned power outages. Numerous factors can disrupt, degrade, or destroy electronic systems. High frequency, high amplitude noise, high ground currents, low power, surges and sags in voltage, harmonic distortion, and other factors will affect the proper functioning of equipment. It is essential that power conditioning technology be used to protect the data center equipment.
63. 63. Power Distribution Units (PDU) 63υ Higher Amps and Single-Phase or Three-Phase υWiring and Cabling All wiring and cabling should be designed and installed in accordance with NFPA 70 of the National Electrical Code, or superseding national or local codes. All wiring and cabling should be run in an orderly and efficient manner, not like the "spaghetti"
64. 64. Electromagnetic Compatibility Electromagnetic interference (EMI) and radio frequency interference (RFI) is radiated and conducted energy from electrical devices that produce electromagnetic fields. The electrical noise currents associated with these can interfere with the signals carried by the electronic components and the cabling of equipment. Sources of EMI and RFI can be inside or outside the data center environment. Common external sources are airports, telecommunications or satellite centers, and similar facilities. Internal sources include the hardware itself. Sun equipment is tolerant of most common EMI/RFI levels. If high levels are suspected, a study should be conducted to determine whether shielding or other remedial actions are necessary. 64
65. 65. 65

66. 66. Electrostatic Discharge Electrostatic discharge (ESD) is the rapid discharge of static electricity between bodies at different electrical potentials and can damage electronic components. ESD can change the electrical characteristics of a semiconductor device, degrading or destroying it. It might also upset the normal operation of an electronic system, causing equipment to malfunction or fail. 66
67. 67. The impact of power disturbances on data center equipment 67υ The source of power system disturbances υ The quality of the AC voltage supplying power to the equipment υ The soundness of the power distribution (wiring) and grounding systems supplying power to the equipment υSite Power Analyses Power disturbances can have numerous effects on sensitive electronic equipment, including data errors, system halts, memory or program loss, and equipment failures. Since it is often difficult to determine whether these problems are caused by power disturbances or by unrelated electronic equipment or software failures, a power system survey and analysis could be required
68. 68. “Air Distribution” 68υ “Monitoring Temperature and RH Levels” υ “Humidification Systems” υ “Air Conditioning Systems” υ “Electrostatic Discharge” υ “Relative Humidity” υ “Temperature Requirements” υ “Reasons for Environmental Control” υ8. HVAC and Other Environmental Controls (heating, ventilation, and air conditioning)
69. 69. Data centers need frequent air exchange 69υ Controls must be adaptable to changes υ Data centers need precise cooling υ Cooling must be delivered where needed υ Need for cooling υReasons for Environmental Control
70. 70. Doors left open 70υ Removal of floor tiles for subfloor work, such as cabling υ Installations, deinstallations, or reconfigurations of hardware υ Failure of any part of the HVAC or support systems υ Component failure (or even

scheduled maintenance) of the environmental support equipment υTemperature Requirements In general, an ambient temperature range in the data center of 70 to 74 F (21 to 23 C) is optimal for system reliability and operator comfort. Most computer equipment can operate within a wide psychometric range, but a temperature level near 72 F (22 C) is best because it is easier to maintain safe associated relative humidity levels.

71. 71. Relative Humidity 71 Relative humidity (RH) is the amount of moisture in a given sample of air at a given temperature in relation to the maximum amount of moisture that the sample could contain at the same temperature.

72. 72. Air Conditioning Systems 72

73. 73. Humidification Systems Humidification can take place within the air conditioners, or by stand-alone units. In some data centers, it might be better to introduce moisture directly into the room where it will mix easily with the ambient temperatures. This can be done with individual humidifiers, separate from the HVAC units. 73

74. 74. These monitoring systems can use SNMP protocol to integrate into overall data center management systems. 74υ Ideally, the monitoring system should be integrated with a tracking system for all parts of the center. υ The monitoring system should have critical alarm capabilities. υ The monitoring system should have historical trend capabilities. The historical psychometric data can be used to analyze seasonal changes and other outside influences. υ The room condition feedback should not be based on one sensor in one part of the room. A single sensor might tell that conditions are perfect, but in truth, they are only perfect in that particular part of the room. υMonitoring Temperature and RH Levels

75. 75. “Verification” 75υ “Labeling and Color Coding” υ “Avoiding Spaghetti” υ “Points of Distribution” υ “Creating a

Network Cabling Infrastructure” υ9. Network Cabling Infrastructure

76. 76. Modular Deυ Determining Connectivity Requirements υ“Creating a Network Cabling Infrastructure” 76 Hierarchy of the Network Structureυsign

77. 77. Network sub-switches 77υ Cat6 and fibre ports for cross-patching υ Network Terminal Servers (NTS) υ“Points of Distribution” A Point Of Distribution (POD) is a rack of devices and patches that manages a certain number of RLUs (which you can think of as a group of devices). PODs allow you to distribute both the physical and logical networking cables and networking equipment into modular and more manageable groups, and allow you to centralize any necessary cross-patching. All of the cabling from a group of devices can connect to the network through the POD. The POD rack contains three things:

78. 78. Avoid messy cable routing on the floor as shown in the following figure. This creates several hazards and liability issues. 78υ Label each cable at both ends so that the floor doesn’t need to be raised to follow cable routing. (See the following section, “Labeling and Color Coding.”) υ Route cables, whenever possible, under the tiles of the raised floors, preferably in cable trays. Don’t lay cable on the ground where it can block air flow and create dust traps. υ Use cable ties to keep cables in ordered bundles. υ Use the correct length of Cat6 or fiber cables to go from point to point. This avoids the need to coil or otherwise bundle excess cable. υ“Avoiding Spaghetti”

79. 79. “Verification” Each patch panel port should be verified and certified by the installer as part of the contract. You should also have cable testers, both Cat65 and fiber, available in the data center. With these you can verify that the patch-panel ports were done correctly and, if you have

questionable cables, you can find out whether they are good or not. This helps to eliminate doubt. 79

80. 80. “Storage” 80υ “Staging Area” υ “Loading Dock” υ10.Shipping, Receiving, and Staging In the data center design, a shipping, receiving, and staging area is an important consideration, particularly if the equipment will involve many reconfigurations in the lifetime of the center. Often, shipping and receiving take place in one area, usually near a loading dock. Staging can happen in the same area, or it could be in a separate location (recommended). Finally, storage facilities must be considered. This chapter has the following sections:

81. 81. “Noise Problems” 81υ “Security Problems” υ “Miscellaneous Disasters” υ “Earthquakes” υ “Flooding” υ “Fire” υ “Personnel Health and Safety” υ “Types of Hazards” υ11.Avoiding Hazards Potential hazards in a data center can range from mildly inconvenient to devastating. Some are difficult to avoid, but knowing what the potential hazards are in the data center area is the first step in preparing to avoid or combat them. This chapter contains the following sections:

82. 82. Noise from data center computers, large machinery, nearby industryυ Vibration caused by construction, large equipment, nearby industry υ Flooding from equipment failure, leaking plumbing, sprinkler systems υ Fire from electrical short circuits υ Tornados 82 Some human-created hazards are: υ Hurricanes υ High winds υ Earthquakes υ Flooding from rain, overflows, runoff υ Fire from electrical storms υTypes of Hazards Some natural hazards are:

83. 83. Avoid long rows of racks or equipment with no breaks 83υ Avoid blockages and doors that won’t open easily from the inside υ Ensure that personnel are able to exit the room or building efficiently υ Monitor air quality in the

room υ Authorized personnel should be trained to respond to emergency situations υ Keep room personnel to the absolute minimum υPersonnel Health and Safety From the very earliest design phases of the data center, the most important concern in disaster avoidance and recovery is for human health and safety. Equipment is important, but it always comes second to people.

84. 84. Carbon dioxide. Not recommended 84υ Halon 1301.Not recommended υ Wet pipe sprinkler υ Dry pipe sprinkler υ FM200 υFire Suppression Systems Modern gas systems are friendlier to hardware and, if the fire is stopped before it can do any serious damage, the data center might be able to continue operations. Water sprinklers are sometimes a viable alternative if saving the building is more important than saving the equipment (a water system will probably cause irreparable damage to the hardware). Gas systems are effective, but are also shorter lived. Once the gas is discharged, there is no second chance, whereas a water system can continue until the fire is out. Water systems are highly recommended in areas that contain a lot of combustible materials such as storerooms.

85. 85. Other steps required to ensure the safety of personnel should be outlined in local building codes. 85υ Can racks be secured to the floor and ceiling as a means of seismic restraint? υ Should equipment be held down with seismic restraints? υ Can the data center be located on lower floors where there would be less sway? υ Is the data center site built to withstand seismic disturbances? υ How often, if ever, do earthquakes occur in the region? υEarthquakes Some parts of the world have little or no history of earth tremors while others are plagued by them. For those building in Iowa, you probably aren't too concerned about earthquakes, whereas those building in California or Tokyo should consider the following:

86. 86. Is a surveillance system (video cameras) an option? 86υ What portions of the data center will be remotely accessible? υ Will people be able to remotely access the data center from anywhere? Will there be access restrictions to certain portions? υ Will the data center only be accessible through the Command Center? υ Where will the Command Center be located? Will it have a separate entrance? υ Are there windows or doors that could prove to be a security risk? Can they be blocked? υ What is the current security of the area? υ What level of security will the area need? υSecurity Problems The security of the data center is critical. Data centers not only contain valuable computer hardware, but the data in the machines is usually worth exponentially more than the 10s or 100s of millions of dollars that the equipment costs. Access should be restricted to only authorized and trained personnel. Several levels of barriers should be in place. The use of “traps” (a space between two doors) is a good idea for security as well as preventing the infiltration of particulate matter.

87. 87. Noise Problems With processors getting faster and disks getting more dense, the cooling requirements in data centers are rising. This means more fans and blowers to move more conditioned air. Noise can be a big problem in some data centers. The use of Command Centers, and devices like network terminal servers that allow remote access to a machine, allow users to work in a less noisy environment. However, you will need to have people in your data center some of the time. Ear protection should be used in particularly noisy rooms, and might even be required. The installation of noise cancelling equipment is useful but expensive. If people are working remotely most of the time, it might not be worth the cost. Ear protection might be adequate. If you do have people in the data center quite

often, the investment in noise cancellation equipment might be worthwhile. 87

88. 88. “Avoiding Contamination” 88υ “Effects of Contaminants” υ “Contaminant Types and Sources” υ12. Environmental Contaminants Particles, gasses, and other contaminants can impact the sustained operations of the computer hardware in a data center. These contaminants can take many forms, some foreseeable and some not.

89. 89. 89

90. 90. Regularly scheduled cleanings: Cleanings should be performed by trained professionals on a regular basis. These cleanings should be done with the same concern and regularity as data backups. 90υ Keeping contaminants out: Keeping contaminants from entering the data center should be done by minimizing traffic through the room, adequate air filtering, avoidance of improper chemical use, and positive pressurization of the room. Also, a properly constructed data center uses only non-shedding and non-gassing materials. If the data center is a retrofit of an existing structure, it might be necessary to change out or seal some existing construction materials. υAvoiding Contamination All surfaces within the controlled zone of the data center should be kept clean. This should be done by:

91. 91. “Pre-Hardware Installation Checklist” 91υ “Construction Criteria” υ “Codes” υ13. Codes and Construction To fully consider and implement the design of the data center, you have to construct a facility that will meet the project scope and also meet code.

92. 92. Fire codυ Electrical codes υ Mechanical codes υ Plumbing codes υThe Quagmire of Codes There are several building code organizations within the United States that have created their own set of codes. For example, there is Building Officials and Code Administrators International, Inc. (BOCA), International Conference of Building Officials

(ICBO), and Southern Building Code Congress International, Inc. (SBCCI). For fire codes, there is the International Fire Code (IFC), which is coordinated with the International Building Code (IBC). However, there is also the National Fire Protection Association (NFPA) who develop and publish NFPA 75 Standard for the Protection of Electronic Computer/Data Processing Equipment and NFPA 1 Fire Prevention Code 2000. Building codes Sewage codes 92υ Energy conservation codes υ Fire sprinkler ordinances υes

93. 93. Managing System Configurations In the world of systems management, it seems that there is always something that needs to be done. An important aspect of systems management is managing the system configurations. Managing the configuration includes managing the version and revision levels of system and application software, the types and versions of systems and adapters, the networks that the system is attached to, and the storage subsystems and their attachment mechanism, and all software, hardware, and firmware patches for all of the above. While this can seem to be a daunting task, particularly in large, fast-growing, complex environments, the configuration can be managed. 93

94. 94. Abstract 94

95. 95. Data Center Service Hardware & Software Network Infrastructure Building & General Specification 95

96. 96. CENTER Juniper Networks Pressυ Data Center and Cloud: Strategy and Planning , Shashi Kiran , Sr. Director ,Cisco υ World Wide technology .Inc υ Raj Jain, Washington University in Saint Louis (class lecture-2013) υResource : & Enterprise Data Center, Design and Methodology, Rob Snevely, 2001 96υAnalyst Conference - May17, 2010

How to Build Data Centers?

What are Data Centers?

Data Centers are a form of value-added service that offers resources for processing and storing data on a large scale for organizations of any size and even professionals may have at hand a structure of great power and flexibility, high security, and also qualified in terms of hardware and software to process and store information.

Currently we can define two main categories of data centers:

A. Private Data Center (PDC)
B. Internet Data Center (DC)
A Private Data Center (PDC) is owned and operated by private corporations, institutions or government agencies with the primary purpose of storing data resulting from processing operations, procedure and also in applications related to the Internet. Furthermore, an Internet Data Center (DC) is usually owned and operated by a provider of telecommunications services by operators of commercial telephony or other types of providers of telecommunications services. Its main objective is to provide various types of connection services, web hosting and equipment users. Services can range from long distance communications, Internet access, content storage, etc.

Services Offered in a Data Center :
Co-location Services: The client hires the racks of physical space and infrastructure for energy and telecommunication, but the servers, systems, management, monitoring and technical support are provided by the client. This relationship can be relaxed and it is customary to establish a contract with the terms and conditions, clearly defining the scope of services of each side. Telecommunications equipment includes.

Benefits For Your Company:

- Security;
- Speed of service;
- Support;
- Expert advice.

Basic data center service is included in a **colocation package of basic services** for the operation of equipment at no extra cost and maintaining the standard throughout the Data Center.

The Data Center Services Offered Are:

• Proactive monitoring and notification;
• DNS server (Domain Name Server) primary and secondary education;
• Technical support 24 x 7 x 365;
• Security building;
• Service reset (on / off equipment);
• Monitoring network;
• Redundant infrastructure;
• Hall of incubation (unpacking and setting)

Additional Data Center Services:

Hiring server colocation, customers receive a range of services. But it can also implement your purchase with options that will provide it with the most comprehensive range of services that a company can receive in colocation.

The client will have:

• Connectivity;
• Provision of access and bandwidth for Internet connection and the external network;
• Room for shared customers;
• cafe customers exclusive;
• IP addressing.

This service is dedicated to businesses that need high-quality infrastructure, connectivity between offices and / or the internet. This data center service is sold in square meters, cages (cages) or half rack and connections (IP, Internet, frame relay, ATM, etc.). From 64 Kbps.

Hosting:

Hosting offers a range of services suitable for companies wishing to leverage investment in hardware and software. **Data center hosting service** allows the customer to use the infrastructure of the data center and edge servers, and rely on highly qualified professionals who offer ongoing support to the customer.
The customer has the option of choosing equipment and software packages tailored according to the needs of business. Everything is custom designed and built to offer the best solution for each client. Thus, it guarantees the purchase of products that your company needs, allowing the client to fully dedicate the actions to focus on their core business.

The allocation of physical space in a rack and the amount available for equipment are calculated according to the defined configuration of servers and equipment hosting. All with the advantage of your company can set the bandwidth.

Customer Benefits:

• Economics of investment in fixed assets;
• Servers of art;
• Constant updating of software / hardware;
• Know-how in technology;
• Speed of service;
• Reliability of services rendered;
• Security;
• Installation of high standard.

Basic Services in The Data Center:

Using a hosting service, optimizes customer investments in hardware and software with exclusive use of the latest generation of **dedicated hosted servers**. Services that are essential for the functioning of the equipment are made available without additional cost and with high standards of our Data Center.

These Include:

• Capacity planning and network server;
• IDS (Intrusion Detection);
• Proactive monitoring and notification;
• Availability of IP address;
• Issue report online;
• Server to relay e-mail;
• DNS server (Domain Name Server) primary and secondary education;
• Technical support 24 x 7 x 365,
• Help Desk;
• Security building;
• Service reset (on / off equipment);
• Ensuring maintenance of logical security of the operating system;
• Full operation of the server until the operating system level;

• Incremental backup.
• Additional space structure in SAN (Storage Area Network);
• Traffic Gbytes per additional months;
• Additional space on internal disk
• Additional memory;
• Raid 1 / 5, with the possibility of protection service to the internal hard drive through replication of data between disks;
• Accounts of additional e-mails;
• Connectivity.

One aspect that must be observed in hiring a service of Data Center, is the type of access (co-location) that the user will have from the server service provider. The type of access will be defined by which the server will be accessed if necessary.

If the **co-location** is hired, the access is done by employees of the data center service provider, locally. If the co-location is remote, access is done through remote control software that will be chosen by the user. In this case the application is installed on the remote access server by staff of the service provider. Eventually one or more tools may need maintenance or there may be a need to install new applications. In such cases, the user must request the service provider to arrange whatever is necessary for the operation. While hosting the server, the user signs a term stating the legality of all software installed on your server.
One can observe that through co-location (location of a server's unique user, installed and operated in the structure of the provider), the user can benefit from a range of resources. A co-location provides high scalability, ie in case of a need for expansion of services or equipment, it can be done immediately, monitored 24 hours a day, 7 days a week (24X7), backup, optimization of costs operation and maintenance, network with high availability and load balancing.

1. Location of Data Center

The choice of location for deployment of the **Data Center** should be made taking into account the region, consistent with the Code of the City zoning, land size, easy access for delivery of equipment, high areas without flooding, existence of infrastructure basic sanitation, water, telephone and electricity.

Criteria For Site Selection of Data Center:

• Being close to points of presence to access networks of optical fiber enabling the connection of two different trunks.

• Availability of energy with the possibility of obtaining two power inputs
• Scalability, to allow increased building area over time.

2. Architecture of Data Center

The Main Components Of On Data Center Areas Are: Social Hall, And Meeting Rooms To Receive Visitors.

Administrative Area of Data Center:

• Operation, maintenance and storage of equipment.
• Living room equipment including servers for hosting and co-location and telecommunications room.
• cafe equipment segments electricity and air conditioning.
• Group Moto generator and fuel tank typically located in an area outside the Data Center.

The Goal Of Space Planning in Data Center Is:

• Have facilities with 60% of the total area devoted to the room of Data Center Equipment.
• Promote the “state of the art data center” premises since the operating system to the level of management of the database.
• Promote facility that reflects the image of a high-tech enterprise, business-risk high-yield investments, functionality and control.

Data Center is usually divided into three areas of physical security in increasing order of restriction of access:

Zone I: Public areas including the Lobby, the area for visitors and administrative areas.
Zone II: Areas of Data Center Operation.
Zone III: Equipment rooms, the heart of the Data Center, where the servers are located, the “shaft” of cables, power distribution units (PDUs), batteries and air conditioning machines.

3. Construction of Data Center

The construction should provide a solid structure composing secure facilities that complement and protect equipment and information residing in the Data Center.

Electricity: The Electrical segment consists of the Uninterrupted Power System (UPS), the Emergency Power System and Power Distribution Units (PDU).

The uninterrupted power system (UPS) has the function of providing energy for all data center equipment, including safety equipment and fire detection and alarm. It consists of sets of compounds by UPS batteries, rectifiers and inverters. These UPSs, redundant, connected in parallel; ensure a continuous supply of power even in case of failure of power transformers, power input or a set of UPS.

The banks of batteries are sized to feed the loads for a period of 15 minutes. This time is sufficient for starting and connection of diesel generators in case of power failure of the Concessionaire.

The power system consists of a group emergency diesel generator which will come into operation and connect to the electrical system of the Data Center automatically.

Generators are rated to withstand all the loads necessary for the operation of the Equipment Data Center during a power failure of the Concessionaire. The goal is to assist the operation 24 hours x 7 days a week, considering the conditions for preventive maintenance, adding new components and replacement operation after unplanned outages.

The power distribution units (PDU) are responsible for conditioning the signal to feed multiple devices at Data Center.

4. Air conditioning in Data Center

The segment of Air Conditioning has the function of maintaining a controlled temperature and humidity in the premises of Data Center. The segment includes the air conditioning system for cooling units and air handling system Distribution of Air Conditioning. It should be connected to emergency power generators.

The Cooling System to provide heating, cooling, humidification and de-humidification of the building.

The Air Treatment System must be separated into three types of area: Room Facilities: Data Center, area offices, equipment rooms Air Conditioning and Electrical. The separation is due to differences in sensible heat and latent heat of each area the conditions of temperature and humidity.

The Distribution System of Air Conditioning Equipment Room to the Data Center will use the system to supply air for the full set beneath the raised floor. This system involves inflating the raised floor at a minimum height of 60 cm., That depending on the amount of conduit, tubing, mats, etc., should have its height adjusted so as to allow air to circulate throughout the room Data Center. The goal is to operate 24 hours a day 7 days a week.

An adequate cooling is essential to maintaining performance and safety of operation of data center services.

A Data Center must ensure that the internal temperature in the areas of production varies by at most 1 ° C. For this, we have:

• Structures cooling N +1, ie for each equipment functioning, there is another reserve (ready to use);
• Modular refrigeration and air exchange;
• Scalability according to demand.

5. Fire Protection System in Data Center

The Data Center is a facility for electronics essentials such as servers and other types of computers and telecommunications equipment. In addition to meeting standards of the local fire department, the fire protection system should seek to avoid damage to the equipment in case of fire.

One of the best solutions for the firefighting equipment rooms is a combination of the Combat System with Pre Action Sprinkler (with dry pipe) above the raised floor system and Fire Fighting for Gas FM 200 below the raised floor.

The combat system with gas will be connected to a sensitive detection system and be the first to be fired. The gas is spread throughout the area, leaving no residue to damage sensitive equipment or to order a cleanup cost of the equipment.

The system of pre-action when triggered, triggers the discharge of water only in the sprinklers that have been operated by heat over the fire.

6. Supervision and Control System in Data Center

The control and supervision system continuously monitors the various segments of Data Center tracking items such as:

• Control of loading and parallelism of the generating sets
• Supervision and control of medium voltage panels
• Supervision and control panels for low-voltage
• Integration with system of generators
• Integration with system rectifiers

The system consists of computers with the latest technology capable of withstanding continuous use, appropriate systems for supervision and control. The same are redundant to each other, allowing high flexibility and performance system.

The Data Center also has a system of closed circuit television and access control that controls the entry and exit in various rooms and areas of physical security at Data Center.

7. Data Center Building Standards

Important factor of a Data Center is to implement and maintain methods of standard implementations of structured cabling seeking possible expansions, certification and ensuring safety and maximum use of the network.

Regarding the standards used, we can highlight the norms created by EIA / TIA (Electronic Industries Association / Telecommunications Industry Association) or ISO / IEC (International Standards Organization / International Electro-technical Commission called ISO / IEC 11801, equivalent to the EIA / TIA reprinted by ISO 568A).

7.1: TIA / EIA TSB 67 Standards

Transmission Performance Specifications for Field Testing of UTP cabling Cat5 (UTP end-to-end Performance Testing System) aiming system of Telecommunications (Telecommunications System Bulletin – TSB) is directed to test specifications for post-installation performance, the specifications include characteristics of field testers, test methods and minimum transmission requirements for UTP cabling systems. It cites factors affecting performance as the characteristics of cable, connecting hardware, patch cords and the cross-connect as well as total number of connections and the quality of the installation. The TIA / EIA TSB-67 refers to two test configurations:

a) Setting up the basic test of link (Basic Link test configuration): The basic test of link is used to check the performance of the cable permanently installed.

This Test Includes The Following Components:
• Up to 90m maximum horizontal cabling: cable includes a telecommunications closet (TC) to a consolidation point and the optional consolidation point to the outlet (cabinet) telecommunications. From one extreme to another on a horizontal cable connection.
• Up to 2m coord (string) to test the main tester for field connection to the site.
• Up to 2m coord test the remote connection to the remote unit of the field tester.

There Are Four Test Parameters In Each Link:
• Mapping (Map Wire) – Consists of 8 confirms the continuity of drivers end-to-end. Indicating possible pairs of short (shorts Between pairs), crossed pairs (crossed pairs), pairs Reverse (reversed pairs) and pairs spliced (split pairs).
• Length (Length) – Method of measuring the length of the cable by electrical means.
• Attenuation (Attenuation) – Method used in measuring the signal loss in the primary channel or link.
• NEXT – Measuring the amount of signal interference issue in a couple others. It is tested on all the endpoints of the link (endpoints, local and remote).

b) Channel Configuration Test
The test channel is used to verify the performance of the channel entirely. The channel has the following components:

• Up to a maximum of 90m of horizontal cable including the cable between the TC and a consolidation point (optional) and the consolidation point to the outlet (cabinet) telecommunications.
• Coord (umbilical connection machine or equipment) from the desktop.
• Cross-connections in telecommunications closets being made through or patch cable mapping coord.
• The total length coords, patch cords and cables and mapping coords area.

7.2: TIA / EIA TSB 72 Standard
Guidelines Centralized Fiber Optic Cabling (Centralized Optical Fiber Cabling). The TSB-72 was created to help in planning a cabling system fiber-to-the-desk (FTTD) from 62.5/125mm, using centralized electronics unlike the traditional method of distribution of equipment to individual floors to extend the connections from the desktop for cross-connect implementation. Use an interconnection between the horizontal and backbone cabling allows for better

flexibility, ease of management and can easily migrate to the cross-connection.

But you must have the maximum length of 90m in horizontal cabling. The distance from the horizontal cabling and backbone combined with the coords of the desktop, patch cords and equipment shall not exceed more than 300m.

The centralized cabling system shall be located within the same building working areas to be served. All displacement and change in activity should be performed at main cross-connect. Horizontal link should be added and removed in the TR. This must be a project of the cabling system enabling the migration to centralized mode pull-through, interconnect or amendment to an implementation of cross-connect. As a method to facilitate this migration, there must be within the scope of the project is having enough space in the TR allowing future growth and provision of additional patch panels and appropriate slacks that allow possible shifts in the cables of cable to the local cross-connect.

This slack can be stored without cables or fiber connectors. To fill the gap it has to prevent the maximum radius for curves in the cables are not violated thereby avoiding possible damage in optical fibers and others. The slack in cables can be stored indoor or on the walls of the telecommunications room, but should use boxes to protect slack optical fibers, due to their limitations and specifications.

With respect to the backbone, they are providing future horizontal links, this minimizes the need for placement of additional backbone cables. The fiber backbone must be capable of supporting current and future networking technologies, and require two fibers for each connection on the desktop.

The standard is required to use the following rules for labeling ANSI/TIA/EIA-606 cabling system centralized ANSI/TIA/EIA-568-A to ensure proper polarity of the fiber and connector specifications and methods of implementing connectorization guidance to the desktop AB and BA direction in the central cross-connect.

7.3: TIA / EIA TSB 75 Standard

Additional Horizontal Cabling Practices for Zones (Additional Horizontal Cabling Practices for Open Offices) methodology to meet modular office

environments increasing flexibility and reducing costs are discriminated as follows:

1. Horizontal Cabling reference to Open Offices (Horizontal Cabling for Open Offices). A termination point horizontal (telecommunications room multi-use) and a point of interconnection or horizontal intermediate (consolidation point) thus creating greater flexibility in open office layouts with modular furniture, where you have frequent changes. The telecommunications room multi-use (MUTO) and the consolidation point should be in a fully accessible, permanent location.
2. Cabinet Telecommunications Multi-Use (MUTO – Multi-User Telecommunications Outlet Assembly) The Cabinet telecommunications multi-use (MUTO) has the function to terminate point for horizontal cabling, consisting of multiple telecommunications closets on the same site. The modular coordinate extends from MUTO terminal equipment without additional intermediate connections. This configuration allows the moving of office plant without affecting the horizontal cabling, followed by the following criteria:
• Can not be installed on the ceiling.
• The maximum length of modular coords should be 20m.
• The modular coordinate MUTO connecting the terminal equipment must be labeled at both ends with a unique identifier.

Should be identified with the coordinate patch of greater length of the work area (modular cord). The length of this coordinate is calculated by formula:
C = (102 – H) / 1.2
W = C – 7 (The length of the cables, work areas may not exceed 20 m)

Where:
C = is the maximum combined length of cable that connects the user equipment to the telecommunications outlet, over the cable that connects the equipment in the telecommunications closet and cable that connects the patch panels;
W = is the length of the work area;
H = is the length of horizontal cable.

If you are already using fiber optic cables in data center, you can use any footage from the horizontal cables, desktop and cabinet telecommunication, keeping in mind that the total length should not exceed more than 100m.

When the optic cabling is centralized, one should follow the guidance of the TIA / EIA TSB 72.

3. Consolidation Point : It is the point of interconnection within the horizontal cabling actually performs a direct connection (straight-through) intermediate between horizontal cabling and cross-connect based on the horizontal cabling that goes to one or the MUTO telecommunications room on the desktop. There should be no cross connections between the cables in data center and you should follow these guidelines:
• Ensure that the total distance of the channel in any way beyond a distance of 100 meters.
• Ensure the fixing of the cables without violating the specifications and characteristics of each material complies with the requirements of minimum radius of curvature.
• Ensure that the consolidation point is at least 15m away from the telecommunications room, avoiding the additional NEXT due to the resonance of the multiple connections link near the cupboard.

7.4: ANSI/TIA/EIA-568-A Standard

Norm that characterizes the minimum specifications for structured cabling, classifying the components of the installation structure as follows:

1. Ease of entry (Entrance facility): Involves the ease of entry of cables, connecting hardware, protection devices and other equipment required for the building. The equipment inside the room can be used for connections to public or private networks.
2. Main cross-connect: The Hall of telecommunications equipment may have the same location of the main cross-connect. Cabling techniques that apply to telecommunications closets (TC) also apply to equipment rooms.
3. Backbone Distribution Interconnection between telecommunications closets, equipment rooms, and may be involved cables, cross-connects (main and intermediate) terminations, jumpers or patch coords for connections:
• There are guidelines to provide the maximum amount of life of at least 10 years to cables supported.
• Assume the distribution system of star topology and can be connected to a main cross-connect, intermediate or other main connection, but you must be careful not to have more than two tiered levels of main cross-connect.
• The recommended approach is to use one of the following features of connections among them, 100MHz UTP, STP-A DE 150MHz, Optical fiber cable 62.5/125 m, singlemode fiber optic cable or coaxial cable 50W

(recognized but not recommended for new installations).
• To make the choice of cabling to be used, there must be criteria for selection of the media depending on their characteristics such as flexibility (considering the supported services), life size and the location and quantity to be used.
• The recommendation of maximum distance from the backbone is also a very important factor for the choice of cabling.

4. Horizontal cross-connect is the name that refers to telecommunications closets (Telecommunications closet functions) which is to hardware in the connection of all horizontal cabling, intermediate cross-connections or even the cable of the backbone .
The cross-connections and interconnections can be said of connections between horizontal and backbone cabling or equipment connecting integrated circuits (hardware).

5. Horizontal distribution is part of the cabling system that connects the wiring from the desktop to the horizontal cross-connect in the TR, the horizontal cabling can find outlets for telecommunications in the area of employment, terminations and patch coords and jumpers in the TR. The horizontal distribution also possesses some important factors such as:
• **General design guidelines:** It has a target to meet the current specifications, thus facilitating the maintenance and replacements also considering the possibility of future equipment installations and changes in service since the horizontal cabling is the least accessible and is the subject to most of the activity of an implementation.
• **Topology:** The horizontal distribution system must have the standard star topology where the points of desktop wiring must be connected in a horizontal cross-connect to the telecommunications room located on the same floor of the desktop.
• **Distances:** The system must meet the average distance of 90m corresponding to the total length of the cable (leaving the work area to the cross-connect in the TR) is allowed where it should not be more than 10m in length for each path of coord area of work, jumping and equipment, but should not exceed 3m in length for path coords and jumpers that are used to connect equipment or the horizontal cabling backbone.
• **Media recognition of horizontal distribution:** You can use the 4-pair UTP 100MHz cable, STP-A 150MHz 2-pair cable or fiber optic 62.5 / 125 m (two fibers) as there are types of cabling for horizontal distribution according to the recommendation of the standards, but there are some observations:
– Coaxial cable although recognized, is not recommended for new installations.

– Hybrid cables (multiple types of media wrapped on the same cable) can be used if each type of media meet with the requirements of transmission and color specifications for this cable is also necessary to make the distinction of UTP multipair.

• **Criteria for selection of media:** For the desktop, it will have to be equipped with at least two telecommunications closets and we could be associated with voice and other data for the first aims must be 4-pair UTP cable 100W category 3 or greater, and for the second 4-pair UTP 100MHz, Category 5, STP-A cable 2 pairs 150MHz or two fiber optic cable, 62.5 / 125 m •.
6. Work area are defined as components of the landscape of work, ranging from cabins or telecommunications closets using 4-pair UTP cable with up to 3m long coords. The desktop is just a reference to the standard since it is quite variable and usually are never permanent and changing.

7.5: ANSI/TIA/EIA-569-A Standard

A major network cabling standards focusing on the specifications of infrastructure, structured cabling, providing specifications and project management for all building facilities. Identifying six infrastructure components: ease of entry, equipment rooms, backbone routes, provision of telecommunications closets, horizontal routes and work areas.

1. Ease of Entry in Data Center(Entrance Facility): It is defined by ease of entry into building or backbone of telecommunications services, and may contain devices interface with public networks. Necessarily the place to be dry and close to the vertical backbone routes.
2.Equipment Room in the Data Center: room whose space is intended to centralized location of equipment common to the employees, their location and design must be considered for the possibility of an increase in equipment and in its accessibility.
3. General Design Considerations for Data Center: Cafe equipment tends to be a centralized space for housing telecommunications equipment (PBX's, servers, routers, among others) of a building, being located near the route of the backbone. Its size has a minimum of 14m ², but to suit the characteristics of specific equipment, there is a need to make to make a project allowing a non-uniform occupation of the building, providing 0.07m in the equipment room space for every 10m usable floor space. If the equipment room is being designed to move, verify that the capacity of the floor will bear the weight of equipment to be installed and check for interference, vibration, altitude, HVAC equipment (dedicated equipment room), lighting, energy and fire prevention.
4. Inter-Building Pathways in Data Center: In a campus environment, routes, inter-building are needed among which it makes the connection of

separate buildings. The lists of standard ANSI/TIA/EIA-569-A basement, ground, aerial and tunnel are the main types of routes used.

5. Underground Inter-Building Backbone Pathways in Data Center: An underground route is considered as a component of the ease of entry. For route planning, you should consider the limitations existing in the topology, ventilation to prevent accumulation of gas, vehicle traffic to determine the thickness of the layer covering the route and whether it should or should not be concrete, if groundwater constituted for conduits, it should have ducts and troughs, including manholes.

Cable Distribution System For Servers in Data Center

The power cords for the servers will be installed under the raised floor and arranged in layers or channels. Fiber and coaxial cables that interconnect the data room of routers and switches to the server room have redundant facility, with a circuit walking under the raised floor and another next to racks of servers. The panels for the distribution of data cables will be distributed throughout the server room.

The design of the cabling is done according to structured cabling standards. 1:10 Distribution System Via Cable to WAN.

Fiber Optic:

The composition of the optical fiber provides conditions for the propagation of light energy through your core; it spreads light by successive reflections.

Optical fibers have some advantages over some of the traditional physical media such as coax and twisted pair. For example:

• Low transmission loss: decrease the number of repeaters.
• High transmission capacity increases the amount of information conveyed.
• Interference immunity and electrical isolation: the data is not corrupted during transmission.
• Security sign: the fiber does not radiate significantly propagated light, giving a high degree of safety information conveyed.

The modern fiber optic bandwidth has very large (x multi-gigahertz km) with low attenuation and low dispersion of the pulses emitted. For these systems, the fiber properties are those that comes with the lowest cost per km per channel that are installed.

The Use Of Optical Fiber Also Has Some Disadvantages Such As:

• Fragility of optical fibers without encapsulation
• Difficulty in fiber-optic connections
• T-type couplers with very large losses
• Lack of standardization of optical components

The transmission capacity (bandwidth) of an optical fiber depends on its length, its geometry and profile of its refractive indices. There are two main classes of fiber: single mode and multi mode fiber has several modes of propagation and in accordance with the profile of the variation of refractive indices of the shell with respect to the core, are classified into: step index and graded-index, the difference between them can be seen in the figure below. Its diameter is quite high, between 50 and 80 microns, making the beam undergoes reflection, limiting the signal range to about 2 Km Because of this multi mode optical fibers are used in local or campus.

Already the single-mode fiber has very small dimensions, and a transmission capacity than multimode fibers, the diameter of 10 microns, allows a propagating wave without reflection. The distance is significantly higher and the bandwidth available is almost unlimited. Singlemode fibers are used, especially in long distance networks, ie networks in metropolitan type Gigabit Ethernet, or type of SDH or DWDM backbones.

There are some characteristics of transmission in optical fibers that strongly influence the performance of fibers with a transmission medium, such as DWDM. In choosing the type of optical fiber for WDM systems in operation, must analyze factors such as attenuation, dispersion and nonlinear effects, because they are essential for good system performance.

Each type of fiber has some behavior for WDM operation which will result in restrictions for this type of operation. These restrictions have a direct impact on system performance, limiting the transmission capacity or decreasing the extent of linkages.

HARDWARE FOR DATA CENTRE

POPULAR FACTS ABOUT HARDWARE FOR DATA CENTRE

Maintaining data or information of or about a company is now the hardest the thing which the company officials come across. With technology advancing, it is very important to keep your data under protection. No company wants their personal data to be available easily to anyone at any time. With the increase in data to be stored, companies now spent more on hardware than software. The data center hardware includes the components that constitute the infrastructure of the data center. It includes the functional and non-functional devices vital for operating a data center.

Why Do We Need Data Centers?

The rapid expansion of cloud computing and people relying on cloud storage platforms to store their information has increased the demand for data centers. The data centers experience huge rush due to the number of users using the same platform. Therefore the data centers should be equipped with latest technologies like optical fiber cables and other devices to manage the flow of incoming and outgoing data.

Hardware Used In Data Centers

The hardware equipment used in the data centers includes routers, switches, firewalls, cables, and modems. Hard drives and tape drives are storage devices used. Cooling systems are also used to draw away the heat produced in the data centers. **The infrastructure used for cooling includes power generators and cooling towers.**

1. **Routers** – These are devices used for networking which forwards packets of data between networks of computers. The destination of the data packet is referred as a node and the routers are responsible for the correct delivery of the right packet at the required node. Thousands and thousands of routers are employed in a data center to forward the packets to different nodes.
2. **Switches** – This is also a networking device which is responsible for connecting devices together on a computer network by using the theory of packet switching techniques. The main advantage of the switches is that it only delivers data which are to be received at the required destination rather than sending all information to all places.
3. **Cables** – Most of the data centers employ optical fiber cables which have higher bandwidth than the normal metal cables which were used before. Another advantage is that it is difficult to tap information from the optical fiber cables. But these cables are hard to repair and very costly.
4. **Modem** – (modulator-demodulator) this is also a network device used in the data centers that encode digital information onto a modulated carrier wave so that it becomes suitable for transmission.

5. **Hard Drive** – This is a device used for storing data and uses the magnetic storage technique. These drives retain the information (digital information) using one or more continuously rotating disk-like structures that are coated with some magnetic material. These drives are commonly used in data centers to store the data that is received.
6. **Tape drives** – This is also storage devices which like hard drives uses magnetic tape upon which data is written and read from. These have sequential storage access, whereas hard drives provide direct access.

DATA CENTRE TOPOLOGIES

The data center network also plays a great role in controlling the traffic and also determines the number of hardware to be used. To understand better, you can check out this study of network traffic characteristics. In fact, the number of hardware needed vary with different network topologies. Therefore, topologies are chosen in such a way that the data is stored efficiently at the same time the method adopted requires comparatively less hardware. Here are some network topologies commonly used.

1. [illegible] DCN

2. ***Three – Tier DCN***
3. ***Dcell***
4. ***BCube***
5. ***Camcube***
6. ***FiConn***
7. ***Jellyfish***
8. ***Scafida***

Need of Cooling Techniques or Devices in Data Centers

The main motive of installing cooling devices or structures in the data centers is to maintain favorable conditions for the hardware to operate. These devices draw away the heat produced. The heat is then diverted to some heat sink.

The data centers experience a lot of traffic of data. Take the example of Facebook. This IT giant has numerous huge data

centers set up in different parts of the world to store a large amount of data that it accepts. With users increasing every minute, the data centers are also expanding to store the details of every user. The company spends a fairly high amount to store and circulate this data.

CLUSTER SYSTEMS

FULLY UNDERSTANDING CLUSTER SYSTEMS

Cluster systems or also known as computer clusters refers to a group or cluster of computers working together. This can be considered as a single system. The computer cluster can be lightly or loosely connected. The cluster systems are divided into nodes and each node has a specific task to accomplish. The task to be performed is assigned using the specially designed software.

The components required for a cluster is connected by local area network with each computer working on its own operating systems. In most of the cluster systems, the computers share the hardware required and sometimes share the operating system as well.

The idea of Computer clusters evolved after looking for a better way to efficiently manage the cost and availability of microprocessors, high-speed networks, and software. If these are to be used separately for each computer, it can cost the company a heavy sum.

BASIC UNDERLYING CONCEPTS

In this technique, the readily available computers are connected using a LAN. The activities and operations carried out by the computers in the network are monitored by software – clustering middleware. A cluster system can be a simple system consisting of two nodes which connects just two personal computers or maybe a supercomputer.

Cluster systems may consist of personal computers' connection but by the correct usage and implementation of cluster architecture, high performance can be obtained.

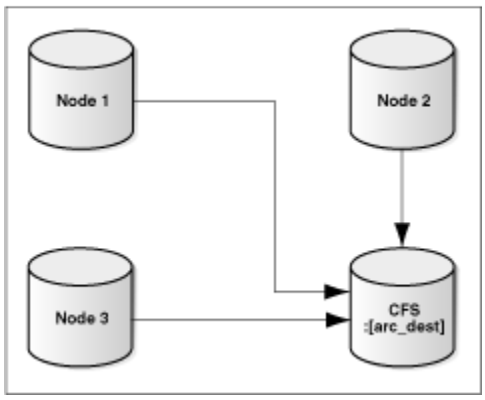
The cluster systems may be employed as a web service or to support intensive scientific calculations. In both the cases mentioned above, a high availability approach is to be used. Load balancing is another technique used in the cluster systems where configurations are done in such a way that cluster-nodes share the workload to improve the performance. For example, a cluster in web server may distribute the queries to different nodes, thus the response time can be minimized. Some cluster systems make use of the famous round-robin method to carry out the operations.

Uses Of Computer Clusters

The computer clusters or cluster systems are used in the weather forecast, in scientific computing and very high-speed clusters are designed for supercomputing.

Data Sharing

The computer clusters were introduced in the 1980s. Clustered file systems are now being used all over and are necessary for modern computer clusters. The message passing Interface (MPI) and Parallel virtual machine (PVM) are two commonly used communication techniques to communicate between the nodes of a cluster.



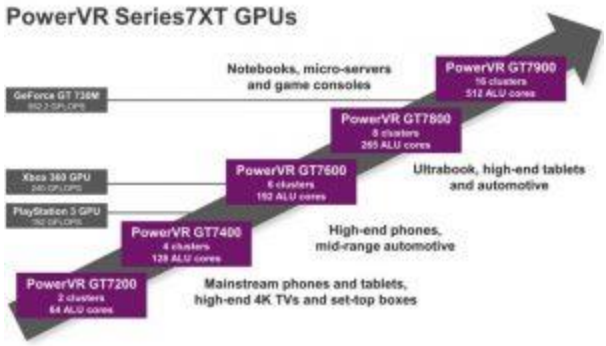
Advantages of Using Cluster Systems

The cluster systems are made keeping in mind the cost factor and performance but the system should also be made in such a way that the system should continue working even if a node malfunctions. The cluster systems can also enable data recovery options at the time of a disaster. It provides parallel processing of data.

Designing and Implementing Clusters

One of the hurdles that the technicians have to encounter is how to tightly couple or connect the nodes of the cluster so that it provides the maximum result. Beowulf system is a type of cluster system where application programs never get to see the nodes (slave computers) but are only allowed to connect or communicate with the master which is a computer that handles certain management and scheduling of the nodes or slave computers. The nodes usually have an operating system of their own version along with a local memory and some disk space.

Gaming Console Clusters



Gaming console clusters are easily available in the market which includes Sony Playstation clusters, Microsoft Xbox clusters, and Nvidia Tesla Personal supercomputer workstation.

In the past cluster computing systems were used to run on different computers with the same OS (operating system). Now the cluster nodes or the slave computers run on different computers with their own operating systems.

Data Center Hardware and Equipment

Data center hardware is all of the physical computing components found inside a data center facility. Essentially, it describes all of the equipment that's crucially needed for the location to carry out its core, assigned IT tasks.

Servers:

A server is essentially a top-of-the-line computer, capable of performing key compute tasks. A typical server is usually several orders of magnitude faster and more powerful than a standard smartphone or laptop. The server is more or less the brains of the operation.

Servers contain many of the same core components as standard home or office computers, including processing chips and hard drives. But, these internal components are often designed to withstand greater rigor and perform more quickly than the chips, hard drives and other internal components for in personal use hardware. For instance, while many computers still have a hard disk drive for

storage and memory, most servers now use solid state drives since they have fewer moving parts and are capable of performing more quickly.

Servers differ from standard computers in a few other ways. In particular, they often lack a standard visual display. They're visually analogous to the computer towers that powered desktops of yore.

Storage:

Some data centers contain hardware specifically for just storing data. Often these serve as either mere backup, just in case something happens to the original copies.

Other facilities have "cold storage" for dealing with older data that can't be deleted fully. For example, older financial records and pictures taken over a decade ago would be good use cases for cold storage.

Typically the cheapest kinds of storage are tape drives, which keep data on tapes similar to cassettes. Hard drives, whether hard disk drives or solid-state drives, will be deployed

occasionally for standalone data storage as well.

Data Center Networking: Switches, Modems, Routers and Cabling

Data center hardware also encompasses all of the equipment needed to move data to and from various computers and hardwares, and to communicate with wider networks like the internet:

- Switches ensure that data packets go to and from the right location, like modernized versions of the giant phone switch boards from decades ago.
- Modems ensure that data can be sent over a given network. Data center modems work essentially just like modems in your house or office that provide wider access to the internet.
- Routers enable multiple servers/computers to all be part of the same local area or wide area network.
- Physical cables will sometimes be deployed between servers and racks, particularly in

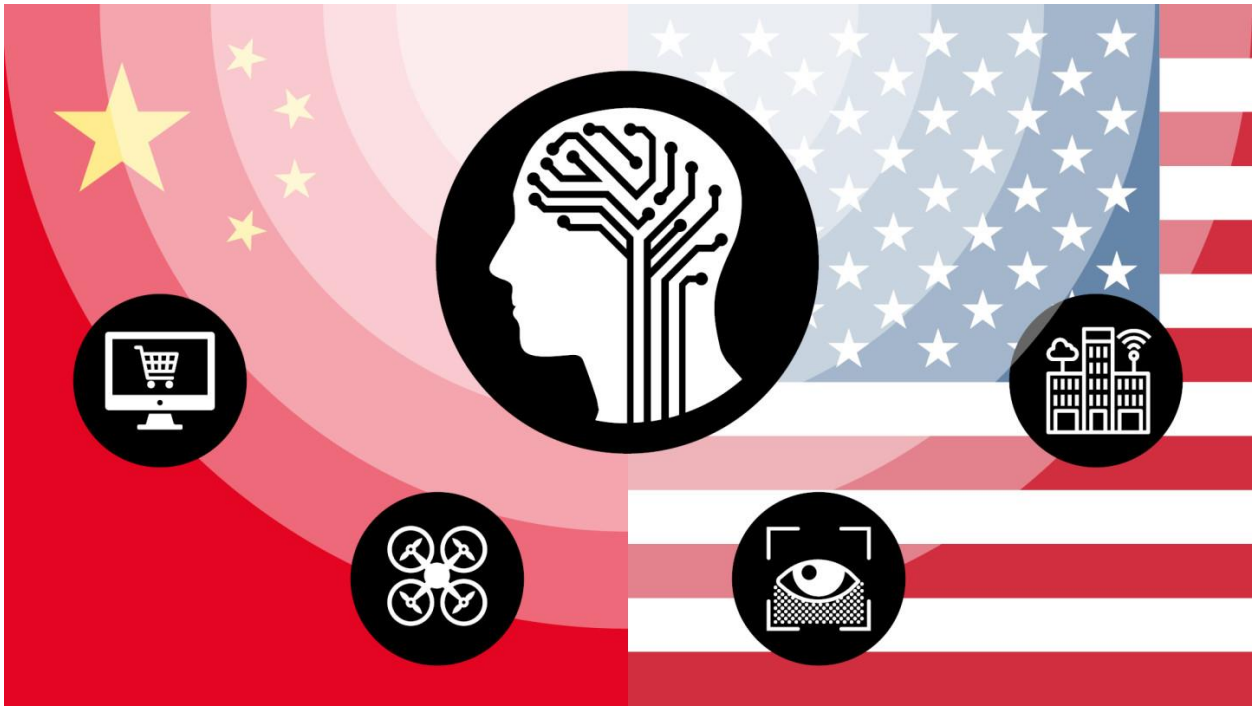
instances where wireless connectivity will not provide enough throughputs or is not reliable enough for the tasks at hand.

What is the Future of Data Center Hardware?

Increasingly, data centers have become less reliant on hardware to function, as virtualized components become more common. In particular, virtualized machines allow one server to essentially function the same as multiple servers, while the trend of software-defined networking allows data centers to be less reliant on physical networking infrastructure like switches and cabling.

Chapter 15:

IS IT WAS IT OR STILL A CENTERALATIC WAR OF BIG DATA CENTERS BETWEEN USA, CHINA, RUSSIA AND OTHER COUNTRIES?



China and USA compete to dominate big data

Beijing plans to be the world leader in the technology by 2030.

Algorithms trained on mountains of Chinese data may soon be making decisions that deeply affect the lives of people in the US.

Take Yitu Technology, a Shanghai-based artificial intelligence start-up that won top honors in two AI competitions in the US last year for its facial recognition technology. The system was built for Chinese law enforcement using data collected by the authorities. Or as the company boasts, it was honed on the “world’s largest portrait system, covering more than 1.5bn people.”

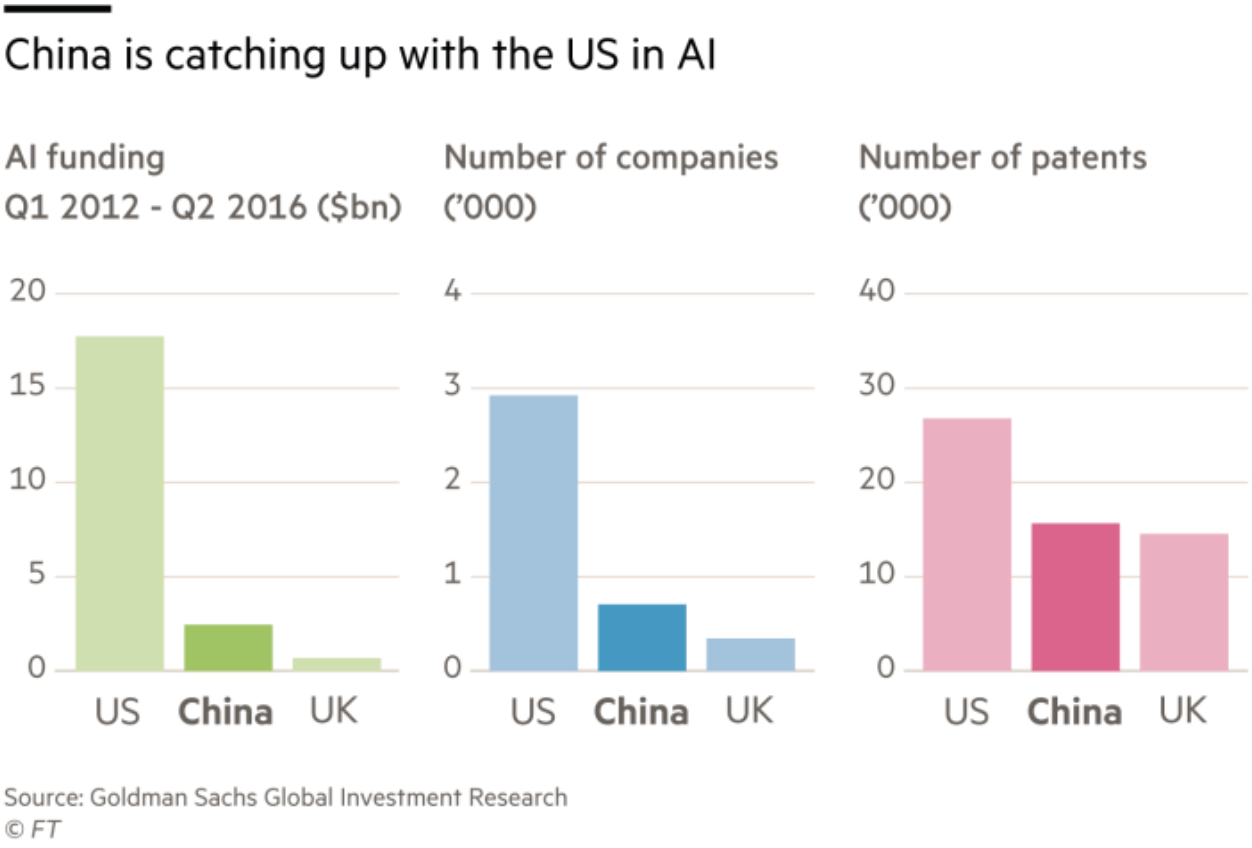
Yitu is now looking for customers in the US to put its software to work. “There are a lot of applications for this technology,” says Wu Shuang, who heads its Silicon Valley research group. This story is the first in a series on the growing competition between the US and China over AI, which will have an impact on the contest for military superiority, the

international trading system and the future of the Chinese Communist party.

It is not alone. Shenzhen-based Malong Technologies has also trained its image recognition algorithms on masses of Chinese data— in its case, by analyzing hundreds of thousands of photos from fashion shows to identify trends for clients in the garment industry. It says it is now trialing the technology with ecommerce companies in the US.

A “key difference in China is there are just more people, more data, more businesses — it’s just bigger,” says chief technology officer Matt Scott, a former Microsoft researcher who moved to China to Co-find the company. “Having access to that data in China, we can export [the technology] around the world.”

Algorithms like these are the advanced guard in a battle that will go a long way to determining economic leadership in the era of Big Data — a contest where China is catching up quickly and now vying with the US to be the dominant force.



The AI revolution is often thought of in terms of robots or drones that can do tasks once performed by humans. But its impact will also be felt from a less visible source — the ability to sweat the data the hardest. Machine learning systems that can find patterns by analyzing large data sets are at the cutting edge of today's artificial intelligence.

For some industries, deep learning — the most advanced form of the technology — has the potential to create value equivalent to as much as 9 per cent of a company's revenues, according to a report in April from McKinsey Global Institute.

That translates into trillions of dollars of potential economic value — and the US and China; are the clear leaders.

“If you look globally, it’s a two-horse race in AI,” says Michael Chui, a McKinsey partner who led the study.

Visitors to the China Public Security Expo in Shenzhen in October are scanned by facial recognition software © Reuters

In China, the AI boom has fed the country’s swelling sense of self-confidence in its expanding technology base. President Xi Jinping has made AI one of the central pillars of the Made in China 2025 plan to transform the country’s economy and has set a goal of being the world leader in the technology by 2030.

At the same time, China's advances are also contributing to an opposite paranoia in the US that its technology exceptionalism can no longer be taken for granted. The Trump administration's plans for a trade war with Beijing are motivated — at least in part — by fear of China's advances in new technology.

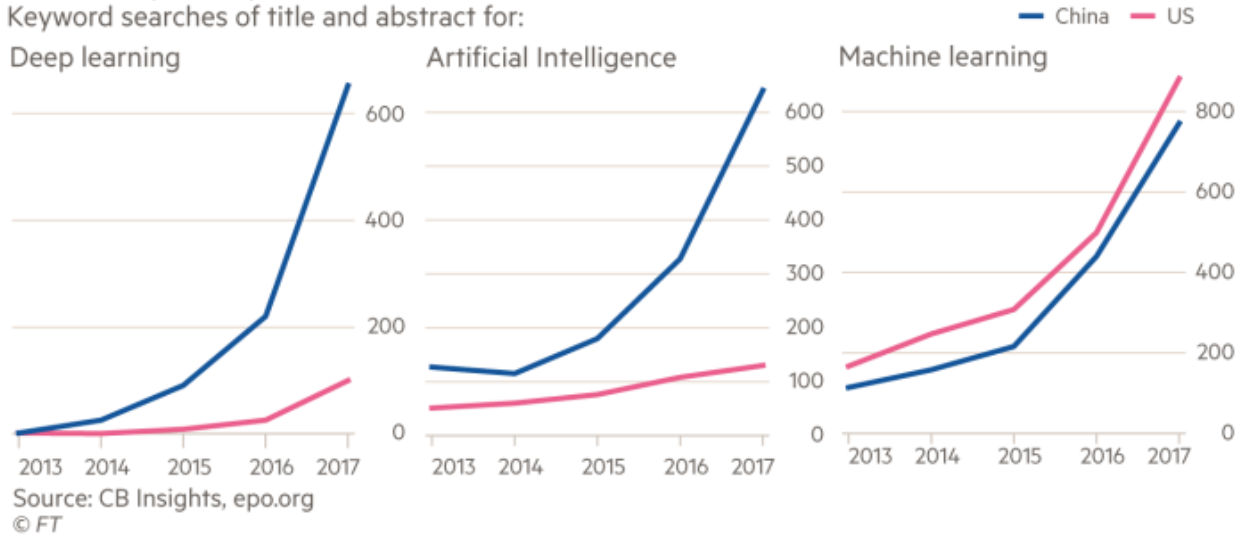
"It's clear that the US government sees itself in a tech arms race with the Chinese government," says Robert Silvers, a partner at legal firm Paul Hastings and former assistant secretary for cyber policy at the department of homeland security. "The US is taking the view that these kinds of technologies are so transformative that the country that gets the lead is going to have not just economic or tech advantage but also national security advantage."

One reason the contest over AI is so charged is that it is connected with a race to find a new military edge. As well as answering mundane customer queries and piloting driverless cars, the same technology can also be deployed to synchronise drone swarms, analyse pictures taken by spy drones and control autonomous boats.

Dominance in AI could bring a step change in warfare, says Sean Gourley, founder of Primer, a Silicon Valley AI start-up whose backers include the CIA's venture capital arm.

Technology shifts like this can undermine the military advantage of great powers.

“It’s likely to be coupled with the reordering of global power. Whoever is best at this will be in a strong position in 10 years’ time,” he says.



Russian president Vladimir Putin found his own way to raise the rhetorical stakes over AI last year: “Whoever becomes the leader in this sphere will become the ruler of the world.” According to most experts, the US still has a clear lead. It takes three things to be a world-class AI power: the most advanced algorithms specialised computing hardware, and a good supply of the raw material that machine learning systems depend on — data.

Last year's Go match, where a system built by Google subsidiary DeepMind trounced leading player Ke Jie in the ancient Chinese board game, was China's wake-up moment in AI, says one Google executive. "Only when the Russians launched the sputnik did the US realise how far they had come," he says. "China had that moment when they lost at AlphaGo."

That view is echoed by some in China. "For top talents, clearly the US will still be the main resource. I think there's no question of that," says Rong Jin, head of machine intelligence technologies at DAMO Academy, Alibaba's research programme. The perception in China is that Americans throw themselves into fundamental research and are heavy duty mathematicians — the disciplines at the heart of AI — while Chinese tend to study coding or engineering.

A Chinese police officer wears a pair of smartglasses with a facial recognition system in Zhengzhou, Henan province © AFP

Yet despite those advantages, China is rapidly narrowing the algorithm gap. When it comes to the output of China's research institutions, "the statistics are definitely rising sharply", says Oren Etzioni, who runs the AI research institute of Microsoft co-founder Paul Allen. He points to other signs of China's growing AI capability, from the reading comprehension test earlier this year in which AI newcomer Alibaba tied for top honours with traditional research power Microsoft, to the strong showing of Chinese researchers in the annual ImageNet competition for image recognition.

On the second category of hardware development, China has been slower to build the sort of homegrown chip industry needed to put it on the leading edge. That has been partly due to a series of decisions that effectively bar the acquisition of US chip companies, which started under the Obama administration and accelerated under President Donald Trump.

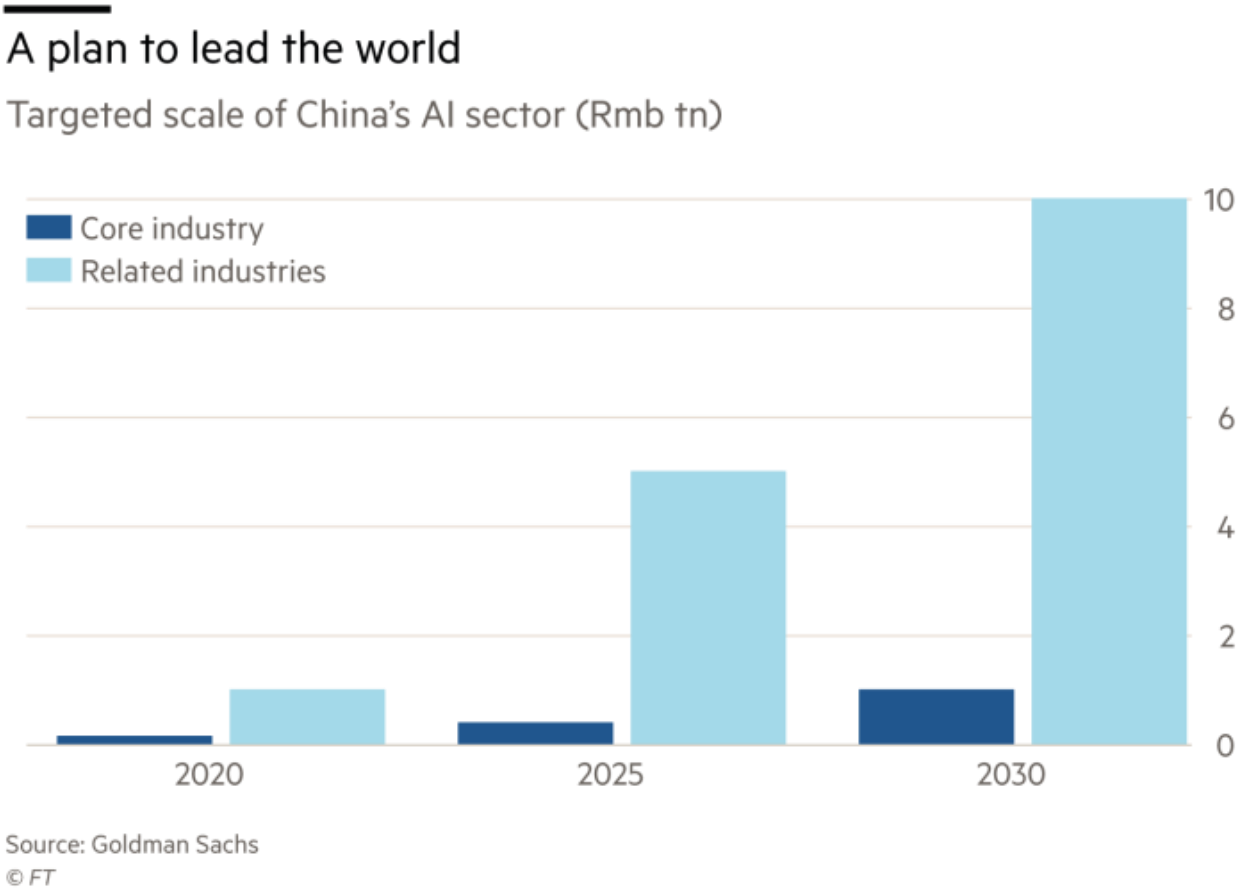
It is in the final area — the availability of raw data — where most experts believe China's AI advantage lies.

China has reams of data on its citizens and is not afraid to use it. This is partly due to a state that monitors everything from birth: facial recognition is so widespread you can be picked up for jaywalking and stopped from stealing tissue at the Temple of Heaven in Beijing.

But it is also a tribute to China's early move online: this is a country where people order, shop, pay and play online, leaving massive data footprints that enable merchants to accurately target ads and promotions. "The density of people is proportional to the density of data," says a leading Chinese machine intelligence scientist.

Chinese attitudes to data privacy are becoming slightly less lax, but regulations are still a million miles from Europe, which is at the other end of the spectrum and will introduce tough privacy rules later this month known as General Data Protection Regulation. Yet American companies like Facebook,

Google and Amazon also have masses of data, says Mr Wu at Yitu.



That suggests that general-purpose AI applications like facial recognition will be the preserve of all "the big platforms", regardless of their country of origin, says James Manyika, a partner at McKinsey. By contrast, more specialised applications could be perfected where the data are the richest. When it comes to manufacturing, for instance, China is "collecting a lot more data", he says.

This data advantage could be greatest in fields where regulation has made access to information harder, or prevented it being collected in the first place, according to some experts. Earlier this year, Google published promising research suggesting it could predict the risk of heart attack by using image-recognition software to study retinal blood vessels. The research relied heavily on UK Biobank, a database drawing on a detailed study of volunteers in Britain beginning in 2006.

Yet only 631 people in the Biobank had medical conditions relevant to the research. That made the data set “relatively small for deep learning”, Google said, reducing the effectiveness of the algorithm it was able to train on the information. Chinese medical AI researchers, by contrast, have been able to tap into far bigger data sets, according to one expert.

If China is rich in data, then it also has the economic opportunities to exploit it — something that has helped lure back many haigui, or returning “sea turtles”, like Mr Jin. AI is being harnessed in law, where machines have replaced stenographers in more than 6,000 courts; on the roads to manage traffic; in hospitals to detect tumours; and in Shanghai subway stations where you can buy tickets by talking to the machine. “AI has the biggest opportunity in China versus any western countries,” says Mr Rong.

Big tech leaders: China's president Xi Jinping with his US counterpart, Donald Trump © AP

Chinese executives talk about a smart city scheme that halved the time it takes to speed ambulances from depot to patient to hospital, by dint of switching traffic flows and traffic lights.

The smart cities scheme also offers another point of difference with the US: collaboration between state and private companies on a large scale. In addition to projects aimed at traffic management, crowd control, finding missing children and elderly and cutting down hospital waiting times— the list goes on— all the big tech players have joint research labs with government.

This is part of a broader experimentation that is lacking in the US, says Mr Wu. “Overall, the Chinese tech scene is more dynamic right now, particularly in terms of trying out new ideas and new products,” he says. “People are just trying out more new things.”

That has not been lost on investors in the US. One leading Silicon Valley venture capitalist puts the difference in AI opportunities bluntly: “The business is bigger and better in China.”

This economic momentum behind AI is closely aligned with a second powerful force: a sense of national mission. That has brought a hydrant of money and clear industrial policy. This state-led strategy is also closely aligned with national champions Baidu, Alibaba and Tencent — all private companies.

Washington has done much less to promote a national agenda. “What’s the national direction around AI and robotics [in the US]? It’s nothing. It’s missing,” says the Silicon Valley investor. “The government is flailing around.”

Worse, the Trump administration’s attempts to clamp down on immigration has upset the US tech industry, which has drawn heavily on overseas talent — not least Indian and Chinese engineers. The heads of AI at Apple, Facebook and

Microsoft, as well as Google's cloud computing division, were all born outside the US.

"We've seen more and more students choosing not to come to the US," says Mr Etzioni. "We're in the process of shooting ourselves in the head." He points to one sign of how the talent pendulum is swinging away from the US: Google and Microsoft have both opened AI research centres in China to tap the AI workforce there.

Yet the expertise advantage that the US has will not disappear overnight. Companies like Yitu are moving in the opposite direction because they believe the US west coast is still the magnet for many of the world's top engineering brains. "Half the AI engineers in Silicon Valley are Chinese," says Mr Wu.

Kai-Fu Lee: 'AI is mostly an enabler, it’s not a nuclear weapon' © Bloomberg

For some, artificial intelligence is a new arena for geopolitical competition with military overtones. But for others, it is the first truly open-source technology, with companies and nations alike sharing ideas to improve the lot of humankind.

“It’s probably the best worldwide collaboration anyone has ever seen,” says Kai-Fu Lee, who headed up Google’s China operations and now runs his own venture capital firm.

Many tech companies like to make the same argument. “We hope we can develop AI tech and share with all players,” Tencent chief operating officer Ren Yuxin told a gathering of developers last year in Chengdu. Tencent and Google last year

agreed to cross-license patents on a range of products and technologies.

“AI is mostly an enabler, it’s not a nuclear weapon,” says Mr Lee. “I worry about the cold war analogy.” He says it is being firmly driven by universities and tech giants rather than defence departments and any competition is at company rather than national level.

James Lewis at the Center for Strategic and International Studies in Washington agrees that the competition “is largely commercial and it’s largely between companies.”

Proponents of this view point to the open-source resources and willingness of scientists to upload their findings immediately rather publish in journals.

Data are not always a company-specific resource, but can be bought from providers — a classic gig economy business that uses shoals of workers to provide and label speech, image and other data in a way a computer can understand.

“There are reams and reams of data,” says Mark Brayan, chief executive of Appen, a Sydney-based company. “But is it the right data? Is it prepared and labelled in the right way? Is it of sufficient quality?”

AI Race between the US, Russia, and China… and where is Germany?

It pretty much does not matter what position one has on Artificial Intelligence (AI) and Machine Learning (ML), but the strong interest in AI and ML can hardly be denied. In fact, doing a quick informal poll among colleagues and friends, it seems that many Thanksgiving table and holiday conversations were dominated by either AI or cryptocurrencies (Bitcoin and the likes). Both of these technologies have the potential for significant disruptions and whether or not they are overhyped (they probably are) they provide huge fundamental value.

While I will write about cryptocurrencies elsewhere, here I want to talk about AI and the global race for leadership in that space. AI is dominating many discussions last but not least due to its sci-fi appeal and the thought experiments of various thinkers and leaders. This interest in the technology has

reached a point where AI is now a topic of politics and important verticals of interest include ethical questions (e.g., killer robots and the terminator scenario), AI security and safety (provable guarantees for the operating envelope and AI hacking), as well as the impact of AI on armed conflicts. In fact the question of leadership and dominance in AI research, development, and deployment has become central over the last few months with various world leaders and politicians acknowledging the potential role of AI. For example, Vladimir Putin recently said:

"Artificial intelligence is the future, not only for Russia but for all humankind.", "Whoever becomes the leader in this sphere will become the ruler of the world."— Vladimir Putin [Source: Wired]

[*Background information and more reads:* AI and its role in armed conflicts, fears AI might trigger WWIII (see also here), and China's strategy to become the front-runner in AI…]

The current discussion is dominated by the idea of an arms race between the US, Russia, and China [see e.g., here, here, and here] due to its potential and risk of a significant power redistribution

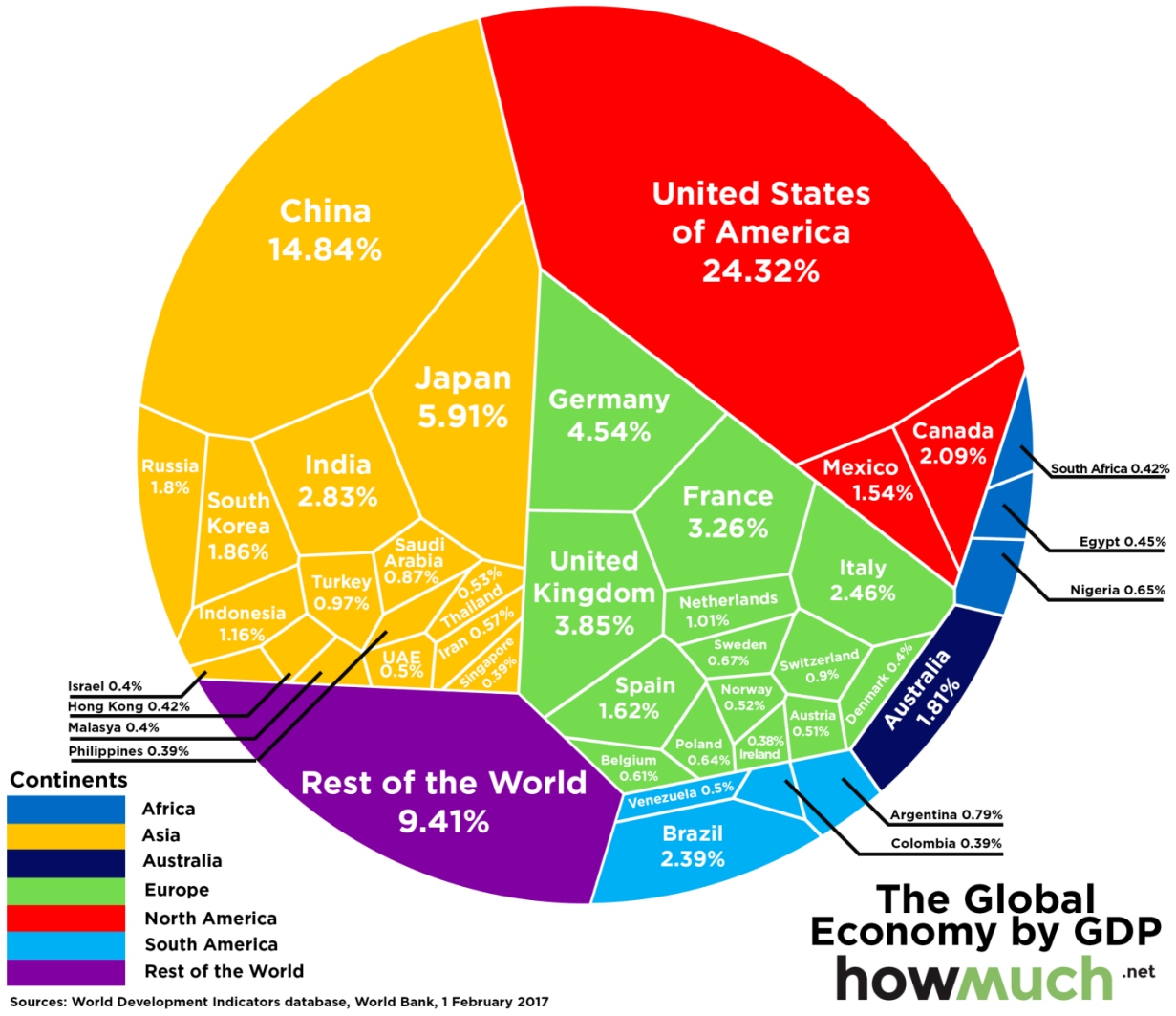
around the world and as such is considered being of highest political relevance and national interests. A significant part of this discussion is dominated by fear and winner-takes-all scenarios particularly fueled by Silicon Valley household names and scientists. So where does all the fear come from? If the entrepreneurs and investors from the valley have understood one thing then it is the power law (sometimes also Pareto effect or 80–20 rule) and the related Matthew effect. The Matthew Effect basically asserts that "*more gains more*" and the power law stipulates that a small fraction, say 20%, of a group is responsible for the majority, say 80%, of an effect. These two effects have been widely observed in society (e.g., wealth distribution in countries etc). With the advent of the internet this effect became significantly more pronounced and we saw an *accelerated version of the power law*: even smaller fractions, say 1%, accounted for even larger majorities of the effect, say 99%. This law has been guiding many investment decisions in the valley and Amazon, Google, Facebook, and partially Uber are great examples of this force being at play. The belief is now that AI will result in a significantly further accelerated version of the power law due to its potential impact: 0.0000001% of the participants will generate 99.9999999999% of the

effect, or put differently, the *winner-takes-all*. Marc Cuban aptly summarized this in a recent interview with CNBC:

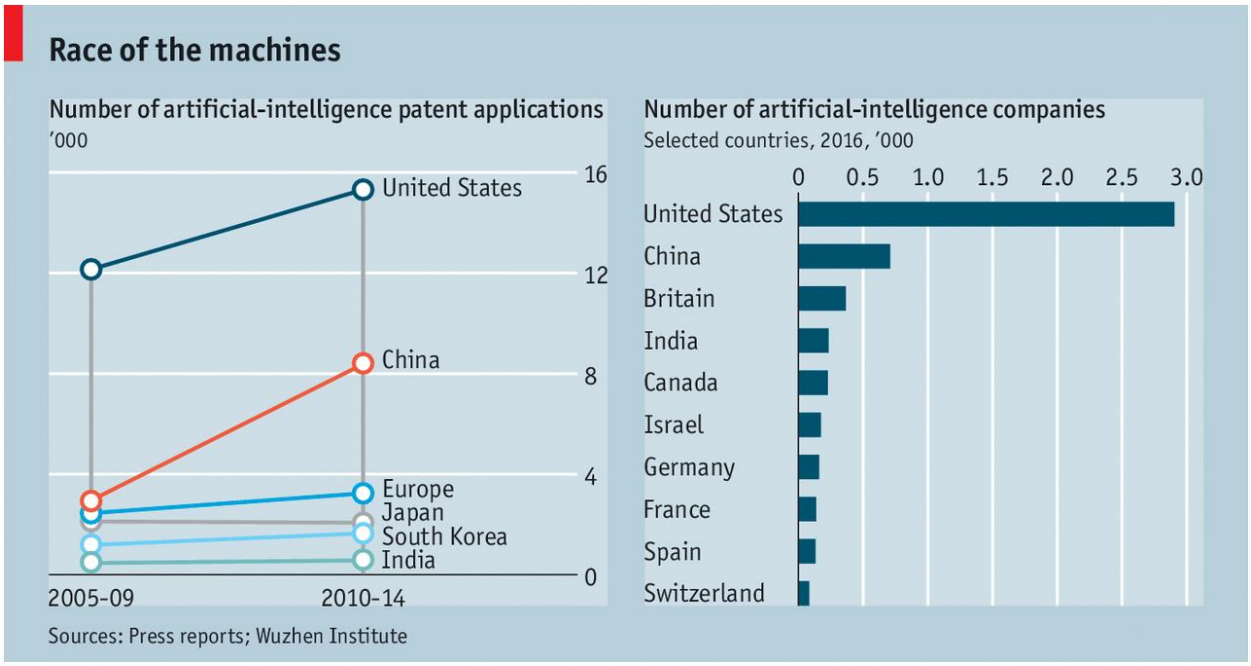
The world's first trillionaire will be an artificial intelligence entrepreneur— Marc Cuban [Source: CNBC]

With such a pronounced difference between the first and the second, the saying "the second is the first loser" reaches a completely new level of severity.

This sounds all very serious. So where is Germany in the picture? Looking at GDP, Germany is the 4th largest economy (after the US,China, and Japan) and ranks 3rd in exports after China and the US. From howmuch.net:



So one would naturally expect that Germany has a significant interest (and the necessary resources) to strive for leadership in AI, at least in specific domains—a domain that immediately comes to mind is manufacturing. However a recent article on economist.com draws a different picture, showing that Germany is significantly falling behind not just the US and China, but also several other countries!



A systemic issue:

It is not hard to understand why Germany's presence in AI is not as strong as expected. Germany never had a startup culture or a strong tech culture and its large companies are too big to be innovative. Moreover, it seems that the mind set is not aligned with innovation: Failure is **not** an option and in fact constitutes life-long stigma whereas in the US a culture of *fast failure* promotes innovation. On top of that, due to labor laws and company structure several large companies suffer

from *legacy leadership* that missed the point of adaptation. I will use the German automotive industry, one of the main, dominating industries, to exemplify the challenges. The German automotive industry focuses on engineering. However, the next evolution in the automotive space, autonomous cars are (mostly) a *software problem* not an engineering problem—we know how to build a great car (engineering problem: solved) but we do not know how to teach it to drive by itself (software or AI problem: unsolved) and in fact highly innovative companies in that space are not traditional car companies (e.g., Apple, nuTonomy, Intel, waymo, Uber). As a function of this basic mindset fallacy and strong resistance to change, German automotive companies have fallen behind now scrambling in a desperate attempt to play catch-up. Competing for top talent (not paying enough: robotics grad pay packages run at $200k/year in Pittsburgh) and technology (not researching enough: the US has a large number of AV startups). Surely, the German giants are still building great cars but their role might be reduced to become "Zulieferer" (Suppliers) for AI companies that build the autonomous technology and/or vehicle. In fact several companies have shifted to providing retro-fit kits for cars and for example in

Tempe, AZ (or Pittsburgh or San Francisco for that matter) you will regularly see retro-fit, autonomous Volvo's on the street functioning as autonomous Uber vehicles collecting important driving data for refinement and training (not always without challenges). Germany banned Uber and tests with autonomous vehicles have been only recently (05/2017) allowed in Germany... The next challenge after *the car* is the *fleet problem*: how to effectively manage a fleet of autonomous vehicles and what are models of use? Same story all over!

Mittelstand is not enough and Legacy Leadership:

A typical argument that is often brought forward is that Germany's strong *Mittelstand* (think of, often family-owned, SMEs with workforce supplied by Germany's apprentice system) will fix the issue by delivering the necessary innovation. In the past this has often been the case and in fact large parts of automotive suppliers come from that space. This was easy: process fine-tuning and kaizen-style hardening of manufacturing. However, when it comes to AI a much more concerted effort with

pooling of top talent is needed. Critical mass is required and smart people bumping into each other, not an R&D function with a “small r” and a “huge D” where innovation is an artifact that has been replaced by most precise, process-engineered, rehashing of Silicon Valley ideas. Some of these heavy weights literally have “being a fast follower” as strategy-mantra: “you go first” as innovation strategy…

The above is exacerbated by German labor laws that force legacy leadership on companies. Employees grow through the ranks, rather than hiring leadership from the outside and at the end of this growth process managers are out-of-touch with current technological advances and innovation. In the US, companies strongly fight this by aggressive firing and (re-)hiring to ensure a dynamic mix. Strong labor laws in Germany though cement legacy leadership: the longer an employee is with a company the harder it is to get rid of the employee and this is almost orthogonal to the actual performance of said employee. Now, these managers with a legacy mindset “hire what they know” and what does not threaten them, their position, or their worldview, effectively leading to

companies that try to solve today's challenges with tools from the 60s…

University-driven public AI research:

The cynical reader might now point towards Japan and their 会社 (Kaishas), these huge conglomerates such as Sony, Toyota, etc. that meticulously groom their employees through years of training within the company. While this is indeed a very similar approach to in-company-growing, the major difference is that the Japanese government and industry interest groups strongly invest in innovation and AI through large-scale initiatives, such as RIKEN-AIP; Japan also has a Google Lunar XPRIZE team…

But everything is going well and the world is jealously looking at Germany because of its strong exports. So how can it be so bad?

These exports and productivity stem from products and services that where created with the aforementioned legacy model at the end of the current S-curve: cars and other relatively pure engineering products. However little has been done at the beginning of the next S-curve with some

notable exceptions such as e.g., KUKA… oh wait, actually they recently got bought by China's Midea (see here and here)… my bad…

What is required is some concerted effort of the German government to strengthen Germany's position in Machine Learning, Artificial Intelligence, and more broadly data-driven methods. Germany did an outstanding job in establishing itself as one of the leaders in *optimization and operations research* with extremely strong groups spread throughout Germany and reasonable industry uptake of these methods. This made and makes a lot of sense as it is aligned with the need for improving manufacturing processes and more generally engineering excellence. The next evolution of these methods however requires a tight integration with AI methods and data and Germany is lacking severly behind (with few exceptions). While in several other countries talent is formed, honed, and groomed through data competitions and a huge number of startups in that space that bite their teeth out on challenging data and prediction problems, defining and redefining best practices at a mind-boggling pace, German companies and data innovation are crippled by German/European privacy laws. To be

clear, I am not against *reasonable* privacy laws and these are only one piece to the puzzle, but surely the current ones in place are not accelerating innovation.

Don't hate the player — hate the game:

In the end, there is a very simple proxy when it comes to innovation in the digital realm (including AI and more broadly the data space): the US has Google, Facebook, Amazon, Apple, Uber, Russia has Yandex, VK.com, China has Tencent, Alibaba, Baidu, DJI, … and Germany? Cheating car manufacturers is the first thing that I regularly hear. I guess after all, it really is about "being first, being smart, or cheat" and since "first" and "smart" were already taken what else is left? The fallout from innovating-through-cheating has not only severely damaged the "Made in Germany" brand but also sends an even stronger message: the German innovation concept has completed its life cycle and is end-of-life and Germany needs a new *mindset* to be competitive in the AI, ML, and data space.

EXECUTIVE SUMMARY

Artificial intelligence (AI) technology is rapidly proliferating around the world. Startling developments keep emerging, from the onset of deepfake videos that blur the line between truth and falsehood, to advanced algorithms that can beat the best players in the world in multiplayer poker. Businesses harness AI capabilities to improve analytic processing; city officials tap AI to monitor traffic congestion and oversee smart energy metering. Yet a growing number of states are deploying advanced AI surveillance tools to monitor, track, and surveil citizens to accomplish a range of policy objectives—some lawful, others that violate human rights, and many of which fall into a murky middle ground.

In order to appropriately address the effects of this technology, it is important to first understand where these tools are being deployed and how they are being used. Unfortunately, such information is scarce. To provide greater clarity, this paper presents an AI Global Surveillance (AIGS) Index—representing one of the first research efforts of its kind. The index compiles empirical data on AI surveillance use for 176 countries around the world. It does not distinguish between legitimate and unlawful uses of AI surveillance. Rather, the purpose of the research is to show how new surveillance capabilities are transforming the ability of governments to monitor and track individuals or systems. It specifically asks:

- Which countries are adopting AI surveillance technology?

- What specific types of AI surveillance are governments deploying?
- Which countries and companies are supplying this technology?

KEY FINDINGS

- AI surveillance technology is spreading at a faster rate to a wider range of countries than experts have commonly understood. At least seventy-five out of 176 countries globally are actively using AI technologies for surveillance purposes. This includes: smart city/safe city platforms (fifty-six countries), facial recognition systems (sixty-four countries), and smart policing (fifty-two countries).
- China is a major driver of AI surveillance worldwide. Technology linked to Chinese companies—particularly Huawei, Hikvision, Dahua, and ZTE—supply AI surveillance technology in sixty-three countries, thirty-six of which have signed onto China's Belt and Road Initiative (BRI). Huawei alone is responsible for providing AI surveillance technology to at least fifty countries worldwide. No other company comes close. The next largest non-Chinese supplier of AI surveillance tech is Japan's NEC Corporation (fourteen countries).
- Chinese product pitches are often accompanied by soft loans to encourage governments to purchase

their equipment. These tactics are particularly relevant in countries like Kenya, Laos, Mongolia, Uganda, and Uzbekistan—which otherwise might not access this technology. This raises troubling questions about the extent to which the Chinese government is subsidizing the purchase of advanced repressive technology.

- But China is not the only country supplying advanced surveillance tech worldwide. U.S. companies are also active in this space. AI surveillance technology supplied by U.S. firms is present in thirty-two countries. The most significant U.S. companies are IBM (eleven countries), Palantir (nine countries), and Cisco (six countries). Other companies based in liberal democracies—France, Germany, Israel, Japan—are also playing important roles in proliferating this technology. Democracies are not taking adequate steps to monitor and control the spread of sophisticated technologies linked to a range of violations.

- Liberal democracies are major users of AI surveillance. The index shows that 51 percent of advanced democracies deploy AI surveillance systems. In contrast, 37 percent of closed autocratic states, 41 percent of electoral autocratic/competitive autocratic states, and 41 percent of electoral democracies/illiberal democracies deploy AI surveillance technology.[1] Governments in full democracies are

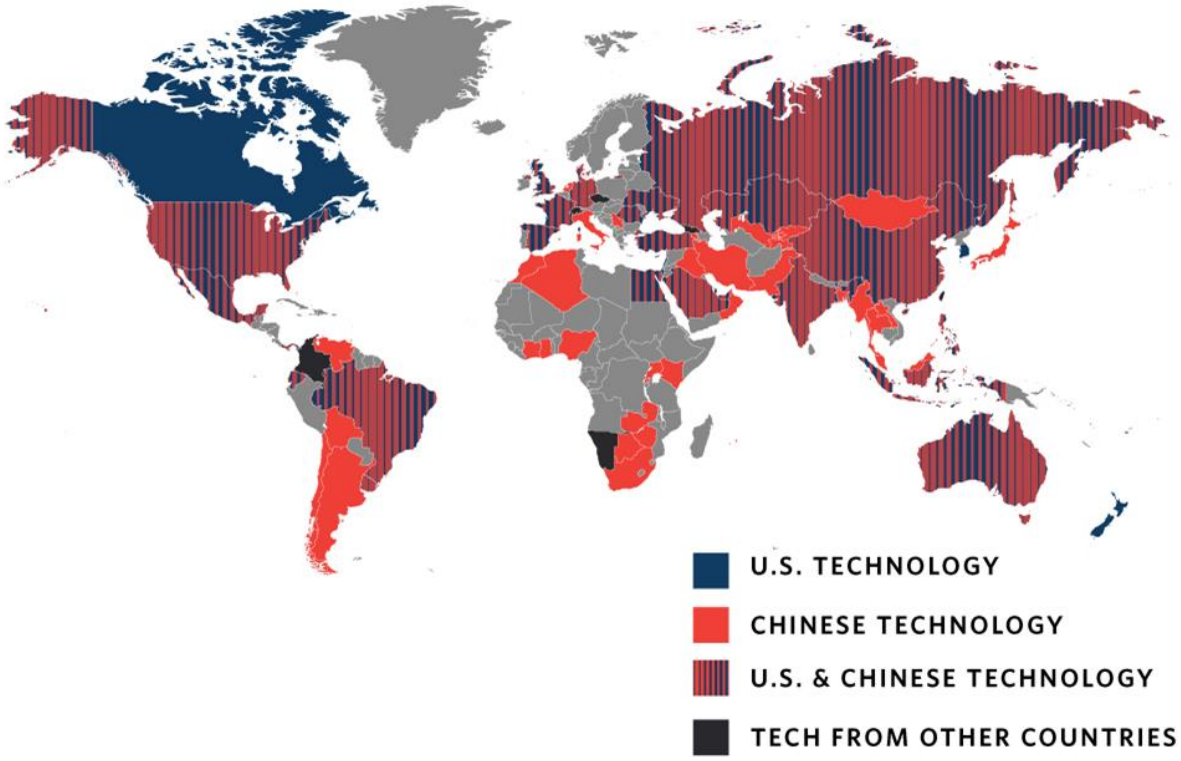
deploying a range of surveillance technology, from safe city platforms to facial recognition cameras. This does not inevitably mean that democracies are abusing these systems. The most important factor determining whether governments will deploy this technology for repressive purposes is the quality of their governance.

- Governments in autocratic and semi-autocratic countries are more prone to abuse AI surveillance than governments in liberal democracies. Some autocratic governments—for example, China, Russia, Saudi Arabia—are exploiting AI technology for mass surveillance purposes. Other governments with dismal human rights records are exploiting AI surveillance in more limited ways to reinforce repression. Yet all political contexts run the risk of unlawfully exploiting AI surveillance technology to obtain certain political objectives.

- There is a strong relationship between a country's military expenditures and a government's use of AI surveillance systems: forty of the world's top fifty military spending countries (based on cumulative military expenditures) also use AI surveillance technology.[2]
- The "Freedom on the Net 2018" report identified eighteen countries out of sixty-five that had accessed AI surveillance technology developed by Chinese companies.[3] The AIGS Index shows that the number of those countries accessing Chinese

AI surveillance technology has risen to forty-seven out of sixty-five countries in 2019.

MAP 1
AI Surveillance Technology Origin



NOTES

- The AIGS Index presents a country-by-country snapshot of AI tech surveillance with the majority of sources falling between 2017 and 2019. Given the opacity of government surveillance use, it is nearly impossible to pin down by specific year which AI platforms or systems are currently in use.
- The AIGS Index uses the same list of independent states included in the Varieties of Democracy (V-Dem) project with two exceptions, totaling

176.[4] The V-Dem country list includes all independent polities worldwide but excludes microstates with populations below 250,000.

- The AIGS Index does not present a complete list of AI surveillance companies operating in particular countries. The paper uses open source reporting and content analysis to derive its findings. Accordingly, there are certain built-in limitations. Some companies, such as Huawei, may have an incentive to highlight new capabilities in this field. Other companies have opted to downplay their association with surveillance technology and have purposely kept documents out of the public domain.

A full version of the index can be accessed online here: https://carnegieendowment.org/files/AI_Global_Surveillance_Index1.pdf. An interactive map keyed to the index that visually depicts the global spread of AI surveillance technology can be accessed here: https://carnegieendowment.org/publications/interactive/ai-surveillance.
All reference source material used to build the index has been compiled into an open Zotero library. It is available here: https://www.zotero.org/groups/2347403/global_ai_surveillance/items.

AI technology was once relegated to the world of science fiction, but today it surrounds us. It powers our smartphones, curates our music preferences, and guides our social media feeds. Perhaps the most notable aspect of AI is its sudden ubiquity.

In general terms, the goal of artificial intelligence is to "make machines intelligent" by automating or replicating behavior that "enables an entity to function appropriately and with foresight in its environment," according to computer scientist Nils Nilsson.[5] AI is not one specific technology. Instead, it is more accurate to think of AI as an integrated system that incorporates information acquisition objectives, logical reasoning principles, and self-correction capacities. An important AI subfield is machine learning, which is a statistical process that analyzes a large amount of information in order to discern a pattern to explain the current data and predict future uses.[6] Several breakthroughs are making new achievements in the field possible: the maturation of machine learning and the onset of deep learning; cloud computing and online data gathering; a new generation of advanced microchips and computer hardware; improved performance of complex algorithms; and market-driven incentives for new uses of AI technology.[7] Unsurprisingly, AI's impact extends well beyond individual consumer choices. It is starting to transform basic patterns of governance, not only by providing governments with unprecedented capabilities to monitor their citizens and shape their choices but also by giving them new capacity to disrupt elections, elevate false information, and delegitimize democratic discourse across borders.

Steven Feldstein

Steven Feldstein is a nonresident fellow in Carnegie's Democracy, Conflict, and Governance Program, where he focuses on issues of democracy, technology, human rights, U.S. foreign policy, conflict trends, and Africa.

The focus of this paper is on AI surveillance and the specific ways governments are harnessing a multitude of tools—from facial recognition systems and big data platforms to predictive policing algorithms—to advance their political goals. Crucially, the index does not distinguish between AI surveillance used for legitimate purposes and unlawful digital surveillance. Rather, the purpose of the research is to shine a light on new surveillance capabilities that are transforming the ability of states—from autocracies to advanced democracies—to keep watch on individuals.

AIGS INDEX—METHODOLOGY

The AIGS Index provides a detailed empirical picture of global AI surveillance trends and describes how governments worldwide are using this technology. It addresses three primary questions:

- Which countries are adopting AI surveillance technology?
- What specific types of AI surveillance are governments deploying?
- Which countries and companies are supplying this technology?

The AIGS Index is contained in Appendix 1. It includes detailed information for seventy-five countries where research indicates governments are deploying AI surveillance technology. The index breaks down AI surveillance tools into the following subcategories: 1) smart city/safe city, 2) facial recognition systems, and 3) smart policing. A full version of the index can be accessed online at https://carnegieendowment.org/files/AI_Global_Surveillance_Index1.pdf. An interactive map keyed to the index that visually depicts the global spread of AI surveillance technology can be accessed at https://carnegieendowment.org/publications/interactive/ai-surveillance.

All reference source material used to build the index has been compiled into an open Zotero library. It is available at https://www.zotero.org/groups/2347403/global_ai_surveillance/items.

The majority of sources referenced by the index occur between 2017 and 2019. A small number of sources date as far back as 2012. The index uses the same list of countries found in the Varieties of Democracy (V-Dem) project with two minor exceptions.[8] The V-Dem country list includes all independent polities worldwide but excludes microstates with populations below 250,000. The research collection effort combed through open-source material, country by country, in English and other languages, including news articles, websites, corporate documents, academic articles, NGO reports, expert submissions, and other public sources. It relied on systematic content analysis for each country incorporating multiple sources to determine the presence of relevant AI surveillance technology and corresponding companies. Sources were

categorized into tiered levels of reliability and accuracy. First-tier sources include major print and news magazine outlets (such as the New York Times, Economist, Financial Times, and Wall Street Journal). Second-tier sources include major national media outlets. Third-tier sources include web articles, blog posts, and other less substantiated sourcing; these were only included after multiple corroboration.
Given limited resources and staffing constraints (one full-time researcher plus volunteer research assistance), the index is only able to offer a snapshot of AI surveillance levels in a given country. It does not provide a comprehensive assessment of all relevant technology, government surveillance uses, and applicable companies. Because research relied primarily on content analysis and literature reviews to derive its findings, there are certain built-in limitations. Some companies, such as Huawei, may have an incentive to highlight new capabilities in this field. Other companies may wish to downplay links to surveillance technology and purposely keep documents out of the public domain.

Field-based research involving on-the-ground information collection and verification would be useful to undertake. A number of countries—such as Angola, Azerbaijan, Belarus, Hungary, Peru, Sri Lanka, Tunisia, and Turkmenistan—provided circumstantial or anecdotal evidence of AI surveillance, but not enough verifiable data to warrant inclusion in the index.

A major difficulty was determining which AI technologies should be included in the index. AI technologies that directly support surveillance objectives—smart city/safe city platforms, facial recognition systems, smart policing systems—are included

in the index. Enabling technologies that are critical to AI functioning but not directly responsible for surveillance programs are not included in the index.

Another data collection challenge is that governments (and many companies) purposely hide their surveillance capabilities. As such, it is difficult to precisely determine the extent to which states are deploying algorithms to support their surveillance objectives, or whether AI use is more speculative than real.

The index does not differentiate between governments that expansively deploy AI surveillance techniques versus those that use AI surveillance to a much lesser degree (for example, the index does not include a standardized interval scale correlating to levels of AI surveillance). This is by design. Because this is a nascent field and there is scant information about how different countries are using AI surveillance techniques, attempting to score a country's relative use of AI surveillance would introduce a significant level of researcher bias. Instead, a basic variable was used: is there documented presence of AI surveillance in a given country? If so, what types of AI surveillance technology is the state deploying? Future research may be able to assess and analyze levels of AI surveillance on a cross-comparative basis.

Finally, instances of AI surveillance documented in the index are not specifically tied to harmful outcomes. The index does not differentiate between unlawful and legitimate surveillance. In part, this is because it is exceedingly difficult to determine what specifically governments are doing in the surveillance realm and what the associated impacts are; there is too much that is unknown and hidden.

The findings indicate that at least seventy-five out of 176 countries globally are actively using AI technologies for surveillance purposes. This includes: smart city/safe city platforms (fifty-six countries), facial recognition systems (sixty-four countries), and smart policing (fifty-two countries). Three key insights emerge from the AIGS Index's findings.

First, global adoption of AI surveillance is increasing at a rapid pace around the world. Seventy-five countries, representing 43 percent of total countries assessed, are deploying AI-powered surveillance in both lawful and unlawful ways. The pool of countries is heterogeneous—they come from all regions, and their political systems range from closed autocracies to advanced democracies. The "Freedom on the Net 2018" report raised eyebrows when it reported that eighteen out of sixty-five assessed countries were using AI surveillance technology from Chinese companies.[9] The report's assessment period ran from June 1, 2017 to May 31, 2018. One year later, the AIGS Index finds that forty-seven countries out of that same group are now deploying AI surveillance technology from China. Unsurprisingly, countries with authoritarian systems and low levels of political rights are investing heavily in AI surveillance techniques. Many governments in the Gulf, East Asia, and South/Central Asia are procuring advanced analytic systems, facial recognition cameras, and sophisticated monitoring capabilities. But liberal democracies in Europe are also racing ahead to install automated border controls, predictive policing, safe cities, and facial recognition systems. In fact, it is striking how many safe city surveillance case studies posted on Huawei's

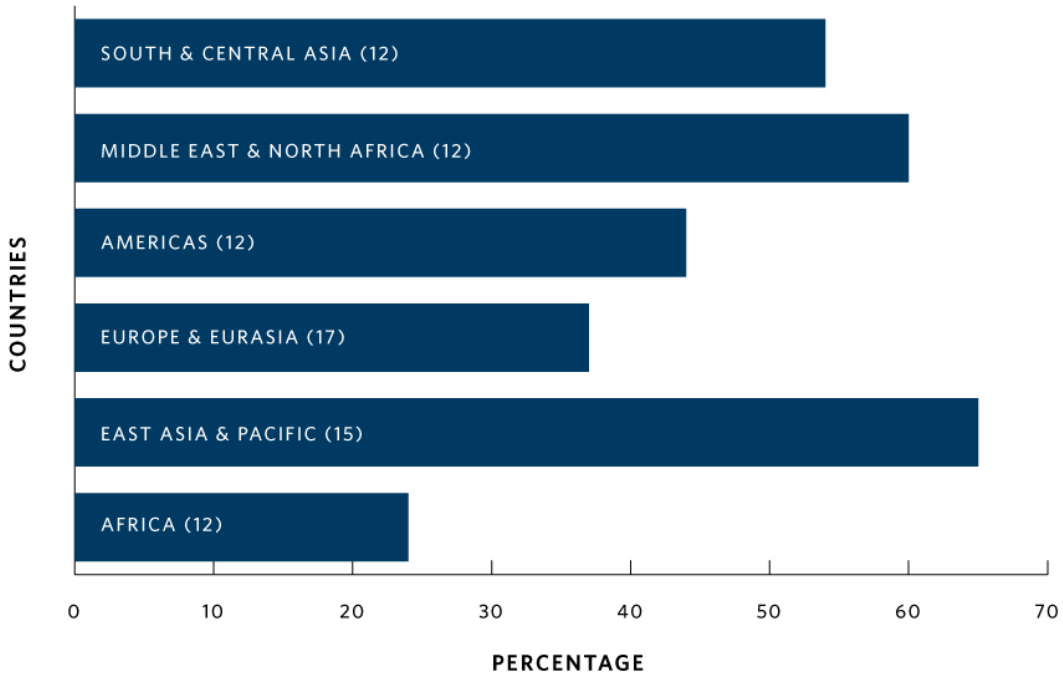
website relate to municipalities in Germany, Italy, the Netherlands, and Spain.

Regionally, there are clear disparities. The East Asia/Pacific and the Middle East/North Africa regions are robust adopters of these tools. South and Central Asia and the Americas also demonstrate sizable take-up of AI surveillance instruments. Sub-Saharan Africa is a laggard—less than one-quarter of its countries are invested in AI surveillance. Most likely this is due to technological underdevelopment (African countries are struggling to extend broadband access to their populations; the region has eighteen of twenty countries with the lowest levels of internet penetration).[10]

Given the aggressiveness of Chinese companies to penetrate African markets via BRI, these numbers will likely rise in the coming years. Figure 1 shows the percentage breakdown by region of countries adopting AI surveillance.

FIGURE 1

Percentage of Countries by Region Adopting AI Surveillance

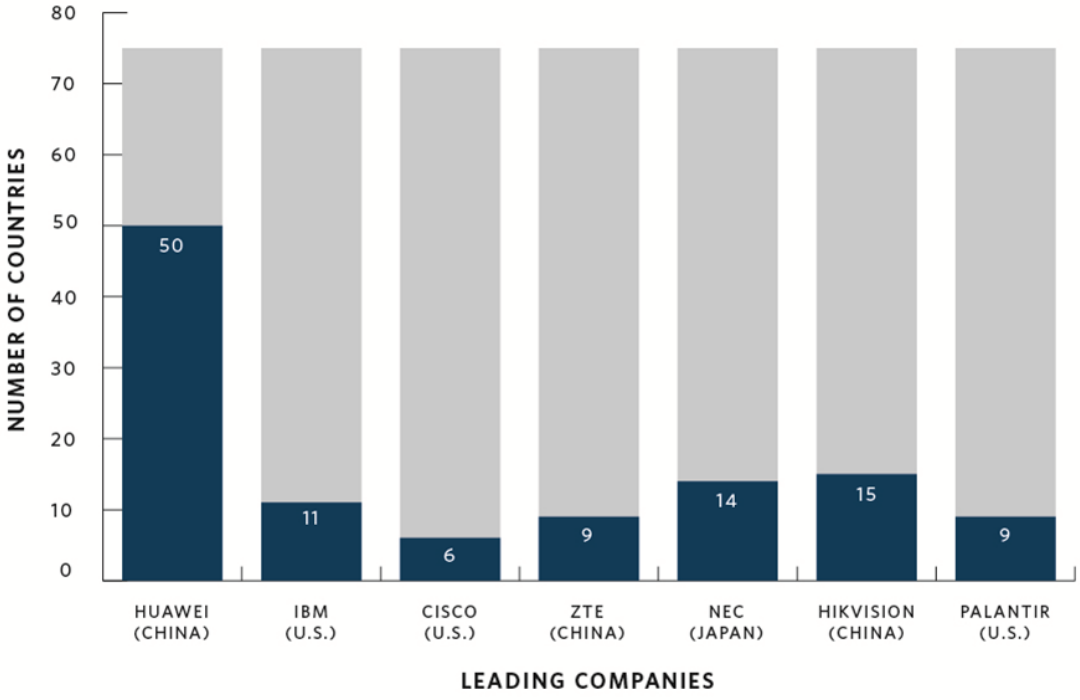


NOTE: The numbers in parentheses indicate how many countries per region have adopted AI surveillance.

Second, China is a major supplier of AI surveillance. Technologies linked to Chinese companies are found in at least sixty-three countries worldwide. Huawei alone is responsible for providing AI surveillance technology to at least fifty countries. There is also considerable overlap between China's Belt and Road Initiative and AI surveillance—thirty-six out of eighty-six BRI countries also contain significant AI surveillance technology. However, China is not the only country supplying advanced surveillance technology. France, Germany, Japan, and the United States are also major players in this sector. U.S. companies, for example, have an active presence in thirty-two

countries. Figure 2 breaks down the leading companies in the sector.

FIGURE 2
Leading Companies Contributing to AI Surveillance



NOTE: The AIGS Index tracks seventy-five countries that employ AI surveillance. The numbers here reflect how many of those countries each company is present in.

Third, liberal democracies are major users of AI surveillance. The index shows that 51 percent of advanced democracies deploy AI surveillance systems. In contrast, 37 percent of closed autocratic states, 41 percent of electoral autocratic/competitive autocratic states, and 41 percent of electoral democracies/illiberal democracies deploy AI surveillance technology. Liberal democratic governments are aggressively using AI tools to police borders, apprehend potential criminals, monitor citizens for bad behavior, and pull out suspected terrorists from crowds. This doesn't necessarily mean that

democracies are using this technology unlawfully. The most important factor determining whether governments will exploit this technology for repressive purposes is the quality of their governance—is there an existing pattern of human rights violations? Are there strong rule of law traditions and independent institutions of accountability? That should provide a measure of reassurance for citizens residing in democratic states.

But advanced democracies are struggling to balance security interests with civil liberties protections. In the United States, increasing numbers of cities have adopted advanced surveillance systems. A 2016 investigation by Axios's Kim Hart revealed, for example, that the Baltimore police had secretly deployed aerial drones to carry out daily surveillance over the city's residents: "From a plane flying overhead, powerful cameras capture aerial images of the entire city. Photos are snapped every second, and the plane can be circling the city for up to 10 hours a day."[11] Baltimore's police also deployed facial recognition cameras to monitor and arrest protesters, particularly during 2018 riots in the city.[12] The ACLU condemned these techniques as the "technological equivalent of putting an ankle GPS [Global Positioning Service] monitor on every person in Baltimore."[13]

On the U.S.-Mexico border, an array of hi-tech companies also purvey advanced surveillance equipment. Israeli defense contractor Elbit Systems has built "dozens of towers in Arizona to spot people as far as 7.5 miles away," writes the Guardian's Olivia Solon. Its technology was first perfected in Israel from a contract to build a "smart fence" to separate Jerusalem from the West Bank. Another company, Anduril Industries, "has developed towers that feature a laser-enhanced camera, radar

and a communications system" that scans a two-mile radius to detect motion. Captured images "are analysed using artificial intelligence to pick out humans from wildlife and other moving objects."[14] It is unclear to what extent these surveillance deployments are covered in U.S. law, let alone whether these actions meet the necessity and proportionality standard.

The United States is not the only democracy embracing AI surveillance. In France, the port city of Marseille initiated a partnership with ZTE in 2016 to establish the Big Data of Public Tranquility project. The goal of the program is to reduce crime by establishing a vast public surveillance network featuring an intelligence operations center and nearly one thousand intelligent closed-circuit television (CCTV) cameras (the number will double by 2020). Local authorities trumpet that this system will make Marseille "the first 'safe city' of France and Europe."[15] Similarly, in 2017, Huawei "gifted" a showcase surveillance system to the northern French town of Valenciennes to demonstrate its safe city model. The package included upgraded high definition CCTV surveillance and an intelligent command center powered by algorithms to detect unusual movements and crowd formations.[16]

The fact that so many democracies—as well as autocracies—are taking up this technology means that regime type is a poor predictor for determining which countries will adopt AI surveillance.

A better predictor for whether a government will procure this technology is related to its military spending. A breakdown of military expenditures in 2018 shows that forty of the top fifty military spending countries also have AI surveillance technology.[17] These countries span from full democracies to dictatorial regimes (and everything in between). They comprise

leading economies like France, Germany, Japan, and South Korea, and poorer states like Pakistan and Oman. This finding is not altogether unexpected; countries with substantial investments in their militaries tend to have higher economic and technological capacities as well as specific threats of concern. If a country takes its security seriously and is willing to invest considerable resources in maintaining robust military-security capabilities, then it should come as little surprise that the country will seek the latest AI tools. The motivations for why European democracies acquire AI surveillance (controlling migration, tracking terrorist threats) may differ from Egypt or Kazakhstan's interests (keeping a lid on internal dissent, cracking down on activist movements before they reach critical mass), but the instruments are remarkably similar. Future research might examine country-level internal security figures and compare them to levels of AI surveillance.

DISTINGUISHING BETWEEN LEGITIMATE AND UNLAWFUL SURVEILLANCE

State surveillance is not inherently unlawful. Governments have legitimate reasons to undertake surveillance that is not rooted in a desire to enforce political repression and limit individual freedoms. For example, tracking tools play a vital role in preventing terrorism. They help security forces deter bad acts and resolve problematic cases. They give authorities the ability to monitor critical threats and react accordingly. But technology has changed the nature of how governments carry out surveillance and what they choose to monitor. The internet has proliferated the amount of transactional data or "metadata"

available about individuals, such as information about sent and received emails, location identification, web-tracking, and other online activities. As former UN special rapporteur Frank La Rue noted in a milestone 2013 surveillance report:

> Communications data are storable, accessible and searchable, and their disclosure to and use by State authorities are largely unregulated. Analysis of this data can be both highly revelatory and invasive, particularly when data is combined and aggregated. As such, States are increasingly drawing on communications data to support law enforcement or national security investigations. States are also compelling the preservation and retention of communication data to enable them to conduct historical surveillance.[18]

It goes without saying that such intrusions profoundly affect an individual's right to privacy—to not be subjected to what the Office of the UN High Commissioner for Human Rights (OHCHR) called "arbitrary or unlawful interference with his or her privacy, family, home or correspondence."[19] Surveillance likewise may infringe upon an individual's right to freedom of association and expression. Under international human rights law, three principles are critical to assessing the lawfulness of a particular surveillance action.

First, does domestic law allow for surveillance? La Rue's successor, David Kaye, issued a report in 2019 that affirmed that legal regulations should be "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly and it must be made accessible to the public." Legal requirements should not be "vague or overbroad," which would allow unconstrained discretion to government officials. The

legal framework itself should be "publicly accessible, clear, precise, comprehensive and non-discriminatory."[20]
Second, does the surveillance action meet the "necessity and proportionality" international legal standard, which restricts surveillance to situations that are "strictly and demonstrably necessary to achieve a legitimate aim"?[21]
Third, are the interests justifying the surveillance action legitimate? Disagreements abound when it comes to determining what constitutes legitimate surveillance and what is an abuse of power. While governments commonly justify surveillance on national security or public order grounds, the OHCHR warns that such restrictions may "unjustifiably or arbitrarily" restrict citizens' rights to freedom of opinion and expression. It contends that legitimate surveillance requires states to "demonstrate the risk that specific expression poses to a definite interest in national security or public order," and that a "robust, independent oversight system" that entrusts judiciaries to authorize relevant surveillance measures and provide remedies in cases of abuse is required.[22] Kaye adds that legitimate surveillance should only apply when the interest of a "whole nation is at stake," and should exclude surveillance carried out "in the sole interest of a Government, regime or power group."[23]
The legal standards required to legitimately carry out surveillance are high, and governments struggle to meet them. Even democracies with strong rule of law traditions and robust oversight institutions frequently fail to adequately protect individual rights in their surveillance programs. Countries with weak legal enforcement or authoritarian systems "routinely shirk these obligations."[24] As the OHCHR's inaugural report on privacy in the digital age concludes, states with "a lack of

adequate national legislation and/or enforcement, weak procedural safeguards and ineffective oversight" bring reduced accountability and heightened conditions for unlawful digital surveillance.[25]

AI surveillance exacerbates these conditions and makes it likelier that democratic and authoritarian governments may carry out surveillance that contravenes international human rights standards. Frank La Rue explains: "Technological advancements mean that the State's effectiveness in conducting surveillance is no longer limited by scale or duration. Declining costs of technology and data storage have eradicated financial or practical disincentives to conducting surveillance. As such, the State now has a greater capability to conduct simultaneous, invasive, targeted and broad-scale surveillance than ever before."[26]

AI surveillance in particular offers governments two major capabilities. One, AI surveillance allows regimes to automate many tracking and monitoring functions formerly delegated to human operators. This brings cost efficiencies, decreases reliance on security forces, and overrides potential principal-agent loyalty problems (where the very forces operating at the behest of the regime decide to seize power for themselves).

Two, AI technology can cast a much wider surveillance net than traditional methods. Unlike human operatives "with limited reserves of time and attention," AI systems never tire or fatigue.[27] As a result, this creates a substantial "chilling effect" even without resorting to physical violence; citizens never know if an automated bot is monitoring their text messages, reading their social media posts, or geotracking their movements around town.[28]

This paper recognizes that AI surveillance technology is "value neutral." In and of themselves, these tools do not foment repression, and their presence does not mean that a government is using them for antidemocratic purposes. The index does not specify, country-by-country, whether these instruments are being used by governments in lawful or illegitimate manners. Rather, the purpose of the index is to identify which countries possess sufficiently advanced tools that allow them to pursue a range of surveillance objectives.

HOW MUCH IS CHINA DRIVING THE SPREAD OF AI SURVEILLANCE?

Empirically, the AIGS Index shows that Chinese companies—led by Huawei—are leading suppliers of AI surveillance around the world. Overall, China is making a sustained push for leadership and primacy in AI.[29] A growing consensus singles out China as a global driver of "authoritarian tech." Experts claim that Chinese companies are working directly with Chinese state authorities to export "authoritarian tech" to like-minded governments in order to spread influence and promote an alternative governance model.[30] But is this accurate?

There is some truth to this argument—a subset of Chinese exports goes directly to countries like Zimbabwe and Venezuela that are gross human rights violators and which would otherwise be unable to access such technology. But AI surveillance is not solely going from one authoritarian country (China) to other authoritarian states. Rather, transfers are happening in a much more heterogeneous fashion. China is exporting surveillance tech to liberal democracies as much as it

is targeting authoritarian markets. Likewise, companies based in liberal democracies (for example, Germany, France, Israel, Japan, South Korea, the UK, the United States) are actively selling sophisticated equipment to unsavory regimes.

Saudi Arabia is a good case in point. Huawei is helping the government build safe cities, but Google is establishing cloud servers, UK arms manufacturer BAE has sold mass surveillance systems, NEC is vending facial recognition cameras, and Amazon and Alibaba both have cloud computing centers in Saudi Arabia and may support a major smart city project.[31] The index shows that repressive countries rarely procure such technology from a single source. In Thailand, government officials repeatedly emphasized the importance of "foreign policy balancing" and not affiliating too strongly with any one side: "Always been that way. That's why we're still a kingdom. We compromise, we negotiate, and we balance."[32]
That being said, there are special reasons why experts are applying greater scrutiny to Chinese companies. Huawei is the leading vendor of advanced surveillance systems worldwide by a huge factor. Its technology is linked to more countries in the index than any other company. It is aggressively seeking new markets in regions like sub-Saharan Africa. Huawei is not only providing advanced equipment but also offering ongoing technological support to set up, operate, and manage these systems.

A recent investigative report by the Wall Street Journal provides an eye-opening example. The reporters found that Huawei technicians in both Uganda and Zambia helped government officials spy on political opponents. This included "intercepting their encrypted communications and social media, and using

cell data to track their whereabouts." Not only did Huawei employees play a "direct role in government efforts to intercept the private communications of opponents," but they also encouraged Ugandan security officials to travel to Algeria so they could study Huawei's "intelligent video surveillance system" operating in Algiers.[33] Uganda subsequently agreed to purchase a similar facial recognition surveillance system from Huawei costing $126 million.[34]

The Australian Strategic Policy Institute's project on Mapping China's Tech Giants indicates that Huawei is responsible for seventy-five "smart city-public security projects," and has seen a colossal increase in its business line: "In 2017, Huawei listed 40 countries where its smart-city technologies had been introduced; in 2018, that reach had reportedly more than doubled to 90 countries (including 230 cities)."[35] Huawei is directly pitching the safe city model to national security agencies, and China's Exim Bank appears to be sweetening the deal with subsidized loans. The result is that a country like Mauritius obtains long-term financing from the Chinese government, which mandates contracting with Chinese firms.[36] The Mauritian government then turns to Huawei as the prime contractor or sub-awardee to set up the safe city and implement advanced surveillance controls.

It is also increasingly clear that firms such as Huawei operate with far less independence from the Chinese government than they claim. Huawei was founded in 1987 by Ren Zhengfei, a former officer in the People's Liberation Army who served in its "military technology division," Anna Fifield at the Washington Post has noted.[37] There are consistent reports that Huawei receives significant subsidies from the Chinese government.[38] There also appear to be strong connections

between Huawei's leadership and China's security and intelligence apparatus. Sun Yafang, for example, chairwoman of Huawei's board from 1999 to 2018, once worked in China's Ministry of State Security.[39] Max Chafkin and Joshua Brustein reported in Bloomberg Businessweek that there are allegations that Ren may have been a "high-ranking Chinese spymaster and indeed may still be."[40] Experts maintain that the Chinese Communist Party increasingly is establishing "party 'cells' in private companies to enable enhanced access and control.[41] Huawei has publicly averred that it would "definitely say no" to any demands by the Chinese government to hand over user data.[42] But this contravenes a 2015 Chinese national security law that mandates companies to allow third-party access to their networks and to turn over source code or encryption keys upon request.[43] Huawei's declared ownership structure is remarkably opaque. A recent academic study by Christopher Balding and Donald C. Clarke concluded that 99 percent of Huawei shares are controlled by a "trade union committee," which in all likelihood is a proxy for Chinese state control of the company."[44]

Even if Chinese companies are making a greater push to sell advanced surveillance tech, the issue of intentionality remains perplexing—to what extent are Chinese firms like Huawei and ZTE operating out of their own economic self-interest when peddling surveillance technology versus carrying out the bidding of the Chinese state? At least in Thailand, recent research interviews did not turn up indications that Chinese companies are pushing a concerted agenda to peddle advanced AI surveillance equipment or encourage the government to build sophisticated monitoring systems. An official from Thailand's Ministry of Interior noted that while AI technology is

"out there" and something the government is thinking more about, "China hasn't offered any AI. It doesn't give AI—Thais have to ask."[45] The smart city/safe city model also garnered skepticism. Somkiat Tangkitvanich, a leading technology expert in Thailand, commented, "the idea of a smart city is a joke." He relayed a recent conversation he had with Thailand's information and communications technologies (ICT) minister: "He [the minister] boasted about the smart city in Phuket. . . . He told me that we are thinking about giving wristbands to tourists so that we can track them, we can help them. Something like that. But it's not really implemented. Smart city in Phuket turns out to be providing free Wi-Fi and internet to tourists!"[46] This serves as a useful reminder that more on-the-ground research is needed to separate hyperbole from fact in this area.

TYPES OF AI SURVEILLANCE

The following sections will describe key AI surveillance techniques and how governments worldwide are deploying them to support specific policy objectives.

States use AI technology to accomplish a broad range of surveillance goals. This section details three primary AI surveillance tools incorporated in the AIGS Index: smart city/safe city platforms, facial recognition systems, and smart policing. It also describes enabling technologies—such as cloud computing and Internet of Things (IOT) networks—that are

integral for AI surveillance tools to function. Enabling technologies are not incorporated in the index.

Importantly, AI surveillance is not a standalone instrument of repression. It forms part of a suite of digital repression tools—information and communications technologies used to surveil, intimidate, coerce, and harass opponents in order to inflict a penalty on a target and deter specific activities or beliefs that challenge the state.[47] (See Appendix 2 for more information.) Table 1 summarizes each technique and its corresponding level of global deployment.

TABLE 1

Summary of AI Surveillance Techniques and Global Prevalence

AI Surveillance Technique	Description	Global Proliferation (out of 75 countries)
Smart Cities/Safe Cities	Cities with sensors that transmit real-time data to facilitate service delivery, city management, and public safety. Often referred to as "safe cities," they incorporate sensors, facial recognition cameras, and police body cameras connected to intelligent command centers to prevent crime, ensure public safety, and respond to emergencies. Only platforms with a clear public safety focus are incorporated in the index.	56 countries
Facial Recognition Systems	Biometric technology that uses cameras (still images or video) to match stored or live footage of individuals with images from databases. Not all systems focus on database matching; some systems assess aggregate demographic trends or conduct broader sentiment analysis via facial recognition crowd scanning.	64 countries
Smart Policing	Data-driven analytic technology used to facilitate investigations and police response; some systems incorporate algorithmic analysis to make predictions about future crimes.	53 countries

SMART CITIES/SAFE CITIES

The World Bank describes smart cities as "technology-intensive" urban centers featuring an array of sensors that gather information in real time from "thousands of

interconnected devices" in order to facilitate improved service delivery and city management.[48] They help municipal authorities manage traffic congestion, direct emergency vehicles to needed locations, foster sustainable energy use, and streamline administrative processes. But there is growing concern that smart cities are also enabling a dramatic increase in public surveillance and intrusive security capabilities. IBM, one of the original coiners of the term, designed a brain-like municipal model where information relevant to city operations could be centrally processed and analyzed.[49] A key component of IBM's smart city is public safety, which incorporates an array of sensors, tracking devices, and surveillance technology to increase police and security force capabilities.

Huawei has been up-front about trumpeting public safety technologies for smart cities. It is marketing "safe cities" to law enforcement communities to "predict, prevent, and reduce crime" and "address new and emerging threats."[50] In a 2016 white paper, Huawei describes a "suite of technology that includes video surveillance, emergent video communication, integrated incident command and control, big data, mobile, and secured public safety cloud" to support local law enforcement and policing as well as the justice and corrections system.[51] Huawei explicitly links its safe city technology to confronting regional security challenges, noting that in the Middle East, its platforms can prevent "extremism"; in Latin America, safe cities enable governments to reduce crime; and that in North America, its technology will help the United States advance "counterextremism" programs.[52]

How do these platforms work in practice to advance surveillance goals? The IT firm Gartner, which partners with Microsoft on smart cities, provides an example:

> Saudi Arabia's Makkah Region Development Authority (MRDA) created a crowd-control system to increase safety and security of Hajj pilgrims. Data is collected via a wristband embedding identity information, special healthcare requirements and a GPS. In addition, surveillance cameras are installed to collect and analyze real-time video along the Al Mashaaer Al Mugaddassah Metro Southern Line (MMMSL), as well as in the holy sites, such as Great Mosque of Mecca, Mount Arafat, Jamarat and Mina.[53]

Unsurprisingly, such systems lend themselves to improper use. Recently, Huawei's safe city project in Serbia, which intends to install 1,000 high-definition (HD) cameras with facial recognition and license plate recognition capabilities in 800 locations across Belgrade, sparked national outrage.[54] Huawei posted a case study (since removed) about the benefits of safe cities and described how similar surveillance technology had facilitated the apprehension of a Serbian hit-and-run perpetrator who had fled the country to a city in China: "Based on images provided by Serbian police, the . . . [local] Public Security Bureau made an arrest within three days using new technologies."[55] Rather than applaud the efficiency of the system, Serbian commentators observed that in a country racked by endemic corruption and encroaching authoritarianism, such technology offers a powerful tool for Serbian authorities to curb dissent and perpetrate abuses. Smart city platforms with a direct public security link are found in at least fifty-six of seventy-five countries with AI surveillance technology.

FACIAL RECOGNITION SYSTEMS

Facial recognition is a biometric technology that uses cameras—both video and still images—to match stored or live footage of individuals with images from a database. Not all facial recognition systems focus on individual identification via database matching. Some systems are designed to assess aggregate demographic trends or to conduct broader sentiment analysis via facial recognition crowd scanning.

Unlike ordinary CCTV, which has been a mainstay of police forces for twenty-five years, facial recognition cameras are much more intrusive. They can scan distinctive facial features in order to create detailed biometric maps of individuals without obtaining consent. Often facial recognition surveillance cameras are mobile and concealable. For example, security forces in Malaysia have entered into a partnership with the Chinese tech company Yitu to equip officers with facial recognition body cameras. This will allow security officials to "rapidly compare images caught by live body cameras with images from a central database."[56]

Huawei is a major purveyor of facial recognition video surveillance, particularly as part of its safe city platforms. It describes the technology's benefits in the Kenya Safe City project:

> As part of this project, Huawei deployed 1,800 HD cameras and 200 HD traffic surveillance systems across the country's capital city, Nairobi. A national police command center supporting over 9,000 police officers and 195 police stations was established to achieve monitoring and case-solving. The system worked during Pope Francis' visit to Kenya in 2015, where more than eight million people welcomed his arrival.

> With Huawei's HD video surveillance and a visualized integrated command solution, the efficiency of policing efforts as well as detention rates rose significantly.[57]

Experts detail several concerns associated with facial recognition.

First, few rules govern access to and the use of image databases (repositories that store captured images from facial recognition cameras). How governments use this information, how long images are stored, and where authorities obtain such images in the first place are opaque issues and vary by jurisdiction. Recent disclosures that U.S. law enforcement agencies (the Federal Bureau of Investigation and Immigration and Customs Enforcement) scanned through millions of photos in state driver's license databases without prior knowledge or consent come as little surprise. The vacuum of legal checks and balances has led to a "surveillance-first, ask-permission-later system," Drew Harrell noted in the Washington Post.[58]

Second, the accuracy of facial recognition technology varies significantly. Certain tests have disclosed unacceptably high false-match rates. A recent independent report of the UK's Metropolitan Police found that its facial recognition technology had an extraordinary error rate of 81 percent.[59] Similarly, Axon, a leading supplier of police body cameras in the United States, announced that it would cease offering facial recognition on its devices. Axon's independent ethics board stated: "Face recognition technology is not currently reliable enough to ethically justify its use."[60]

But other assessments demonstrate much more favorable results. Evaluations conducted between 2014 and 2018 of 127 algorithms from thirty-nine developers by the U.S. National Institute for Standards and Technology showed that "facial

recognition software got 20 times better at searching a database to find a matching photograph." The failure rate in the same period dropped from 4.0 percent to 0.2 percent.[61]

One reason for the discrepancy is that under ideal conditions, facial recognition can perform very well. But when unexpected variables are thrown in—poor weather or fuzzy database images—then failure rates start to shoot up. Facial recognition technology also has been unable to shake consistent gender and racial biases, which lead to elevated false positives for minorities and women—"the darker the skin, the more errors arise—up to nearly 35 percent for images of darker skinned women" noted Steve Lohr in the New York Times.[62]

Citizens are starting to fight back against facial recognition systems. Protesters in Hong Kong, for example, have covered up their faces and disabled their smartphone facial recognition logins to prevent law enforcement access. They have also turned the tables on the police by taking pictures of unbadged officers and using facial recognition image searching to expose the officers' identities online.[63]

Facial recognition systems are rapidly spreading around the world. The index identifies at least sixty-four countries that are actively incorporating facial recognition systems in their AI surveillance programs.

SMART POLICING

The idea behind smart policing is to feed immense quantities of data into an algorithm—geographic location, historic arrest levels, types of committed crimes, biometric data, social media feeds—in order to prevent crime, respond to criminal acts, or even to make predictions about future criminal activity. As Privacy International notes: "With the proliferation of

surveillance cameras, facial recognition, open source and social media intelligence, biometrics, and data emerging from smart cities, the police now have unprecedented access to massive amounts of data." Therefore, one major component to smart policing is to create automated platforms that can disaggregate immense amounts of material, facilitate data coming in from multiple sources, and permit fine-tuned collection of individual information.

One area that has received considerable recent attention is predictive policing. The technique accelerated in the United States after the National Institute of Justice started issuing grants for pilot predictive policing projects in 2009. At its core, these programs claim to predict with remarkable accuracy, based on massive data aggregation, where future crimes will be committed and which individuals are likely to commit those crimes. Predictive policing has exploded in popularity. The PredPol predictive analytics program, for example, is deployed "by more than 60 police departments around the country."[64] But there are growing concerns about algorithmic bias and prejudice, as well as the effectiveness of these predictions. Recent reporting by Caroline Haskins for Vice describes how PredPol's predictive crime forecasting algorithm operates. Predpol's software generates crime forecasts for police officers "on a scale as small as 500 by 500 square feet," which can pinpoint specific houses. It assumes that "certain crimes committed at a particular time are more likely to occur in the same place in the future."[65] PredPol reveals that "historical event datasets are used to train the algorithm for each new city (ideally 2 to 5 years of data). PredPol then updates the algorithm each day with new events as they are received from the department." New predictions are highlighted in special red

boxes superimposed on Google Maps representing high-risk areas that warrant special attention from police patrols.[66] A key shortcoming in PredPol's methodology is that it generates future predictions based on data from past criminal activity and arrests. Certain minority neighborhoods that have suffered from "overpolicing" and biased police conduct show up with higher frequency in PredPol's dashboard. This may not represent fine-tuned algorithmic crime prediction as much as it involves the perpetuation of structurally biased policing.

China has enthusiastically embraced predictive policing as part of its Xinjiang crackdown. Human Rights Watch reports on the creation of an Integrated Joint Operations Platform (IJOP), which collects data from CCTV cameras, facial recognition devices, and "wifi sniffers" (devices that eavesdrop on activities or communications within wireless networks). IJOP procures additional data from license plates and identification cards scanned at checkpoints, as well as health, banking, and legal records.[67] Chinese authorities are supplementing IJOP with mandatory DNA samples from all Xinjiang residents aged twelve to sixty-five.[68] This information is fed into IJOP computers, and algorithms sift through troves of data looking for threatening patterns. Once IJOP flags an individual, that person is picked up by security forces and detained for questioning.[69]

Smart policing techniques are used in at least fifty-three of seventy-five countries with AI surveillance.

A second category of technology is not directly responsible for supporting surveillance programs, but provides critical capabilities that are essential for implementing applications. Advanced video surveillance and facial recognition cameras could not function without cloud computing capabilities. As one expert put it, if video surveillance is the "eyes" then cloud services are the "brains" that "connect cameras and hardware to the cloud computing models via 5G networks."[70] However, cloud computing in isolation is not inherently oriented toward surveillance. Therefore, these secondary technologies are placed in an "enabling technologies" category and described below.[71]

They are not included in the AIGS Index.

TABLE 2
AI Surveillance—Enabling Technologies

AI Surveillance Technique	Description
Automated Border Control (ABC) Systems	Biometric systems powered by facial recognition to automatically control airport or border access; can also include biometric passports and IDs
Cloud Computing	Infrastructural components, networks, and sensors that enable AI processing and operations (for example, cloud servers, data centers, IOT networks)
Internet of Things	Devices connected via the internet that allow data to be shared for analytic processing in the cloud
Other AI Technology	Other relevant AI technologies such as digital government, AI training centers, and AI research institutes

AUTOMATED BORDER CONTROL SYSTEMS

These are found primarily in international airports and border crossings. According to the consulting firm Accenture, ABC

systems use “multi-model biometric matching”—facial image recognition combined with e-passports or other biometric documents—to process passengers.[72]

The process initiates when a passenger steps in front of a multi-camera wall. Digital mirrors located adjacent to the cameras attract passengers’ eyes for image capture. A risk assessment is then performed through automated testing of identities against an individual’s passport and certain security watch-lists.[73] Those who are not cleared by the automated system must go into secondary screening with human agents.
Governments are piloting new features, such as automated lie detection technology, in ABC systems. For example, the European Union is testing a technology called iBorderCtrl in three countries—Greece, Hungary, and Latvia—to screen migrants at border crossings. Individuals are asked questions about their countries of origin and circumstances of departure. The answers are then evaluated by an AI-based lie-detecting system.[74] Travelers found to have honestly answered questions are given a code allowing them to cross. All others are transferred to human border guards for additional questioning. The technology behind iBorderCtrl is based on “affect recognition science,” which purports to read facial expressions and infer emotional states in order to render legal judgments or policy decisions. Psychologists have widely criticized these tools, maintaining that it is difficult to rely on facial expressions alone to accurately determine a person’s state of mind.[75] Despite scientific skepticism about these techniques, governments continue to explore their use.

CLOUD COMPUTING

Governments and companies are increasingly storing data in massive off-site locations—known as the cloud—that are accessible through a network, usually the internet.[76] Cloud computing is a general use technology that includes everything from turn-by-turn GPS maps, social network and email communications, file storage, and streaming content access. The National Institute of Standards and Technology defines cloud computing as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[77] In basic terms, cloud computing data centers function as the backbone of the internet, instantly storing, communicating, and transporting produced information. As such, cloud computing is essential to effectively running AI systems. Microsoft, IBM, Amazon, Huawei, and Alibaba have all established these data centers to facilitate AI operations.

A growing number of countries have fully embraced cloud computing and outsourced all of their data storage needs to a single corporate platform. In 2018, for example, Iceland signed a service agreement with Microsoft for the company to be the

sole IT supplier for the country's entire public sector.[78] The cloud computing trend is not without problems. For one, cloud servers present enticing targets for cyber hackers. Security firms like NSO Group claim they are able to penetrate cloud servers and access a target's "location data, archived messages or photos," leading many to question whether cloud computing companies can keep personal information, corporate secrets, classified government material, or health records safe (however they generally represent a more secure method of storage than legacy on-site data storage facilities).[79] A related concern is forced data disclosures—even if cloud servers remain technically secure, governments may coerce companies into disclosing certain data (such as email communications or text messages of regime critics) held in the cloud.

INTERNET OF THINGS

The IOT is based on the reality that more and more devices will be connected to each other via the internet, allowing data to be shared for analytic processing in the cloud.[80] A major IOT hurdle is lack of interoperability between devices. At present, iPhones, Alexa speakers, Nest thermostats, and OnStar auto systems function from different platforms and use different information sources. The IOT's goal is to "help tame this Tower of Babel" and ensure device integration and data aggregation (although companies like Amazon, Apple, and Google are also setting up distinct ecosystems that only have limited interoperability with other platforms).[81] While the IOT will bring greater efficiencies, it may also transform traditional non-networked devices, such as smart speakers, into omnipresent surveillance instruments:

> The Internet of Things promises a new frontier for networking objects, machines, and environments in ways that we [are] just beginning to understand. When, say, a television has a microphone and a network connection, and is reprogrammable by its vendor, it could be used to listen in to one side of a telephone conversation taking place in its room—no matter how encrypted the telephone service itself might be. These forces are on a trajectory towards a future with more opportunities for surveillance.[82]

Controversy surrounding IOT technology is growing. In early 2019, Amazon disclosed that thousands of its workers listened to conversations recorded by Echo smart speakers. In some cases, its workers debated whether recordings of possible crimes should be turned over to law enforcement authorities.[83] Amazon analyzed these transcripts without the knowledge or consent of its customers. Similarly, it came to light that Google and Facebook contractors have been regularly listening to recordings between their platforms and individual consumers.[84]

IOT-powered mobile surveillance is another possibility for this class of technology. A new device was recently demonstrated that plugs into a Tesla Model S or Model 3 car and turns its built-in cameras "into a system that spots, tracks, and stores license plates and faces over time," journalist Andy Greenberg described. When the owner has parked the car, "it can track nearby faces to see which ones repeatedly appear." The purpose of the device is to warn car owners against thieves and vandals. But as the device's inventor Truman Kain acknowledges, "it turns your Tesla into an AI-powered surveillance station" and provides "another set of eyes, to help out and tell you it's seen a

license plate following you over multiple days, or even multiple turns of a single trip."[85]

CONCLUSION

The spread of AI surveillance continues unabated. Its use by repressive regimes to engineer crackdowns against targeted populations has already sounded alarm bells. But even in countries with strong rule of law traditions, AI gives rise to troublesome ethical questions. Experts express concerns about facial recognition error rates and heightened false positives for minority populations. The public is increasingly aware about algorithmic bias in AI training datasets and their prejudicial impact on predictive policing algorithms and other analytic tools used by law enforcement. Even benign IOT applications—smart speakers, remote keyless entry locks, automotive intelligent dash displays—may open troubling pathways for surveillance. Pilot technologies that states are testing on their borders—such as iBorderCtrl's affective recognition system—are expanding despite criticisms that they are based on faulty science and unsubstantiated research. The cumulative impact gives pause. Disquieting questions are surfacing regarding the accuracy, fairness, methodological consistency, and prejudicial impact of advanced surveillance technologies. Governments have an obligation to provide better answers and fuller transparency about how they will use these new intrusive tools.

The purpose of the index and working paper is to highlight emergent trends for a technology that is not well understood yet

will increasingly shape modern life. The good news is that there is ample time to initiate a much-needed public debate about the proper balance between AI technology, government surveillance, and the privacy rights of citizens. But as these technologies become more embedded in governance and politics, the window for change will narrow.

ACKNOWLEDGMENTS

Special thanks to Luke Lamey, undergraduate at Georgetown University, for research assistance in compiling references for the AI Global Surveillance Index. Many thanks also go to Jon Bateman (Carnegie Endowment for International Peace), Thomas Carothers (Carnegie Endowment for International Peace), Adrian Shabhaz (Freedom House), Brian Wampler (Boise State University), and Nick Wright (Intelligent Biology, Georgetown University) for generously giving their time to read through prior drafts of this paper and to offer invaluable feedback and advice.

CA: Closed Autocracy

EA: Electoral Autocracy

ED: Electoral Democracy

LD: Liberal Democracy

AF: Africa

EAP: East Asia & Pacific

EUR: Europe & Eurasia

MENA: Middle East & North Africa

SCA: South & Central Asia

WHA: Western Hemisphere

SOURCE NOTES

- **Aggregate Regime Score:** Comprised of evenly weighted average

of country scores from Freedom in the World 2019, https://freedomhouse.org/report/countries-world-freedom-2019; the EIU Democracy Index 2018, https://www.eiu.com/topic/democracy-index; and the V-Dem Dataset version 9, Electoral Democracy Index, https://www.v-dem.net/en/data/data-version-9/.

- **Regime Type:** The AIGS Index uses a four-part regime category typology established by V-Dem and Regimes of the World: closed autocracies, electoral autocracies, electoral democracies, and liberal democracies. Countries with an average score less than 2.5 are labeled closed autocracies. Countries with an average score between 4.9 and 2.5 are labeled electoral autocracies. Countries with an average score between 7.9 and 5.0 are labeled electoral democracies. Countries with an average score of 8 or higher are labeled liberal democracies. See Anna Lührmann, Marcus Tannenberg, and Staffan I. Lindberg, "Regimes of the World (RoW): Opening New Avenues for the Comparative Study of Political Regimes," Politics & Governance 6, no. 1 (2018).
- "Freedom on the Net 2018" data can be accessed at https://freedomhouse.org/report/freedom-net/freedom-net-2018.
- Military expenditure data compiled by the Stockholm International Peace Research Institute and can be accessed at https://www.sipri.org/databases/milex.

- Data on Belt and Road Initiative participating countries compiled from Dan Kliman and Abigail Grace's report, "Power Play: Addressing China's Belt and Road Strategy," Center for a New American Security, 2018, https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Power-Play-Addressing-Chinas-Belt-and-Road-Strategy.pdf?mtime=20180920093003, as well as from open source reporting.
- Data for China's overseas direct investment compiled by the Economist Intelligence Unit, "China Going Global Investment Index 2017," http://pages.eiu.com/rs/753-RIQ-438/images/ODI_in_China_2017_English.pdf.

APPENDIX 2: TAXONOMY OF DIGITAL REPRESSION

Appendix 2
Taxonomy of Digital Repression

Techniques of Digital Repression

Surveillance	Censorship	Disinformation	Cyber Attacks & Hacking	Internet Shutdowns	Targeted Arrests & Violence
AI surveillance (facial recognition systems, intelligent video surveillance, smart policing, smart cities/safe cities) ***Communications surveillance*** (internet/ social media monitoring, mobile phone tapping/ SIM registration, location monitoring, intrusion spyware, packet inspection, network interception, cable tapping, telecom surveillance) ***Surveillance laws*** (intelligence/national security laws, data disclosure, data retention, data localization)	***Political & social content blocked/filtered; use of friction & flooding*** ***Social media/ICT apps blocked*** ***Content removal*** ***Censorship laws/ directives:*** Religion/blasphemy Cyber crime False news Political/hate speech Lèse-majesté Security/terrorism Sedition Copyright infringement Defamation/libel Indecency/anti-LGBT ***Financial targeting of groups***	Government/pro-government outlets peddle disinformation, false content Cyber trolling, social media manipulation/harassment by pro-government actors (astroturfing, bots, sockpuppets, impersonation) Election manipulation (for example, data exploitation)	State-sponsored technical attacks which manipulate software, data, computer systems, or networks to degrade operational capabilities or collect information ***Categories:*** Attacks harming operational capacity Intrusion and surveillance attacks ***Illustrative Tools:*** Vandalism Distributed denial of service Man-in-the-middle Phishing Advanced persistent threat Spoofing Border Gateway Protocol	***Internet or electronic communications disrupted*** Total internet shutdowns Partial shutdowns (restricted website/social media access, blackouts, slowdowns, throttling) ***Infrastructure restrictions*** (internet firewall; closed ICT infrastructure)	ICT user charged, arrested, imprisoned, or in prolonged detention for political/social content ICT user physically attacked or killed

NOTES

[1] The index uses a four-part regime category typology established by V-Dem and Regimes of the World: closed autocracies, electoral autocracies, electoral democracies, and liberal democracies. As the authors describe it, “In closed autocracies, the chief executive is either not subjected to elections or there is no meaningful, de-facto competition in elections. Electoral autocracies hold de-facto multiparty elections for the chief executive, but they fall short of democratic standards due to significant irregularities, limitations on party competition or other violations of Dahl’s

institutional requisites for democracies. To be counted as electoral democracies, countries not only have to hold defacto free and fair and multiparty elections, but also . . . achieve a sufficient level of institutional guarantees of democracy such as freedom of association, suffrage, clean elections, an elected executive, and freedom of expression. A liberal democracy is, in addition, characterized by its having effective legislative and judicial oversight of the executive as well as protection of individual liberties and the rule of law." Anna Lührmann, Marcus Tannenberg, and Staffan I. Lindberg, "Regimes of the World (RoW): Opening New Avenues for the Comparative Study of Political Regimes," Politics & Governance 6, no. 1 (2018).

[2] Information from the Stockholm International Peace Research Institute (SIPRI), SIPRI Military Expenditure Database, 2019, https://www.sipri.org/databases/milex.

[3] "Freedom on the Net 2018: The Rise of Digital Authoritarianism," Freedom House, October 30, 2018. https://freedomhouse.org/report/freedom-net/freedom-net-2018/rise-digital-authoritarianism.

[4] In contrast to V-Dem, the AIGS Index does not evaluate the Palestinian Territories (West Bank, Gaza) and Zanzibar due to inconsistent treatment by external sources. See Michael Coppedge, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, et al, "V-Dem Country-Year Dataset 2019," Varieties of Democracy (V-Dem) Project, 2019, https://doi.org/10.23696/vdemcy19.

[5] Nils J. Nilsson, The Quest for Artificial Intelligence (United States: Cambridge University Press, 2009), 4.

[6] "Preparing for the Future of Artificial Intelligence," whitehouse.gov, October

2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
[7] Ibid.
[8] As described in the executive summary, in contrast to V-Dem, the AIGS Index does not evaluate the Palestinian Territories (West Bank, Gaza) and Zanzibar due to inconsistent treatment by external sources. "V-Dem Country-Year Dataset 2019," 2019.
[9] "Freedom on the Net 2018: The Rise of Digital Authoritarianism," October 30, 2018, 9.
[10] "Digital 2019: Internet Trends in Q3 2019," DataReportal—Global Digital Insights, July 19, 2019. https://datareportal.com/reports/digital-2019-internet-trends-in-q3.
[11] Kim Hart, "Baltimore Wrestles With Aerial Surveillance," Axios, July 31, 2019, https://www.axios.com/baltimore-wrestles-with-aerial-surveillance-to-reduce-crime-2d973591-0b33-4e25-94a7-c3f553dc2934.html.
[12] Kevin Rector and Alison Knezevich, "Maryland's Use of Facial Recognition Software Questioned by Researchers, Civil Liberties Advocates," Baltimore Sun, October 18, 2016, https://www.baltimoresun.com/news/crime/bs-md-facial-recognition-20161017-story.html.
[13] "Persistent Surveillance's Cynical Attempt to Profit Off Baltimore's Trauma," ACLU of Maryland, June 8, 2018, https://www.aclu-md.org/en/press-releases/persistent-surveillances-cynical-attempt-profit-baltimores-trauma.
[14] Olivia Solon, "'Surveillance Society': Has Technology at the US-Mexico Border Gone Too Far?," Guardian, June 13, 2018, https://www.theguardian.com/technology/2018/jun/13/

mexico-us-border-wall-surveillance-artificial-intelligence-technology.

[15] Alvaro Artigas, "Surveillance, Smart Technologies and the Development of Safe City Solutions: The Case of Chinese ICT Firms and Their International Expansion to Emerging Markets," IBEI Working Paper, 2017, https://www.ibei.org/surveillance-smart-technologies-and-the-development-of-safe-city-solutions-the-case-of-chinese-ict-firms-and-their-international-expansion-to-emerging-markets_112561.pdf.

[16] Theodore Terschlusen, "Valencienne: Demain Les Caméras de Vidéosurveillance Seront Intelligentes Et…Chinoises," September 2, 2017, https://webcache.googleusercontent.com/search?q=cache:FS-IyIma564J:www.lavoixdunord.fr/116566/article/2017-02-09/demain-les-cameras-de-videosurveillance-seront-intelligentes-et-chinoises+&cd=1&hl=en&ct=clnk&gl=us.

[17] SIPRI Military Expenditure Database, 2019.

[18] "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue," A/HRC/23/40, April 17, 2013, https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf.

[19] "The Right to Privacy in the Digital Age," Report of the Office of the United Nations High Commissioner for Human Rights, A/HRC/27/37, June 30, 2014, https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session27/Documents/A.HRC.27.37_en.pdf. See also article 12 of the Universal Declaration of Human Rights.

[20] "Surveillance and Human Rights, Report of the Special Rapporteur on the Promotion and Protection of the Right to

Freedom of Opinion and Expression," A/HRC/41/35, May 28, 2019, https://undocs.org/A/HRC/41/35.
[21] "Necessary and Proportionate," Necessary and Proportionate: International Principles on the Application of Human Rights to Communications Surveillance, March 4, 2016. https://necessaryandproportionate.org/principles.
[22] "The Right to Privacy in the Digital Age," A/HRC/27/37.
[23] "Surveillance and Human Rights," A/HRC/41/35.
[24] Steven Feldstein, "Can a U.N. Report Help Rein in Expansive and Abusive Digital Surveillance?," World Politics Review, July 9, 2019, https://www.worldpoliticsreview.com/articles/28016/can-a-u-n-report-help-rein-in-expansive-and-abusive-digital-surveillance.
[25] "The Right to Privacy in the Digital Age," A/HRC/27/37.
[26] "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue," A/HRC/23/40.
[27] Steven Feldstein, "The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression," Journal of Democracy 30, no. 1 (2019): 42.
[28] Nicholas Wright notes: "People will know that the omnipresent monitoring of their physical and digital activities will be used to predict undesired behavior, even actions they are merely contemplating. . . . In order to prevent the system from making negative predictions, many people will begin to mimic the behaviors of a "responsible" member of society. These may be as subtle as how long one's eyes look at different elements on a phone screen. This will improve social control not only by forcing people to act in certain ways, but also by changing the way they think." Nicholas Wright, "How Artificial Intelligence

Will Reshape the Global Order," Foreign Affairs, July 10, 2018, https://www.foreignaffairs.com/articles/world/2018-07-10/how-artificial-intelligence-will-reshape-global-order.

[29] See Jeffrey Ding, "Deciphering China's AI Dream," Future of Humanity Institute, University of Oxford, 2018, https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf.

[30] See for example: "China's Digital Authoritarianism: Surveillance, Influence, and Political Control," U.S. House of Representatives, Permanent Select Committee on Intelligence, committee hearing, May 16, 2019, https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=109462; Robert Morgus and Justin Sherman, "Authoritarians Are Exporting Surveillance Tech, And With It Their Vision for the Internet," Council on Foreign Relations, December 5, 2018, https://www.cfr.org/blog/authoritarians-are-exporting-surveillance-tech-and-it-their-vision-internet; "Freedom on the Net 2018: The Rise of Digital Authoritarianism," October 30, 2018; and Paul Mozur, Jonah M. Kessel, and Melissa Chan, "Made in China, Exported to the World: The Surveillance State," New York Times, April 24, 2019, https://www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-police-government.html.

[31] "Yanbu: A Smart Industrial Oil Kingdom City — Huawei Publications," Huawei, 2019, https://e.huawei.com/us/publications/global/ict_insights/201708310903/manufacturing/201712061133; Sebastian Moss, "Google Cloud Continues to Grow, Is Coming to Saudi Arabia," Data Centre Dynamics, April 24, 2018, https://www.datacenterdynamics.com/news/google-cloud-continues-to-grow-is-coming-to-saudi-arabia/; Rob

Evans, "BAE 'Secretly Sold Mass Surveillance Technology to Repressive Regimes,'" Guardian, June 14, 2017, https://www.theguardian.com/business/2017/jun/15/bae-mass-surveillance-technology-repressive-regimes; Triska Hamid, "NEC Profits from Middle East Cyber Fears," The National, December 10, 2013, https://www.thenational.ae/business/nec-profits-from-middle-east-cyber-fears-1.267420; and Alaa Shahine, Erik Schatzker, Vivian Nereim, and Glen Carey, "Saudis Are Talking to Amazon, Alibaba About New City, Prince Says," Bloomberg, October 26, 2017, https://www.bloomberg.com/news/articles/2017-10-26/saudis-are-talking-to-amazon-alibaba-over-new-city-prince-says.

[32] Korkit Danchaivichit (deputy secretary general of the National Telecommunications and Broadcasting Commission), interview with the author, May 17, 2019.

[33] Joe Parkinson, Nicholas Bariyo and Josh Chin, "Huawei Technicians Helped African Governments Spy on Political Opponents," Wall Street Journal, August 14, 2019, https://www.wsj.com/articles/huawei-technicians-helped-african-governments-spy-on-political-opponents-11565793017.

[34] "Uganda's Cash-Strapped Cops Spend $126 Million on CCTV from Huawei," Reuters, August 16, 2019, https://www.reuters.com/article/us-uganda-crime-idUSKCN1V50RF.

[35] Danielle Cave Thomas, Samantha Hoffman, Alex Joske, and Fergus Ryan, Elise, "Mapping China's Tech Giants," Australian Strategic Policy Institute, Issues Paper, Report No.

15/2019, https://www.aspi.org.au/report/mapping-chinas-tech-giants.

[36] "Building a Safe Mauritius, the Inspiration for Heaven," Huawei, 2019, https://e.huawei.com/topic/leading-new-ict-en/mauritius-safecity-case.html.

[37] Anna Fifield, "Bloodthirsty Like a Wolf: Inside the Military Style Discipline at China's Tech Titan Huawei," Washington Post, December 13, 2018, https://www.washingtonpost.com/world/asia_pacific/bloodthirsty-like-a-wolf-inside-the-military-style-discipline-at-chinas-tech-titan-huawei/2018/12/12/76055116-fd85-11e8-a17e-162b712e8fc2_story.html?noredirect=on&utm_term=.fca5427820cf.

[38] The EU's former top trade official, Karel De Gucht, is on record observing: "They [Huawei] get subsidies. If you have a line of a couple of tens of billions with the bank that you can use at your discretion this is a huge subsidy, no?" Shawn Donnan and Christian Oliver, "EU Commissioner Attacks China's Telecoms Subsidies," Financial Times, March 27, 2014, https://www.ft.com/content/d6d0bcc6-b5cb-11e3-b40e-00144feabdc0. Likewise, a 2012 U.S. congressional report from the House Intelligence Committee noted that Huawei reaps the "benefit of billions of dollars in Chinese government financing," and that both Huawei and ZTE "provide a wealth of opportunities for Chinese intelligence agencies to insert malicious hardware or software implants into critical telecommunications components and systems." "Investigative Report on the U.S. National Security Issues Posed by Chinese Telecommunications Companies Huawei and ZTE," U.S. House

of Representatives, Permanent Select Committee on Intelligence,
October 8, 2012, https://republicans-intelligence.house.gov/sites/intelligence.house.gov/files/documents/huawei-zte%20investigative%20report%20(final).pdf, 3.
[39] Max Chafkin and Joshua Brustein, "Why America Is So Scared of China's Biggest Tech Company," Bloomberg Businessweek, March 23,
2018, https://www.bloomberg.com/news/features/2018-03-22/why-america-is-so-scared-of-china-s-biggest-tech-company.
[40] Ibid.
[41] Zhang Lin, "Chinese Communist Party Needs to Curtail Its Presence in Private Businesses," South China Morning Post, November 25, 2018, https://www.scmp.com/economy/china-economy/article/2174811/chinese-communist-party-needs-curtail-its-presence-private.
[42] Arjun Kharpal, "Huawei CEO: No Matter My Communist Party Ties, I'll 'Definitely' Refuse If Beijing Wants Our Customers' Data," CNBC, January 15,
2019, https://www.cnbc.com/2019/01/15/huawei-ceo-we-would-refuse-a-chinese-government-request-for-user-data.html.
[43] Paul Mozur, "China's Internet Controls Will Get Stricter, to Dismay of Foreign Business," New York Times, November 7, 2016, https://www.nytimes.com/2016/11/08/business/international/china-cyber-security-regulations.html.
[44] Christopher Balding and Donald C. Clarke, "Who Owns Huawei?" SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, April 17,
2019, https://papers.ssrn.com/abstract=3372669.

[45] Author interview with an official from Thailand's Ministry of Interior, May 14, 2019.
[46] Somkiat Tangkitvanich (president, Thailand Development Research Institute), interview with the author, May 14, 2019.
[47] Christian Davenport, "State Repression and Political Order," Annual Review of Political Science 10 (2007): 2.
[48] "Smart Cities," World Bank, January 8, 2015, https://www.worldbank.org/en/topic/digitaldevelopment/brief/smart-cities.
[49] "Smart Cities: Utopian Vision, Dystopian Reality," Privacy International, October 2017, http://www.privacyinternational.org/report/638/smart-cities-utopian-vision-dystopian-reality.
[50] "Huawei Smart City White Paper," Huawei Enterprise, 2016, https://e.huawei.com/en/material/onLineView?MaterialID=9b0000e57fa94a2dbc0e43f5817ca767.
[51] Ibid.
[52] "The Road to Collaborative Public Safety," Huawei, 2017, http://e-file.huawei.com/~/media/EBG/Download_Files/Publications/en/Safe%20City%20Extra.pdf.
[53] "Three Rules When Using AI to Add Value to Your IoT Smart Cities," Gartner, January 29, 2018, https://www.gartner.com/doc/reprints?id=1-4XYENKG&ct=180501&st=sb.
[54] Bojan Stojkovski, "Big Brother Comes to Belgrade," Foreign Policy, June 18, 2019, accessed July 29, 2019, https://foreignpolicy.com/2019/06/18/big-brother-comes-to-belgrade-huawei-china-facial-recognition-vucic/.
[55] "Huawei Safe City Solution: Safeguards Serbia," Huawei Enterprise, August 23, 2018, http://archive.li/pZ9HO.

[56] Feldstein, "The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression," 40.
[57] "Video Surveillance as the Foundation of 'Safe City' in Kenya," Huawei, 2019, https://www.huawei.com/en/industry-insights/technology/digital-transformation/video/video-surveillance-as-the-foundation-of-safe-city-in-kenya. See also Victor Kapiyo and Grace Githaiga, "Kenya, Communications Surveillance in the Digital Age," Global Information Society Watch," 2014, https://www.giswatch.org/en/country-report/communications-surveillance/kenya.
[58] Drew Harrell, "FBI, ICE Find State Driver's License Photos Are a Gold Mine for Facial-Recognition Searches," Washington Post, July 7, 2019, https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/.
[59] Rowland Manthorpe and Alexander J Martin, "81% of 'Suspects' Flagged by Met's Police Facial Recognition Technology Innocent, Independent Report Says," Sky News, July 4, 2019, https://news.sky.com/story/met-polices-facial-recognition-tech-has-81-error-rate-independent-report-says-11755941.
[60] Charlie Warzel, "A Major Police Body Cam Company Just Banned Facial Recognition," New York Times, June 27, 2019, https://www.nytimes.com/2019/06/27/opinion/police-cam-facial-recognition.html.
[61] "NIST Evaluation Shows Advance in Face Recognition Software's Capabilities," U.S. National Institute for Standards and Technology, November 30, 2018, https://www.nist.gov/news-events/news/2018/11/nist-

evaluation-shows-advance-face-recognition-softwares-capabilities.

[62] Steve Lohr, "Facial Recognition Is Accurate, If You're a White Guy," New York Times, February 9, 2018, https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

[63] Paul Mozur, "In Hong Kong Protests, Faces Become Weapons," New York Times, July 26, 2019, https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html.

[64] Randy Rieland, "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?" Smithsonian, March 5, 2018. https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/.

[65] Caroline Haskins, "Revealed: This Is Palantir's Top-Secret User Manual for Cops," Vice, July 12, 2019, https://www.vice.com/en_us/article/9kx4z8/revealed-this-is-palantirs-top-secret-user-manual-for-cops.

[66] "The Three Pillars of Predictive Policing," PredPol, 2018, https://www.predpol.com/law-enforcement/.

[67] "China: Big Data Fuels Crackdown in Minority Region," Human Rights Watch, February 26, 2018, https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region.

[68] Cate Cadell, "From Laboratory in Far West, China's Surveillance State Spreads," Reuters, August 14, 2018, https://www.reuters.com/article/us-china-monitoring-insight-idUSKBN1KZ0R3; "China: Minority Region Collects DNA from Millions," Human Rights Watch, December 13,

2017, https://www.hrw.org/news/2017/12/13/china-minority-region-collects-dna-millions.
[69] Feldstein, "The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression," 45.
[70] Jeffrey Ding, "ChinAI #59: Is Winter Coming for Hikvision?," accessed July 28, 2019, https://chinai.substack.com/p/chinai-59-is-winter-coming-for-hikvision.
[71] For an in-depth look at the AI machine learning value chain, including an incisive examination of data storage modalities, see: Charlotte Stanton et al., "What the Machine Learning Value Chain Means for Geopolitics," Carnegie Endowment for International Peace, August 5, 2019, https://carnegieendowment.org/files/7-1-19_Stanton_etal_Machine_Learning.pdf.
[72] "Accenture Automated Border Clearance Solutions," Accenture, 2019, https://www.accenture.com/us-en/service-border-management-automated-border-clearance-summary.
[73] "EGate Solutions: Automated Border Control (ABC)," Gemalto, April 12, 2019, https://www.gemalto.com/govt/coesys/eborder-abc.
[74] "IBorderCtrl: The Project," iBorderCtrl, 2016, https://www.iborderctrl.eu/The-project.
[75] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," Psychological Science in the Public Interest, July 17, 2019, https://doi.org/10.1177/1529100619832930. See also Lucien Begault, "Automated Technologies and the Future of Fortress Europe," Amnesty International, March 28, 2019, https://www.amnesty.org/en/latest/news/2019/03/automated-technologies-and-the-future-of-fortress-europe/.

[76] Patricia Moloney Figliola and Eric A Fischer, “Overview and Issues for Implementation of the Federal Cloud Computing Initiative: Implications for Federal Information Technology Reform Management,” Congressional Research Service, January 20, 2015, https://fas.org/sgp/crs/misc/R42887.pdf

[77] Peter Mell and Timothy Grance, “The NIST Definition of Cloud Computing,” NIST Special Publication 800-145, September 2011, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

[78] Esat Dedezade, “Iceland to Become the First ‘Cloud-First-Nation,’” Microsoft News Centre Europe, September 19, 2018, https://news.microsoft.com/europe/features/iceland-to-become-the-first-cloud-first-nation/.

[79] Mehul Srivastava and Tim Bradshaw, “Israeli Group’s Spyware ‘Offers Keys to Big Tech’s Cloud’,” Financial Times, July 19, 2019, https://www.ft.com/content/95b91412-a946-11e9-b6ee-3cdf3174eb89.

[80] “Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence,” Report of the 2015-2016 Study Panel, Stanford University, September 2016, https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai100report10032016fnl_singles.pdf.

[81] Ibid.

[82] Jonathan L. Zittrain, Matthew G. Olsen, David O'Brien, and Bruce Schneier, "Don’t Panic: Making Progress on the “Going Dark” Debate." Berkman Center Research Publication 2016-1, https://cyber.harvard.edu/pubrelease/dont-panic/Dont_Panic_Making_Progress_on_Going_Dark_Debate.pdf

[83] Matt Day, Giles Turner, and Natalia Drozdiak, "Amazon Workers Are Listening to What You Tell Alexa," Bloomberg, April 10, 2019, https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio.

[84] Greg Bensinger, "Google Employs Humans to Listen to Some Voice-Assistant Recordings," Washington Post, July 11, 2019, https://www.washingtonpost.com/technology/2019/07/11/google-employs-humans-listen-some-voice-assistant-recordings/.

[85] Andy Greenberg, "This Tesla Mod Turns a Model S Into a Mobile 'Surveillance Station,'" Wired, August 9, 2019, https://www.wired.com/story/tesla-surveillance-detection-scout/.

End of document

China and the U.S Are Fighting a Major Battle over Killer Robots and the Future of AI

Chinese soldiers march past Tiananmen Square before a military parade on September 3, 2015 in Beijing, China.

Russia started sabotaging the discussion from the very first session. Throughout the morning of Aug. 21, its diplomats at the United Nations in Geneva took the floor, nitpicking language in a document meant to pave the way for an eventual ban on lethal autonomous weapons, also known as killer robots, an emerging category of weapons that would be able to fight on their own and decide who to target and kill.

“They were basically trying to waste time,” says Laura Nolan of the International Committee for Robot Arms Control, who watched with frustration in the hall.

But while Russia vigorously worked to derail progress, it had a quieter partner: China.

“I very much get the impression that they’re working together in some way,” says Nolan. “[The Chinese] are letting the Russians steamroll the process, and they’re happy to hang back.”

China has stayed coy at these discussions, which have taken place at least once a year since 2014. Its delegates

contribute just the minimum, and often send ambiguous signals on where they stand. They have called killer robots a "humanitarian concern," yet have stepped in to water down the text being debated.

Stakes are high for the emerging military power. The robots in question — while not yet humanoid, techno-thriller Terminators — would nevertheless be deadly: Imagine dozens of drones swarming like bees on the attack, or intelligent vehicles patrolling a border with shoot-to-kill orders.

At times, Beijing has given some hope to activists demanding a ban on such weapons. According to the Campaign to Stop Killer Robots, the coalition Nolan's organization is a part of, China last year joined 28 other states in saying it would support prohibiting fully autonomous weapons — but, Beijing clarified, just against their use on the battlefield, not their development nor production. That has raised eyebrows among experts skeptical of its intentions.

“They’re simultaneously working on the technology while trying to use international law as a limit against their competitors,” observes Peter Singer, a specialist on 21st century warfare.

Quite a few countries at these meetings might levy the same accusation against the United States. While Washington has not obstructed the talks, it has not appeared keen to move things forward, either.

People take part in a demonstration as part of the campaign "Stop Killer Robots" organised by German NGO "Facing Finance" to ban what they call killer robots on March 21, 2019 in front of the Brandenburg Gate in Berlin.

Part of the reluctance from major military powers over a ban stems from the extent artificial intelligence (AI) has affected their defense industries. In addition to the U.S. and China, these states also include the U.K., Australia, Israel, South Korea, and a few others. But it is China that has become the most formidable challenger in the AI competition against the American superpower.

President Xi Jinping has called for the country to become a world leader in AI by 2030, and has placed military innovation firmly at the center of the program, encouraging the People's Liberation Army (PLA) to work with startups in the private sector, and with universities.

Chinese AI companies are also making substantial contributions to the effort. Commercial giants such as SenseTime, Megvii, and Yitu sell smart surveillance cameras, voice recognition capabilities, and big data services to the government and for export. Such technology has most notably been used to police the country's far western territory of Xinjiang, where the U.N. estimates up to 1 million Uighurs, an ethnic minority, have been detained

in camps and where facial recognition devices have become commonplace.

“These technologies could easily be a key component for autonomous weapons,” says Daan Kayser of PAX, a European peace organization. Once a robot can accurately identify a face or object, only a few extra lines of code would transform it into an automatic killing machine.

In addition to technology from commercial companies, the PLA has said it plans to develop new types of combat forces, including AI and unmanned — in other words autonomous or near-autonomous — combat systems.

The country’s domestic arms industry has obliged. A few examples include manufacturer Ziyan’s new Blowfish A2 drone. The company boasts it can carry a machine gun, independently fly as a swarm group without human operators, and “engage the target autonomously.” On land, Norinco’s Cavalry, an unmanned ground vehicle with a machine gun and rocket launchers, advertises near autonomous features. And by sea, Chinese military researchers are building unmanned submarines. The 912

Project, a classified program, hopes to develop underwater robots over the next few years.

“Killer robots don’t exist yet, but what we see is a trend towards increasing autonomy,” says Kayser of PAX. “We’re very close to crossing that line, and a lot of the projects that countries are working on — of course they don’t say they’re going to be killer robots. But if we see terms like ‘autonomy in targeting’ — that’s getting very close to something that would be an autonomous weapon.”

All things considered, China’s behavior at the U.N. makes practical sense. Like other states, it is already developing intelligent weapons. The technology is fast outpacing the process at the U.N., where discussions will continue for another two years, if not longer. Without any clear international legal parameters, major militaries are feeling the pressure to invest in autonomous capabilities on the assumption that others are.

Such thinking especially characterizes the discourse around AI and autonomous weapons systems between China and the U.S.

“Essentially you have two sides that are worried about the other gaining an advantage,” says Singer. “That then has the ironic result of them both plowing resources into it, competing against each other, and becoming less secure.”

The other frontier unbound by international law is space. Here, China sees some opportunities to leapfrog American technology. It’s also where Beijing believes the U.S. would be most vulnerable in any conflict because of its dependence on information technology such as GPS, which not only helps soldiers and civilians get around, but services like stock exchanges and ATMs.

The country’s Shiyan-7 satellite, able to maneuver and dock with larger space objects, would in theory, experts say, also be able to latch on to and disable enemy space assets. More recently, China has been testing satellite SJ-17. It moves around with precision at very high altitudes — 22,000 miles above Earth. Satellites in orbit fly at tens of thousands of miles per hour. They possess the kinetic potency to shatter anything in their path, essentially acting as kamikazes against another country’s satellite. The U.S. military worries

this is what China has in mind when developing satellites that can move so unusually in space.

Advanced space weapons, killer robots, and the U.S. and China preparing for World War III! It may all sound surreal, like a spectacular science fiction, but in the staid halls of the U.N., over the draft documents bureaucrats pass around, they are exactly what countries are anticipating. What makes their work more challenging than past international weapons bans is the preemptive nature of it, and the technology involved that would make enforcement and verification difficult, if not impossible.

Kayser knows time is running out. “An AI arms race would have no winners,” he says. Preventing one from happening would depend on the major powers. He isn’t optimistic.

“They are not taking their responsibility to ensure that international peace and security is maintained. They are actually taking steps that are dangerous and risky for international peace.”

The Big Data Artificial Intelegence Hot War That Threatens Us All?

2201

In the spring of 2016, an artificial intelligence system called AlphaGo defeated a world champion Go player in a match at the Four Seasons hotel in Seoul. In the US, this momentous news required some unpacking. Most Americans were unfamiliar with Go, an ancient Asian game that involves placing black and white stones on a wooden board. And the technology that had emerged victorious was even more foreign: a form of AI called machine learning, which uses large data sets to train a computer to recognize patterns and make its own strategic choices.

Still, the gist of the story was familiar enough. Computers had already mastered checkers and chess; now they had learned to dominate a still more complex game. Geeks cared, but most people didn't. In the White House, Terah Lyons, one of Barack Obama's science and technology policy advisers, remembers her team cheering on the fourth floor of the Eisenhower Executive Building. "We saw it as a win for technology," she says. "The next day the rest of the White House forgot about it."

In China, by contrast, 280 million people watched AlphaGo win. There, what really mattered was that a machine owned by a California company, Alphabet, the parent of Google, had conquered a game invented more than 2,500 years ago in Asia. Americans don't even play Go. And yet they had somehow figured out how to vanquish it? Kai-Fu Lee, a pioneer in the field of AI, remembers being asked to comment on the match by nearly every major television

station in the country. Until then, he had been quietly investing in Chinese AI companies. But when he saw the attention, he started broadcasting his venture fund's artificial intelligence investment strategy. "We said, OK, after this match, the whole country is going to know about AI," he recalls. "So we went big."

In Beijing, the machine's victory cracked the air like a warning shot. That impression was only reinforced when, over the next few months, the Obama administration published a series of reports grappling with the benefits and risks of AI. The papers made a series of recommendations for government action, both to stave off potential job losses from automation and to invest in the development of machine learning. A group of senior policy wonks inside China's science and technology bureaucracy, who had already been working on their own plan for AI, believed they were seeing signs of a focused, emerging US strategy—and they needed to act fast.

In May 2017, AlphaGo triumphed again, this time over Ke Jie, a Chinese Go master, ranked at the top of the world. Two months later, China unveiled its Next Generation Artificial Intelligence Development Plan, a document that laid out the country's strategy to become the global leader in AI by 2030. And with this clear signal from Beijing, it was as if a giant axle began to turn in the machinery of the industrial state. Other Chinese government ministries soon issued their own plans, based on the strategy sketched out by Beijing's planners. Expert advisory groups and industry alliances cropped up, and local governments all over China began to fund AI ventures.

China's tech giants were enlisted as well. Alibaba, the giant online retailer, was tapped to develop a "City Brain" for a new Special Economic Zone being planned about 60 miles southwest of Beijing. Already, in the city of Hangzhou, the company was soaking up data from thousands of street cameras and using it to control traffic lights with AI, optimizing traffic flow in much the way AlphaGo had optimized for winning moves on the Go board; now Alibaba would help design AI into a new megacity's entire infrastructure from the ground up.

On October 18, 2017, China's president, Xi Jinping, stood in front of 2,300 of his fellow party members, flanked by enormous red drapes and a giant gold hammer and sickle. As Xi laid out his plans for the party's future over nearly three and a half hours, he named artificial intelligence, big data, and the internet as core technologies that would help transform China into an advanced industrial economy in the coming decades. It was the first time many of these technologies had explicitly come up in a president's speech at the Communist Party Congress, a once-in-five-years event.

In the decisive span of a few months, the Chinese government had given its citizens a new vision of the future, and made clear that it would be coming fast. "If AlphaGo was China's Sputnik moment, the government's AI plan was like President John F. Kennedy's landmark speech calling for America to land a man on the moon," Kai-Fu Lee writes in his new book, *AI Superpowers*.

Meanwhile, as Beijing began to build up speed, the United States government was slowing to a walk. After President

Trump took office, the Obama-era reports on AI were relegated to an archived website. In March 2017, Treasury secretary Steven Mnuchin said that the idea of humans losing jobs because of AI “is not even on our radar screen.” It might be a threat, he added, in “50 to 100 more years.” That same year, China committed itself to building a $150 billion AI industry by 2030.

Only slowly, pushed mainly by the Pentagon, has the Trump administration begun to talk about, and fund, national AI initiatives. In May, secretary of defense James Mattis read an article in *The Atlantic* by Henry Kissinger, who warned that AI was moving so quickly it could soon subvert human intelligence and creativity. The result, he warned, could be the end of the Enlightenment; he called for a government commission to study the issue.

Many AI experts pooh-poohed Kissinger’s article for extrapolating too broadly and darkly from the field’s narrow accomplishments. Mattis, however, pulled the article into a memo for President Trump. That month, Michael Kratsios, Trump’s top adviser on technology, organized a summit on the subject of AI. In an interview with WIRED this summer, Kratsios said the White House was fully committed to AI research and to figuring out “what the government can do, and how it can do it even more.” In June, Ivanka Trump tweeted out a link to the Kissinger piece, praising its account of “the ongoing technological revolution whose consequences we have failed to fully reckon with.”

But if the Trump White House was relatively slow to grasp the significance and potential of AI, it was quick to rivalry.

By midsummer, talk of a “new cold war arms race” over artificial intelligence was pervasive in the US media.

At the dawn of a new stage in the digital revolution, the world’s two most powerful nations are rapidly retreating into positions of competitive isolation, like players across a Go board. And what’s at stake is not just the technological dominance of the United States. At a moment of great anxiety about the state of modern liberal democracy, AI in China appears to be an incredibly powerful enabler of authoritarian rule. Is the arc of the digital revolution bending toward tyranny, and is there any way to stop it?

After the end of the Cold War, conventional wisdom in the West came to be guided by two articles of faith: that liberal democracy was destined to spread across the planet, and that digital technology would be the wind at its back. The censorship, media consolidation, and propaganda that had propped up Soviet-era autocracies would simply be inoperable in the age of the internet. The World Wide Web would give people free, unmediated access to the world’s information. It would enable citizens to organize, hold governments accountable, and evade the predations of the state.

No one had more confidence in the liberalizing effects of technology than the tech companies themselves: Twitter was, in one executive’s words, “the free

speech wing of the free speech party"; Facebook wanted to make the world more open and connected; Google, cofounded by a refugee from the Soviet Union, wanted to organize the world's information and make it accessible to all.

As the era of social media kicked in, the techno-optimists' twin articles of faith looked unassailable. In 2009, during Iran's Green Revolution, outsiders marveled at how protest organizers on Twitter circumvented the state's media blackout. A year later, the Arab Spring toppled regimes in Tunisia and Egypt and sparked protests across the Middle East, spreading with all the virality of a social media phenomenon—because, in large part, that's what it *was*. "If you want to liberate a society, all you need is the internet," said Wael Ghonim, an Egyptian Google executive who set up the primary Facebook group that helped galvanize dissenters in Cairo.

It didn't take long, however, for the Arab Spring to turn into winter—in ways that would become eerily familiar to Western countries in a few years. Within a few weeks of President Hosni Mubarak's departure, Ghonim saw activists start to turn on each other. Social media was amplifying everyone's worst instincts. "You could easily see the voices in the middle become more and more irrelevant, the voices on the extremes becoming more and more heard," he recalls. The activists who were vulgar or attacked other groups or responded with rage got more likes and shares. That gave them more influence, and it gave otherwise moderate people a model to emulate. Why post something conciliatory if no one on Facebook will read it? Instead, post

something full of vitriol that millions will see. Ghonim began to become dispirited. The tools that had brought the protesters together, he said, were now tearing them apart.

Political opinions have become more polarized, populations have become more tribal, and civic nationalism is disintegrating.

Ultimately, Egypt elected a government run by the Muslim Brotherhood, a traditionalist political machine that had played little part in the initial Tahrir Square groundswell. Then in 2013 the military staged a successful coup. Soon thereafter, Ghonim moved to California, where he tried to set up a social media platform that would favor reason over outrage. But it was too hard to peel users away from Twitter and Facebook, and the project didn't last long. Egypt's military government, meanwhile, recently passed a law that allows it to wipe its critics off social media.

Of course, it's not just in Egypt and the Middle East that things have gone sour. In a remarkably short time, the exuberance surrounding the spread of liberalism and technology has turned into a crisis of faith in both. Overall, the number of liberal democracies in the world has been in steady decline for a decade. According to Freedom House, 71 countries last year saw declines in their political rights and freedoms; only 35 saw improvements.

While the crisis of democracy has many causes, social media platforms have come to seem like a prime culprit. The recent wave of antiestablishment politicians and nativist

political movements—Donald Trump in the United States; Brexit in the UK; the resurgent right wing in Germany, Italy, or across Eastern Europe—has revealed not only a deep disenchantment with the global rules and institutions of Western democracy, but also an automated media landscape that rewards demagoguery with clicks. Political opinions have become more polarized, populations have become more tribal, and civic nationalism is disintegrating.

Which leaves us where we are now: Rather than cheering for the way social platforms spread democracy, we are busy assessing the extent to which they corrode it.

In China, government officials watched the Arab Spring with attentiveness and unease. Beijing already had the world's most sophisticated internet control system, dynamically blocking a huge swath of foreign web domains, including Google. Now it garlanded its Great Firewall with even more barbed wire. China developed new ways to surgically turn off internet access in zones within cities, including a major block of downtown Beijing where it feared demonstrations. It also digitally walled off the entire province of Xinjiang after violent protests there that spread via the internet. Beijing may even have dabbled with creating a nationwide internet "kill switch."

This bowdlerized version of the internet doesn't sound at all like the original dream of the World Wide Web, but it has

thrived nonetheless. By now, there are roughly 800 million people who surf the internet, exchange chat messages, and shop online behind the Great Firewall—nearly as many people as live in the United States and Europe combined. And for many Chinese, rising middle-class prosperity has made online censorship considerably easier to bear. Give me liberty, the line might go, or give me wealth.

China's authoritarianism, which has doubled down under Xi's leadership, certainly hasn't hindered the Chinese tech industry. Over the past decade, China's leading tech companies have come to dominate their home markets and compete globally. They've expanded through acquisitions in Southeast Asia. Baidu and Tencent have set up research centers in the US, and Huawei sells advanced networking equipment in Europe. The old Silk Road is being strung with Chinese fiber-optic cables and network equipment.

More than any other country, China has shown that, with a few adjustments, autocracy is quite compatible with the internet age. But those adjustments have caused the internet itself to start to break apart, like two continents cracking along a shelf. There's the freewheeling, lightly regulated internet dominated by the geeks of Silicon Valley. And then there's China's authoritarian alternative, powered by massive, home-grown tech giants as innovative as their Western counterparts.

In China and the West alike, power comes from controlling data, making sense of it,

and using it to influence how people behave.

Today, China doesn't just play defense against viral dissent by redacting troublesome parts of the internet; the government actively wields technology as a tool of control. In cities across China, including in Xinjiang, authorities are trying out facial-recognition software and other AI-powered technologies for security. In May, facial-recognition cameras at Jiaxing Sports Center Stadium in Zhejiang led to the arrest of a fugitive who was attending a concert. He had been wanted since 2015 for allegedly stealing more than $17,000 worth of potatoes. China's Police Cloud System is built to monitor seven categories of people, including those who "undermine stability." The country also aspires to build a system that will give every citizen and every company a social credit score: Imagine your FICO score adjusted to reflect your shopping habits, your driving record, and the appropriateness of your politics.

The fundamental force driving this change—this pivot from defense to offense—is a shift in how power flows from technology. In the beginning, the communications revolution made computers affordable to the masses. It wired devices together in a giant global network and shrank them down to the size of your hand. It was a revolution that empowered the individual—the lone programmer with the power to create in her pocket, the academic with infinite research at his fingertips, the dissident with a new and powerful way of organizing resistance.

Today's stage of the digital revolution is different. That supercomputer in your pocket is also a homing device. It's tracking your every "like," keeping a record of everyone you talk to, everything you buy, everything you read, and everywhere you go. Your fridge, your thermostat, your smartwatch, and your car are increasingly sending your information back to headquarters too. In the future, security cameras will track the ways our eyes dilate, and sensors on the wall will track our body temperature.

In today's digital world, in China and the West alike, power comes from controlling data, making sense of it, and using it to influence how people behave. That power will only grow as the next generation of mobile networks goes live. Remember how it felt like magic to be able to browse real web pages on the second-generation iPhone? That was 3G, the mobile standard that became widespread in the mid-2000s. Modern 4G networks are several times faster. 5G will be vastly faster still. And when we can do things faster, we do them more, which means data piles up.

It's already hard for most people to comprehend, much less control, all the information collected about them. And the leverage that accrues to data aggregators will just increase as we move into the era of AI.

Vladimir Putin is a technological pioneer when it comes to cyberwarfare and disinformation. And he has an opinion

about what happens next with AI: “The one who becomes the leader in this sphere will be the ruler of the world.”

In a way, Putin’s line is a bit overwrought. AI is not a hill that one nation can conquer or a hydrogen bomb that one country will develop first. Increasingly, AI is simply how computers work; it’s a broad term describing systems that learn from examples—or follow rules—to make independent decisions. Still, it’s easily the most important advance in computer science in a generation. Sundar Pichai, the CEO of Google, has compared it to the discovery of electricity or fire.

A country that strategically and smartly implements AI technologies throughout its workforce will likely grow faster, even as it deals with the disruptions that AI is likely to cause. Its cities will run more efficiently, as driverless cars and smart infrastructure cut congestion. Its largest businesses will have the best maps of consumer behavior. Its people will live longer, as AI revolutionizes the diagnosis and treatment of disease. And its military will project more power, as autonomous weapons replace soldiers on the battlefield and pilots in the skies and as cybertroops wage digital warfare. “I can’t really think of any mission that doesn’t have the potential to be done better or faster if properly integrated with AI,” says Will Roper, an assistant secretary of the US Air Force.

And these benefits may compound with interest. So far, at least, AI appears to be a centralizing force, among companies and among nations. The more data you gather, the better the systems you can build; and better systems allow you to collect more data. “AI will become

concentrated, because of the inputs required to pull it off. You need a lot of data and you need a lot of computing power," says Tim Hwang, who leads the Harvard-MIT Ethics and Governance of AI Initiative.

The Chinese government can access personal data for reasons of public or national security without the same legal constraints a democracy would face.

China has two fundamental advantages over the US in building a robust AI infrastructure, and they're both, generally, advantages that authoritarian states have over democratic ones. The first is the sheer scope of the data generated by Chinese tech giants. Think of how much data Facebook collects from its users and how that data powers the company's algorithms; now consider that Tencent's popular WeChat app is basically like Facebook, Twitter, and your online bank account all rolled into one. China has roughly three times as many mobile phone users as the US, and those phone users spend nearly *50 times* as much via mobile payments. China is, as The Economist first put it, the Saudi Arabia of data. Data privacy protections are on the rise in China, but they are still weaker than those in the US and much weaker than those in Europe, allowing data aggregators a freer hand in what they can do with what they collect. And the government can access personal data for reasons of public or national security without the same legal constraints a democracy would face.

Of course, data isn't everything: Any technological system depends on a whole stack of tools, from its software to its processors to the humans who curate noisy inputs and analyze results. And there are promising subfields of AI, such as reinforcement learning, that generate their own data from scratch, using lots of computing power. Still, China has a second big advantage as we move into the era of AI, and that's the relationship between its largest companies and the state. In China, the private-sector companies at the cutting edge of AI innovation feel obliged to keep Xi's priorities in mind. Under Xi, Communist Party committees within companies have expanded. Last November, China tapped Baidu, Alibaba, Tencent, and iFlytek, a Chinese voice-recognition software company, as the inaugural members of its "AI National Team." The message was clear: Go forth, invest, and the government will ensure that your breakthroughs have a market not just in China, but beyond.

During the original Cold War, the US relied on companies like Lockheed, Northrop, and Raytheon to develop cutting-edge strategic technology. Technically, these companies were privately owned. In practice, their vital defense mission made them quasipublic entities. (Indeed, long before the phrase "too big to fail" was ever used to describe a bank, it was applied to Lockheed.)

Fast forward to today, and the companies at the forefront of AI—Google, Facebook, Amazon, Apple, and Microsoft—don't exactly wear flag pins on their lapels. This past spring, employees at Google demanded that the company pull out of a Pentagon collaboration called Project Maven. The idea

was to use AI for image recognition in Defense Department missions. Ultimately, Google's management caved. Defense Department officials were bitterly disappointed, especially given that Google has a number of partnerships with Chinese technology companies. "It is ironic to be working with Chinese companies as though that is not a direct channel to the Chinese military," says former secretary of defense Ashton Carter, "and not to be willing to operate with the US military, which is far more transparent and which reflects the values of our society. We're imperfect for sure, but we're not a dictatorship."

The Cold War wasn't inevitable in 1945. The United States and Soviet Union had been allies during World War II, but then a series of choices and circumstances over a five-year period set the conflict on its self-perpetuating track. Similarly, as we can now see in the cold glare of hindsight, it was never inevitable that the digital revolution would inherently favor democracy. Nor is it inevitable today that AI will favor global authoritarianism to the permanent disadvantage of liberalism. If that scenario comes to pass, it will be because a series of choices and circumstances precipitated it.

In the original Cold War, two ideological foes created rival geopolitical blocs that were effectively non-interoperable. The US was boxed out of the Soviet bloc, and vice versa.

The same could easily happen again, to disastrous effect. A new cold war that gradually isolates the Chinese and American tech sectors from each other would starve the US of much of the fuel it now relies on for innovation: American companies depend heavily on the Chinese market for their profits and for engineering and software talent. At the same time, it could actually create the kinds of dangers that hawks warn about now: It would increase the risk that one side could surprise the other with a decisive strategic breakthrough in AI or quantum computing.

Right now, maintaining a degree of openness with China is the best defense against the growth of a techno--authoritarian bloc. That's not the way American leaders are headed, though.

A little over six months after Donald Trump's inauguration—and his invocation of "American carnage"—the administration launched a sweeping investigation into China's trade practices and alleged theft of US technology via cyberspace. That investigation has mushroomed into a steadily escalating trade war, with the US launching tariffs on billions of dollars of Chinese goods and new investment and export restrictions on technologies that China considers key to AI and to its advanced manufacturing ambitions.

For US security hawks, the prospect that China might dominate both 5G and AI is a nightmare scenario.

The confrontation is about much more than trade. The Trump administration has made it official US policy to

protect the “national security innovation base”—White House shorthand for America’s leading technology and talent—from China and other foreign economic predators. In January, Axios published a leaked White House presentation that recommended the US work with its allies to build a 5G network that excludes China, to prevent Beijing from grabbing “the commanding heights of the information domain.” The presentation likened the 21st-century struggle for data dominance to the WWII-era race to construct an atom bomb. Then in April, the US Commerce Department hit ZTE, a leading Chinese telecommunications equipment firm that was gearing up to work on China’s 5G network, with a seven-year ban on doing business with US suppliers; the department said ZTE had violated the terms of a sanctions settlement. (The US later lifted the ban.)

For US security hawks, the prospect that China might dominate both 5G and AI is a nightmare scenario. At the same time, Washington’s escalating pushback against China’s tech ambitions has made Xi even more determined to wean his country off Western technology.

This is a very different philosophy from the one that has guided the technology sector for 30 years, which has favored deeply enmeshed hardware and software supply chains. Shortly before Trump’s inauguration, Jack Ma, the chair of Alibaba, pledged to create a million jobs in the United States. By September 2018, he was forced to admit that the offer was off the table, another casualty in the growing list of companies and projects that are now unthinkable.

Global work in AI has long taken place in three spheres: research departments, corporations, and the military. The first sphere has always been marked by openness and cooperation; to a lesser extent, so has the second. Academics freely share their work. Microsoft has trained many of China's best AI researchers and helped nurture many promising AI startups, and Alibaba, Baidu, and Tencent employ US engineers at their research hubs in Silicon Valley and Seattle. An AI-driven breakthrough in Shanghai—say, in diagnosing disease through more accurate scans of medical images—can save lives in Shawnee. But national security concerns have a way of overriding commercial considerations. For now, the political momentum appears to be driving the two countries' tech sectors apart to such a degree that even collaboration between researchers and corporations could be stifled. The schism could well define how the struggle between democracy and authoritarianism plays out.

Imagine its 2022: America's confrontational economic policies have continued, and China has refused to yield. Huawei and ZTE have been banned from the networks of the US and key Western allies. Through investment and theft, Beijing has reduced its reliance on US semiconductors. Rival tech superpowers have failed to develop common standards. US and Chinese academics increasingly deposit their cutting-edge AI research in

government safes instead of sharing it at international conferences. Other countries—like France and Russia—have tried to build homegrown technology industries centered on AI, but they lag far behind.

The world's nations can commit to American technology: buying Apple phones, using Google search, driving Teslas, and managing a fleet of personal robots made by a startup in Seattle. Or they can commit to China: using the equivalents built by Alibaba and Tencent, connecting through the 5G network constructed by Huawei and ZTE, and driving autonomous cars built by Baidu. The choice is a fraught one. If you are a poor country that lacks the capacity to build your own data network, you're going to feel loyalty to whoever helps lay the pipes at low cost. It will all seem uncomfortably close to the arms and security pacts that defined the Cold War.

And we may be seeing the first evidence of this. In May 2018, about six months after Zimbabwe finally got rid of the despot Robert Mugabe, the new government announced that it was partnering with a Chinese company called CloudWalk to build an AI and facial-recognition system. Zimbabwe gets to expand its surveillance state. China gets money, influence, and data. In July, nearly 700 dignitaries from China and Pakistan gathered in Islamabad to celebrate the completion of the Pak-China Optical Fibre Cable, a 500-mile-long data line connecting the two countries through the Karakoram Mountains, built by Huawei and financed with a loan from China's Export-Import Bank. Documents obtained by Pakistan's *Dawn* newspaper revealed a future plan for high-

speed fiber to help wire up cities across Pakistan with surveillance cameras and vehicle-monitoring systems, part of a “Safe Cities” initiative launched in 2016 with help from Huawei and other Chinese firms. China has effectively constructed its own Marshall Plan, one that may, in some cases, build surveillance states instead of democracies.

It’s not hard to see the appeal for much of the world of hitching their future to China. Today, as the West grapples with stagnant wage growth and declining trust in core institutions, more Chinese people live in cities, work in middle-class jobs, drive cars, and take vacations than ever before. China’s plans for a tech-driven, privacy-invading social credit system may sound dystopian to Western ears, but it hasn’t raised much protest there. In a recent survey by the public relations consultancy Edelman, 84 percent of Chinese respondents said they had trust in their government. In the US, only a third of people felt that way.

For the past century, democracies have proven more resilient and successful than dictatorships, even if democracies have made stupid decisions along the way.

No one can be certain what happens next. In the US, in the wake of controversies surrounding the 2016 election and user privacy, a growing number of Republicans and Democrats want to regulate America’s tech giants and rein them in. At the same time, China has stiffened its resolve to become an AI superpower and export its techno--

authoritarian revolution—which means the US has a vital national interest in ensuring that its tech firms remain world leaders. For now, there is nothing close to a serious debate about how to address this dilemma.

As for China, it remains unclear how much digital intrusion people there will tolerate in the name of efficiency and social cohesion—to say nothing of people in other countries that are tempted by Beijing's model. Regimes that ask people to trade freedom for stability tend to invite dissent. And Chinese growth is slowing. For the past century, democracies have proven more resilient and successful than dictatorships, even if democracies, particularly in an age of algorithms, have made some stupid decisions along the way.

It is at least conceivable that Trump's aggressive policies could, counterintuitively, lead to a rapprochement with Beijing. If Trump threatens to take something off the table that China truly cannot afford to lose, that could pressure Beijing to dial back its global tech ambitions and open its domestic market to US firms. But there is another way to influence China, one more likely to succeed: The US could try to wrap Beijing in a technology embrace. Work with China to develop rules and norms for the development of AI. Establish international standards to ensure that the algorithms governing people's lives and livelihoods are transparent and accountable. Both countries could, as Tim Hwang suggests, commit to developing more shared, open databases for researchers.

But for now, at least, conflicting goals, mutual suspicion, and a growing conviction that AI and other advanced

technologies are a winner-take-all game are pushing the two countries' tech sectors further apart. A permanent cleavage will come at a steep cost and will only give techno-authoritarianism more room to grow.

Artificial intelligence arms race

From Wikipedia, the free encyclopedia

An **artificial intelligence arms race** is a competition between two or more states to have its military forces equipped with the best "artificial intelligence" (AI). Since the mid-2010s, many analysts have argued that such a global arms race for better artificial intelligence has already begun.

Contents

- 1Stances toward military artificial intelligence
  - 1.1Russia
  - 1.2China
  - 1.3United States
  - 1.4United Kingdom
  - 1.5Israel
  - 1.6South Korea
- 2Trends
- 3Proposals for international regulation
- 4Other reactions to autonomous weapons
- 5Disassociation
- 6See also
- 7References
- 8Further reading

Stances toward military artificial intelligence

Russia

Russian General Viktor Bondarev, commander-in-chief of the Russian air force, has stated that as early as February 2017, Russia has been working on AI-guided missiles that can decide to switch targets mid-flight.[1] Reports by state-sponsored Russian media on potential military uses of AI increased in mid-2017.[2] In May 2017, the CEO of Russia's Kronstadt Group, a defense contractor, stated that "there already exist completely autonomous AI operation systems that provide the means for UAV clusters, when they fulfill missions autonomously, sharing tasks between them, and interact", and that it is inevitable that

"swarms of drones" will one day fly over combat zones.[3] Russia has been testing several autonomous and semi-autonomous combat systems, such as Kalashnikov's "neural net" combat module, with a machine gun, a camera, and an AI that its makers claim can make its own targeting judgements without human intervention.[4] In September 2017, during a National Knowledge Day address to over a million students in 16,000 Russian schools, Russian President Vladimir Putin stated "Artificial intelligence is the future, not only for Russia but for all humankind... Whoever becomes the leader in this sphere will become the ruler of the world".[5][6]

The Russian government has strongly rejected any ban on lethal autonomous weapons systems, suggesting that such a ban could be ignored.[7][8]

China

Further information: Artificial intelligence industry in China

According to a February 2019 report by Gregory C. Allen of the Center for a New American Security, "China's leadership – including President Xi Jinping – believes that being at the forefront in AI technology is critical to the future of global military and economic power competition."[9] Chinese military officials have said that their goal is to incorporate commercial AI technology to "narrow the gap between the Chinese military and global advanced powers."[9] The close ties between Silicon Valley and China, and the open nature of the American research community, has made the West's most advanced AI technology easily available to China; in addition, Chinese industry has numerous home-grown AI accomplishments of its own, such as Baidu passing a notable Chinese-language speech recognition capability benchmark in 2015.[10] As of 2017, Beijing's roadmap aims to create a $150 billion AI industry by 2030.[11] Before 2013, Chinese defense procurement was mainly restricted to a few conglomerates; however, as of 2017, China often sources

sensitive emerging technology such as drones and artificial intelligence from private start-up companies.[12] One Chinese state has pledged to invest $5 billion in AI. Beijing has committed $2 billion to an AI development park.[13] The *Japan Times* reported in 2018 that annual private Chinese investment in AI is under $7 billion per year. AI startups in China received nearly half of total global investment in AI startups in 2017; the Chinese filed for nearly five times as many AI patents as did Americans.[14]

China published a position paper in 2016 questioning the adequacy of existing international law to address the eventuality of fully autonomous weapons, becoming the first permanent member of the U.N. Security Council to broach the issue.[15] In 2018, Xi called for greater international cooperation in basic AI research.[16] Chinese officials have expressed concern that AI such as drones could lead to accidental war, especially in the absence of international norms.[17]

United States

The *Sea Hunter*, an autonomous US warship, 2016

In 2014, former Secretary of Defense Chuck Hagel posited the "Third Offset Strategy" that rapid advances in artificial intelligence will define the next generation of warfare.[18] According to data science and analytics firm Govini, The U.S. Department of Defense increased investment in artificial intelligence, big data and cloud computing from $5.6 billion in 2011 to $7.4 billion in 2016.[19] However, the civilian NSF budget for AI saw no increase in 2017.[11]

The U.S. has many military AI combat programs, such as the *Sea Hunter* autonomous warship, which is designed to operate for extended periods at sea without a single crew member, and to even guide itself in and out of port.[4] As of 2017, a temporary US Department of Defense directive requires a human operator to be kept in the loop when it comes to the taking of human life by

autonomous weapons systems.[20] *Japan Times* reported in 2018 that the United States private investment is around $70 billion per year.[14]

United Kingdom

In 2015, the UK government opposed a ban on lethal autonomous weapons, stating that "international humanitarian law already provides sufficient regulation for this area", but that all weapons employed by UK armed forces would be "under human oversight and control".[21]

Israel

Israel's Harpy anti-radar "fire and forget" drone is designed to be launched by ground troops, and autonomously fly over an area to find and destroy radar that fits pre-determined criteria.[22]

South Korea

The South Korean Super aEgis II machine gun, unveiled in 2010, sees use both in South Korea and in the Middle East. It can identify, track, and destroy a moving target at a range of 4 km. While the technology can theoretically operate without human intervention, in practice safeguards are installed to require manual input. A South Korean manufacturer states, "Our weapons don't sleep, like humans must. They can see in the dark, like humans can't. Our technology therefore plugs the gaps in human capability", and they want to "get to a place where our software can discern whether a target is friend, foe, civilian or military".[23]

Trends

According to Siemens, worldwide military spending on robotics was US$5.1 billion in 2010 and US$7.5 billion in 2015.[24][25]

China became a top player in artificial intelligence research in the 2010s. According to the *Financial Times*, in 2016, for the first time, China published more AI papers than the entire European Union. When restricted to number of AI papers in the top 5% of

cited papers, China overtook the United States in 2016 but lagged behind the European Union.[11] 23% of the researchers presenting at the 2017 American Association for the Advancement of Artificial Intelligence (AAAI) conference were Chinese.[26] Eric Schmidt, the former chairman of Alphabet, has predicted China will be the leading country in AI by 2025.[27]

AAAI presenters:[26]

Country	in 2012	in 2017
US	41%	34%
China	10%	23%
UK	5%	5%

Proposals for international regulation

As early as 2007, scholars such as AI professor Noel Sharkey have warned of "an emerging arms race among the hi-tech nations to develop autonomous submarines, fighter jets, battleships and tanks that can find their own targets and apply violent force without the involvement of meaningful human decisions".[28][29] As early as 2014, AI specialists such as Steve Omohundro have been warning that "An autonomous weapons arms race is already taking place".[30] Miles Brundage of the University of Oxford has argued an AI arms race might be somewhat mitigated through diplomacy: "We saw in the various historical arms races that collaboration and dialog can pay dividends".[31] Over a hundred experts signed an open letter in 2017 calling on the UN to address the issue of lethal autonomous weapons;[32][33] however, at a November 2017 session of the UN Convention on Certain Conventional Weapons (CCW), diplomats could not agree even on how to define such weapons.[34] The Indian ambassador and chair of the CCW stated that agreement on rules remained a distant prospect. As of 2017, twenty-two countries have called for a full ban on lethal autonomous weapons.[35]

Many experts believe attempts to completely ban killer robots are likely to fail.[36] A 2017 report from Harvard's Belfer Center predicts that AI has the potential to be as transformative as nuclear weapons.[31][37][38] The report further argues that "Preventing expanded military use of AI is likely impossible" and that "the more modest goal of safe and effective technology management must be pursued", such as banning the attaching of an AI dead man's switch to a nuclear arsenal.[38] Part of the impracticality is that detecting treaty violations would be extremely difficult.[39][40]

Other reactions to autonomous weapons

A 2015 open letter calling for the ban of lethal automated weapons systems has been signed by tens of thousands of citizens, including scholars such as physicist Stephen Hawking, Tesla magnate Elon Musk, and Apple's Steve Wozniak.[34]

Professor Noel Sharkey of the University of Sheffield has warned that autonomous weapons will inevitably fall into the hands of terrorist groups such as the Islamic State.[41]

Disassociation

Many Western tech companies are leery of being associated too closely with the U.S. military, for fear of losing access to China's market.[10] Furthermore, some researchers, such as DeepMind's Demis Hassabis, are ideologically opposed to contributing to military work.[42]

For example, **Project Maven** is a Pentagon project involving using machine learning and engineering talent to distinguish people and objects in drone videos,[43] established in a memo by the U.S. Deputy Secretary of Defense on 26 April 2017.[44] Also known as the Algorithmic Warfare Cross Functional Team,[45] it is, according to Lt. Gen. of the United States Air Force Jack Shanahan in November 2017, a project "designed to be that pilot project, that pathfinder, that spark that kindles the flame front of artificial intelligence across the rest of the [Defense] Department".[46] Its chief, U.S. Marine Corps Col. Drew Cukor, said: "People and computers will work symbiotically to increase the ability of weapon systems to detect objects."[47] At the second Defense One Tech Summit in July 2017, Cukor also said that the investment in a "deliberate workflow process" was funded by the Department [of Defense] through its "rapid acquisition authorities" for about "the next 36 months".[48]

In June 2018, company sources at Google said that top executive Diane Greene told staff that the company would not follow-up Project Maven after the current contract expires in March 2019.[43]

See also

- Arms race
- Artificial intelligence
- Artificial general intelligence
- Ethics of artificial intelligence
- Existential risk from artificial general intelligence
- Global issue
- Lethal autonomous weapon
- Military robot
- Unmanned combat aerial vehicle

References

1. ^ *"Russia is building a missile that can makes its own decisions"*. *Newsweek. 20 July 2017. Retrieved 24 December 2017.*
2. ^ *"Why Elon Musk is right about the threat posed by Russian artificial intelligence"*. *The Independent. 6 September 2017. Retrieved 24 December 2017.*
3. ^ *"Russia is developing autonomous "swarms of drones" it calls an inevitable part of future warfare"*. *Newsweek. 15 May 2017. Retrieved 24 December2017.*
4. ^ Jump up to:[a] [b] *Smith, Mark (25 August 2017). "Is 'killer robot' warfare closer than we think?". BBC News. Retrieved 24 December 2017.*
5. ^ *"Artificial Intelligence Fuels New Global Arms Race". WIRED. Retrieved 24 December 2017.*
6. ^ *Clifford, Catherine (29 September 2017). "In the same way there was a nuclear arms race, there will be a race to build A.I., says tech exec". CNBC. Retrieved 24 December 2017.*
7. ^ *"Russia rejects potential UN 'killer robots' ban, official statement says". Institution of Engineering and Technology. 1 December 2017. Retrieved 12 January 2018.*
8. ^ *"Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, Russian Federation, November 2017" (PDF). Retrieved 12 January 2018.*
9. ^ Jump up to:[a] [b] *Allen, Gregory. "Understanding China's AI Strategy". Center for a New American Security. Center for a New American Security. Retrieved 15 March 2019.*

10.^ Jump up to:[a] [b] *Markoff, John; Rosenberg, Matthew (3 February 2017). "China's Intelligent Weaponry Gets Smarter". The New York Times. Retrieved 24 December 2017.*
11.^ Jump up to:[a] [b] [c] *"China seeks dominance of global AI industry". Financial Times. 15 October 2017. Retrieved 24 December 2017.*
12.^ *"China enlists start-ups in high-tech arms race". Financial Times. 9 July 2017. Retrieved 24 December 2017.*
13.^ *Metz, Cade (12 February 2018). "As China Marches Forward on A.I., the White House Is Silent". The New York Times. Retrieved 14 February 2018.*
14.^ Jump up to:[a] [b] *"The artificial intelligence race heats up". The Japan Times. 1 March 2018. Retrieved 5 March 2018.*
15.^ *"Robots with Guns: The Rise of Autonomous Weapons Systems". Snopes.com. 21 April 2017. Retrieved 24 December 2017.*
16.^ *Pecotic, Adrian (2019). "Whoever Predicts the Future Will Win the AI Arms Race". Foreign Policy. Retrieved 16 July 2019.*
17.^ *Vincent, James (6 February 2019). "China is worried an AI arms race could lead to accidental war". The Verge. Retrieved 16 July 2019.*
18.^ *"US risks losing AI arms race to China and Russia". CNN. 29 November 2017. Retrieved 24 December 2017.*
19.^ *Davenport, Christian (3 December 2017). "Future wars may depend as much on algorithms as on ammunition, report says". Washington Post. Retrieved 24 December 2017.*
20.^ *"US general warns of out-of-control killer robots". CNN. 18 July 2017. Retrieved 24 December 2017.*
21.^ *Gibbs, Samuel (20 August 2017). "Elon Musk leads 116 experts calling for outright ban of killer robots". The Guardian. Retrieved 24 December 2017.*
22.^ *"'Killer robots': autonomous weapons pose moral dilemma | World| Breakings news and perspectives from around the globe | DW | 14.11.2017". DW.COM. 14 November 2017. Retrieved 12 January 2018.*
23.^ *Parkin, Simon (16 July 2015). "Killer robots: The soldiers that never sleep". BBC. Retrieved 13 January 2018.*
24.^ *"Getting to grips with military robotics". The Economist. 25 January 2018. Retrieved 7 February 2018.*
25.^ *"Autonomous Systems: Infographic". www.siemens.com. Retrieved 7 February 2018.*
26.^ Jump up to:[a] [b] *Kopf, Dan (2018). "China is rapidly closing the US's lead in AI research". Quartz. Retrieved 7 February 2018.*
27.^ *"The battle for digital supremacy". The Economist. 2018. Retrieved 19 March 2018.*
28.^ *"Ban on killer robots urgently needed, say scientists". The Guardian. 13 November 2017. Retrieved 24 December 2017.*
29.^ *Sharkey, Noel (17 August 2007). "Noel Sharkey: Robot wars are a reality". The Guardian. Retrieved 11 January 2018.*

30.^ *Markoff, John (11 November 2014). "Fearing Bombs That Can Pick Whom to Kill". The New York Times. Retrieved 11 January 2018.*
31.^ Jump up to:[a] [b] *"AI Could Revolutionize War as Much as Nukes". WIRED. Retrieved 24 December 2017.*
32.^ *Gibbs, Samuel (20 August 2017). "Elon Musk leads 116 experts calling for outright ban of killer robots". The Guardian. Retrieved 11 January 2018.*
33.^ *"An Open Letter to the United Nations Convention on Certain Conventional Weapons". Future of Life Institute. Retrieved 14 January 2018.*
34.^ Jump up to:[a] [b] *"Rise of the killer machines". Asia Times. 24 November 2017. Retrieved 24 December 2017.*
35.^ *"'Robots are not taking over,' says head of UN body on autonomous weapons". The Guardian. 17 November 2017. Retrieved 14 January 2018.*
36.^ *"Sorry, Banning 'Killer Robots' Just Isn't Practical". WIRED. 22 August 2017. Retrieved 14 January 2018.*
37.^ *McFarland, Matt (14 November 2017). "'Slaughterbots' film shows potential horrors of killer drones". CNNMoney. Retrieved 14 January 2018.*
38.^ Jump up to:[a] [b] Allen, Greg, and Taniel Chan. "Artificial Intelligence and National Security." Report. Harvard Kennedy School, Harvard University. Boston, MA (2017).
39.^ Antebi, Liran. "Who Will Stop the Robots?." Military and Strategic Affairs 5.2 (2013).
40.^ Shulman, C., & Armstrong, S. (2009, July). Arms control and intelligence explosions. In 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July (pp. 2-4).
41.^ *Wheeler, Brian (30 November 2017). "Terrorists 'certain' to get killer robots". BBC News. Retrieved 24 December 2017.*
42.^ *Metz, Cade (15 March 2018). "Pentagon Wants Silicon Valley's Help on A.I." The New York Times. Retrieved 19 March 2018.*
43.^ Jump up to:[a] [b] *"Google 'to end' Pentagon Artificial Intelligence project". BBC News. 2 June 2018. Retrieved 3 June 2018.*
44.^ *Robert O. Work (26 April 2017). "Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)" (PDF). Retrieved 3 June 2018.*
45.^ *"Google employees resign in protest against Air Force's Project Maven". Fedscoop. 14 May 2018. Retrieved 3 June 2018.*
46.^ *Allen, Gregory C. (21 December 2017). "Project Maven brings AI to the fight against ISIS". Bulletin of the Atomic Scientists. Retrieved 3 June2018.*
47.^ *Ethan Baron (3 June 2018). "Google Backs Off from Pentagon Project After Uproar: Report". Military.com. Mercury.com. Retrieved 3 June 2018.*
48.^ *Cheryl Pellerin (21 July 2017). "Project Maven to Deploy Computer Algorithms to War Zone by Year's End". DoD News, Defense Media Activity. United States Department of Defense. Retrieved 3 June 2018.*

Further reading

- Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race", *Foreign Affairs*, vol. 98, no. 3 (May/June 2019), pp. 135–44. "Today's AI technologies are powerful but unreliable. Rules-based systems cannot deal with circumstances their programmers did not anticipate. Learning systems are limited by the data on which they were trained. AI failures have already led to tragedy. Advanced autopilot features in cars, although they perform well in some circumstances, have driven cars without warning into trucks, concrete barriers, and parked cars. In the wrong situation, AI systems go from supersmart to superdumb in an instant. When an enemy is trying to manipulate and hack an AI system, the risks are even greater." (p. 140.)

Chapter 16:

Computers & Information Technology

Terminologies, Acronyms, Idioms and Glossaries!

A

access point:
A device that allows wireless-equipped computers and other devices to communicate with a wired network.
accessibility:
As specified in Section 508 of the 1998 Rehabilitation Act, the process of designing and developing Web sites and other technology that can be navigated and understood by all people, including those with visual, hearing, motor, or cognitive impairments. This type of design also can benefit people with older/slower software and hardware.
ActiveX:
A technology from Microsoft that links desktop applications to the World Wide Web. Using ActiveX tools, interactive web content can be created. Example: In addition to viewing Word and Excel documents from within a browser, additional functionality such as animation, credit card transactions, or spreadsheet calculations.
address:
Identifies the location of an Internet resource. Examples: an e-mail address (sales@dataprise.com); a web address (http://www.dataprise.com); or an internet address (192.168.100.1).
alias:
A short, easy to remember name created for use in place of a longer, more complicated name; commonly used in e-mail applications. Also referred to as a "nickname".
anonymous FTP:
Archive sites where Internet users can log in and download files and programs without a special username or password. Typically, you enter anonymous as a username and your e-mail address as a password.
Anti-Spam
To prevent e-mail spam, both end users and administrators of e-mail

systems use various anti-spam techniques. Some of these techniques have been embedded in products, services and software to ease the burden on users and administrators. No one technique is a complete solution to the spam problem, and each has trade-offs between incorrectly rejecting legitimate e-mail vs. not rejecting all spam, and the associated costs in time and effort. Dataprise Cloud-Based Anti-SPAM e-mail service eliminates the problem almost entirely. Our state-of-the-art solution lets users see only the e-mail they want — and filters out all of the viruses and e-solicitations they don't want before they reach user's computers and mobile devices.

applet:

A program capable of running on any computer regardless of the operating system. Many applets can be downloaded from various sites on the Internet.

application:

A program designed for a specific purpose, such as word processing or graphic design.

ASCII file:

A file that can be opened and read by standard text editor programs (for example, Notepad or Simple Text) on almost any type of computer. Also referred to as "plain text files". Examples: documents saved in ASCII format within word processors like Microsoft Word or WordPerfect; e-mail messages created by a program like Outlook; or HTML files.

AT command set:

An industry standard set of commands beginning with the letters "AT" that are used to control a modem. Example: ATDT tells the modem to dial (D) using touch-tone dialing (T). ATDP specifies pulse dialing (P). Also referred to as the "Hayes Command Set".

attachment:

In this context, a file that is sent along with an e-mail message. ASCII (plain text) files may be appended to the message text, but other types of files are encoded and sent separately (common formats that can be selected include MIME, BinHex, and Uuencode).

authentication:

The process of identifying yourself and the verification that you're who

you say you are. Computers where restricted information is stored may require you to enter your username and password to gain access.

B

backbone:
A term that is often used to describe the main network connections that comprise the Internet or other major network.
bandwidth:
A measurement of the amount of data that can be transmitted over a network at any given time. The higher the network's bandwidth, the greater the volume of data that can be transmitted.
BCP:
Business Continuity Plan, or "BCP," is a set of documents, instructions, and procedures which enable a business to respond to accidents, disasters, emergencies, and/or threats without any stoppage or hindrance in its key operations. It is also called a business resumption plan, disaster recovery plan, or recovery plan.
BI:
Business Intelligence - A recognized industry term for organizational analytics, including historical, current, and predictive views of business operations.
binary file:
A file that cannot be read by standard text editor programs like Notepad or Simple Text. Examples: documents created by applications such as Microsoft Word or WordPerfect or DOS files with the extension ".com" or ".exe".
BinHex:
A common file format for Macintosh computers; it enables a binary file to be transferred over the Internet as an ASCII file. Using a program like Stuffit, a file can be encoded and renamed with an ".hqx" extension. The recipient uses a similar program to decode the file.

bit:
A binary digit (either 0 or 1); it is the most basic unit of data that can be recognized and processed by a computer.
blended learning:
Instruction that combines aspects of both face-to-face (F2F) and online learning experiences. An increasing number of courses at OSU now offer this type of mix.
blog:
Refers to a weblog, a web page that contains journal-like entries and links that are updated daily for public viewing.
bluetooth:
A wireless networking technology that allows users to send voice and data from one electronic device to another via radio waves.
BMP:
Bitmap file; a common image format on Windows computers. Files of this type usually have the suffix ".bmp" as part of their name.
bookmark:
A feature available in certain programs like Internet Explorer, Firefox, and Acrobat Reader; it is a shortcut you can use to get to a particular web page (IE and Firefox) or to a specified location within a document (PDF).
boolean logic:
A form of algebra in which all values are reduced to either true/false, yes/no, on/off, or 1/0.
bounce:
A term applied to an e-mail message when it is returned to you as undeliverable.
bridge:
A device used for connecting two Local Area Networks (LANs) or two segments of the same LAN; bridges forward packets without analyzing or re-routing them.
broadband connection:
A high-speed Internet connection; at present, cable modems and DSL (Digital Subscriber Lines) are the two technologies that are most commonly available to provide such access.

browser:
A program used to access World Wide Web pages. Examples: Firefox, Safari or Internet Explorer.
buffer:
On a multitasking system, a certain amount of RAM that is allocated as a temporary holding area so that the CPU can manipulate data before transferring it to a particular device.
buffered:
Data that is collected but not made immediately available. Compare to a language translator who listens to a whole statement before repeating what the speaker has said rather than providing a word-by-word translation. Example: Streaming media data viewable using a tool like RealMedia Player is buffered.
business continuity:
Business continuity is the activity performed by an organization to ensure that critical business functions will be available to customers, suppliers, regulators, and other entities that must have access to those functions. These activities include many daily chores such as project management, system backups, change control, and help desk. Business Continuity is not something implemented at the time of a disaster; Business Continuity refers to those activities performed daily to maintain service, consistency, and recoverability. To learn more please click **here**.
business continuity plan:
Business Continuity Plan or "BCP" is a set of documents, instructions, and procedures which enable a business to respond to accidents, disasters, emergencies, and/or threats without any stoppage or hindrance in its key operations. It is also called a business resumption plan, disaster recovery plan, or recovery plan. Also see above explanation. To learn more please click **here**.
BYOD:
Bring Your Own Device or "BYOD" is a business and technology policy that allows employees to bring in personal mobile devices and use these devices to access company data, email, etc.
byte:
A group of adjacent binary digits that a computer processes as a unit to form a character such as the letter "C". A byte consists of eight bits.

C

cable modem:
A special type of modem that connects to a local cable TV line to provide a continuous connection to the Internet. Like an analog modem, a cable modem is used to send and receive data, but the difference is that transfer speeds are much faster. A 56 Kbps modem can receive data at about 53 Kbps, while a cable modem can achieve about 1.5 Mbps (about 30 times faster). Cable modems attach to a 10Base-T Ethernet card inside your computer.

cache:
Refers to: 1) a region of computer memory where frequently accessed data can be stored for rapid access; or 2) a optional file on your hard drive where such data also can be stored. Examples: Internet Explorer and Firefox have options for defining both memory and disk cache. The act of storing data for fast retrieval is called "caching".

captcha:
A challenge-response test in the form of an image of distorted text the user must enter that to determine whether the user is human or an automated bot.

carrier services
As authorized agents for the biggest names in the telecommunications industry, Dataprise will deliver the most appropriate and cost-effective carrier solutions for your organization. Dataprise will design, implement and support all of your Data, Internet, Voice and Conferencing solutions. To learn more please click **here**.

case-sensitive:
Generally applies to a data input field; a case-sensitive restriction means lower-case letters are not equivalent to the same letters in upper-case. Example: "data" is not recognized as being the same word as "Data" or "DATA".

CBT:
Computer-Based Training; a type of training in which a student learns a particular application by using special programs on a computer. Sometimes referred to as "CAI" (Computer-Assisted Instruction) or "CBI" (Computer-Based Instruction), although these two terms may also be used to describe a computer program used to assist a teacher or trainer in classroom instruction.
CD-R drive:
A type of disk drive that can create CD-ROMs and audio CDs. CD-R drives that feature multi session recording allow you to continue adding data to a compact disk which is very important if you plan on using the drive for backup.
CD-ROM:
Compact Disk, Read Only Memory; a high-capacity secondary storage medium. Information contained on a CD is read-only. Special CD-ROM mastering equipment available in the OIT Multimedia Lab can be reserved for creating new CDs.
CD-RW, CD-R disk:
A CD-RW disk allows you to write data onto it multiple times instead of just once (a CD-R disk). With a CD-R drive you can use a CD-RW disk just like a floppy or zip disk for backing up files, as well as for creating CD-ROMs and audio CDs.
CGI:
Common Gateway Interface; a mechanism used by most web servers to process data received from a client browser (e.g., a user). CGI scripts contain the instructions that tell the web server what to do with the data.
chat:
Real-time communication between two or more users via networked-connected computers. After you enter a chat (or chat room), any user can type a message that will appear on the monitors of all the other participants. While most ISPs offer chat, it is not supported by OIT. However, the campus CMS (Carmen) supported by TELR does provide the capability for live chat among students participating in online courses.
client:
A program or computer that connects to and requests information from a

server. Examples: Internet Explorer or Firefox. A client program also may be referred to as "client software" or "client-server software".

client-server technology:

Refers to a connection between networked computers in which the services of one computer (the server) are requested by the other (the client). Information obtained is then processed locally on the client computer.

cloud:

(See below): a common shorthand for a provided cloud computing service (or even an aggregation of all existing cloud services) is "The Cloud".To learn more please click **here**.

cloud computing:

A general term used to describe Internet services such as social networking services (e.g., Facebook and Twitter), online backup services, and applications that run within a Web browser. Could computing also includes computer networks that are connected over the Internet for server redundancy or cluster computing purposes.

CMS:

'Content Management System' is the collection of procedures used to manage work flow in a collaborative environment. In a CMS, data can be defined as nearly anything: documents, movies, pictures, phone numbers, scientific data, and so forth. CMSs are frequently used for storing, controlling, revising, semantically enriching, and publishing documentation. Serving as a central repository, the CMS increases the version level of new updates to an already existing file. Version control is one of the primary advantages of a CMS.

compress:

The process of making a file smaller so that it will save disk space and transfer faster over a network. The most common compression utilities are Winrar for PC or compatible computers (.zip files) and or Stuffit (.sit files) for Macintosh computers.

connect:

A term that commonly refers to accessing a remote computer; also a message that appears at the point when two modems recognize each other.

cookie:

A small piece of information you may be asked to accept when connecting to certain servers via a web browser. It is used throughout your session as a

means of identifying you. A cookie is specific to, and sent only to the server that generated it.

courseware:
Software designed specifically for use in a classroom or other educational setting.

CPU:
Central processing unit; the part of a computer that oversees all operations and calculations.

CSP:
Cloud Service Provider; a business model for providing cloud services.

CSS:
Cascading Style Sheet; A set of rules that define how web pages are displayed using CSS, designers can create rules that define how page

cursor:
A special symbol that indicates where the next character you type on your screen will appear. You use your mouse or the arrow keys on your keyboard to move the cursor around on your screen.

cyberspace:
A term describing the world of computers and the society that uses them

D

DaaS:
Desktop-as-a-Service - Also called virtual desktop or hosted desktop services, it is the outsourcing of a virtual desktop infrastructure (VDI) to a third- party service provider. To learn more please click **here**.

daemon:
A special small program that performs a specific task; it may run all the time watching a system, or it can take action only when a task needs to be performed. Example: If an e-mail message is returned to you as undeliverable, you may receive a message from the mailer daemon.

database:
A collection of information organized so that a computer application can

quickly access selected information; it can be thought of as an electronic filing system. Traditional databases are organized by fields, records (a complete set of fields), and files (a collection of records). Alternatively, in a Hypertext database, any object (e.g., text, a picture, or a film) can be linked to any other object.

data center:

A data center (data centre / datacentre / datacenter) is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and security devices.

decompress:

Opposite of compressing a file; the process of restoring the file to its original size and format. The most common programs for decompressing files are Winrar for PC and compatible computers (.zip files) and Stuffit Expander (.sit files) for Macintosh computers.

defragmentation:

The process of rewriting parts of a file to contiguous sectors on a hard drive to increase the speed of access and retrieval.

degauss:

A process used to remove magnetism from a computer monitors. Note flat-panel displays do not have a degauss button since magnetism doesn't build up in them.

desktop:

On computers like IBM PC or compatibles and Macintoshes, the backdrop where windows and icons for disks and applications reside.

DHCP:

Dynamic Host Configuration Protocol; a protocol that lets a server on a local network assign temporary IP addresses to a computer or other network devices.

dialog box:

Sometimes referred to as a window; on a graphical user interface system, an enclosed area displayed by a program or process to prompt a user for entry of information in one or more boxes (fields).

Dial-Up Adapter:
A network component within Windows that enables you to connect to a dial up server via a modem. Users running dial-up connections on Windows computers must have Dial-Up Adapter installed and properly configured.

dial up connection:
A connection from your computer that goes through a regular telephone line. You use special communications software to instruct your modem to dial a number to access another computer system or a network. May also be referred to as "dial up networking".

digital asset:
Intellectual content which has been digitized and can be referenced or retrieved online; for example, PowerPoint slides, audio or video files, or files created in a word processing application, etc.

digitize:
Sometimes referred to as digital imaging; the act of translating an image, a sound, or a video clip into digital format for use on a computer. Also used to describe the process of converting coordinates on a map to x,y coordinates for input to a computer. All data a computer processes must be digitally encoded as a series of zeroes and ones.

DIMM:
Dual In-line Memory Module; a small circuit board that can hold a group of memory chips. A DIMM is capable of transferring 64 bits instead of the 32 bits each SIMM can handle. Pentium processors require a 64-bit path to memory so SIMMs must be installed two at a time as opposed to one DIMM at a time.

directory:
An area on a disk that contains files or additional divisions called "subdirectories" or "folders". Using directories helps to keep files organized into separate categories, such as by application, type, or usage.

disaster recovery
Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of technology infrastructure critical to an organization after a natural or human-induced disaster. Disaster recovery is a subset of business continuity. While business continuity involves planning for keeping all aspects of a business functioning in the

midst of disruptive events, disaster recovery focuses on the IT or technology systems that support business functions. Dataprise's specialist Disaster Recovery Consulting Team can help you devise a near bulletproof Disaster Recovery Plan, so that you can have total piece of mind that your critical systems and processes are safe, and/or can recover from any potential data loss situation. To learn more please click **here**.

disaster recovery planning

Also referred to as "DRP". Please see above explanation.

discussion group:

Another term for an online newsgroup or forum.

distance education:

May also be referred to as "online learning" or "eLearning." A means of instruction that implies a course instructor and students are separated in space and perhaps, in time. Interaction may be synchronous (facilitated) or asynchronous (self-paced). Students can work with various course materials, or they may use tools like chat or discussion groups to collaborate on projects.

distance learning:

The goal of distance education; distance learning and distance education are often used interchangeably.

dither:

A means by which the illusion of new colors and shades is created by varying the pattern of dots; the more dither patterns a device or program supports, the more shades of gray it can represent. Also referred to as halftoning in the context of printing.

DNS:

Domain Name System; a service for accessing a networked computer by name rather than by numerical, (IP) address.

domain:

Part of an Internet address. The network hierarchy consists of domains and subdomains. At the top are a number of major categories (e.g., com, edu, gov); next are domains within these categories (e.g., ohio-state); and then there are subdomains. The computer name is at the lowest level of the hierarchy.

download:
The process of transferring one or more files from a remote computer to your local computer. The opposite action is upload.
dpi:
Dots per inch; a measure of a printer's resolution. The higher the number, the better the print quality. A minimum of 300 dpi usually is required for professional quality printing.
DRaaS:
Disaster Recovery as a Service; a service that helps recover data in the event of a server failure or natural disaster.
drag and drop:
The act of clicking on one icon and moving it on top of another icon to initiate a specific action. Example: Dragging a file on top of a folder to copy it to a new location.
DSL:
Digital Subscriber Line; an always on broadband connection over standard phone lines.
DVD:
Digital video disk; a type of compact disc that holds far more information than the CD-ROMs that are used for storing music files. A DVD can hold a minimum of 4.7 GB, enough for a full-length movie. MPEG-2 is used to compress video data for storage on a DVD. DVD drives are backward-compatible and can play CD-ROMs.
DVD-RW, DVD-R disk:
A DVD-RW disk allows you to write data onto it multiple times instead of just once like on a DVD-R disk. A DVD disk can hold a minimum of 4.7GB which is enough to store a full-length movie. Other uses for DVDs include storage for multimedia presentations that include both sound and graphics.

E

EAP:
Extensible Authentication Protocol; a general protocol for authentication that also supports multiple authentication methods.
EGA:
Extended Graphics Adapter; a card (or board) usually found in older PCs that enables the monitor to display 640 pixels horizontally and 350 vertically.
eLearning:
Electronic learning; applies to a wide scope of processes including Web-based learning, computer-based instruction, virtual classrooms, and digital collaboration. Content may be delivered in a variety of ways including via the Internet, satellite broadcast, interactive TV, and DVD- or CD-ROMs.
e-mail:
Electronic mail; the exchange of messages between users who have access to either the same system or who are connected via a network (often the Internet). If a user is not logged on when a new message arrives, it is stored for later retrieval.
e-mail archiving:
Email archiving is typically a stand-alone IT application that integrates with an enterprise email server, such a Microsoft Exchange. In addition to simply accumulating email messages, these applications index and provide quick, searchable access to archived messages independent of the users of the system, using different technical methods of implementation. The reasons a company may opt to implement an email archiving solution include protection of mission critical data, record retention for regulatory requirements or litigation, and reducing production email server load. Dataprise Cloud-based e-mail archiving service offers you the latest storage technologies in a secure, redundant and easy-to-use format. We take care of all the fine details, from configuring our archiving software to automatically transferring the files to our secure remote servers.

emoticon:
A combination of keyboard characters meant to represent a facial expression. Frequently used in electronic communications to convey a particular meaning, much like tone of voice is used in spoken

communications. Examples: the characters :-) for a smiley face or ;-) for a wink.

emulation:
Refers to the ability of a program or device to imitate another program or device; communications software often include terminal emulation drivers to enable you to log on to a mainframe. There also are programs that enable a Mac to function as a PC.

encryption:
The manipulation of data to prevent accurate interpretation by all but those for whom the data is intended.

EPS:
Encapsulated PostScript; a graphics format that describes an image in the PostScript language.

Ethernet:
A popular network technology that enables data to travel at 10 megabits per second. Campus microcomputers connected to a network have Ethernet cards installed that are attached to Ethernet cabling. An Ethernet connection is often referred to as a "direct connection" and is capable of providing data transmission speeds over 500 Kbps.

Ethernet card:
An adapter card that fits into a computer and connects to Ethernet cabling; different types of adaptor cards fit specific computers. Microcomputers connected to the campus network have some type of Ethernet card installed. Example: computers in campus offices or in dorms rooms wired for ResNet. Also referred to as "Ethernet adapter".

expansion card:
Also referred to as an expansion board; a circuit board you can insert into a slot inside your computer to give it added functionality. A card can replace an existing one or may be added in an empty slot. Some examples include sound, graphics, USB, Firewire, and internal modem cards.

extension:
A suffix preceded by a period at the end of a filename; used to describe the file type. Example: On a Windows computer, the extension ".exe" represents an executable file.

F

female connector:
A cable connector that has holes and plugs into a port or interface to connect one device to another.
field:
A single piece of information within a database (e.g., an entry for name or address). Also refers to a specific area within a dialog box or a window where information can be entered.
file:
A collection of data that has a name (called the filename). Almost all information on a computer is stored in some type of file. Examples: data file (contains data such as a group of records); executable file (contains a program or commands that are executable); text file (contains data that can be read using a standard text editor).
filter:
Refers to: 1) a program that has the function of translating data into a different format (e.g., a program used to import or export data or a particular file); 2) a pattern that prevents non-matching data from passing through (e.g., email filters); and 3) in paint programs and image editors, a special effect that can be applied to a bit map.
finger:
A type of directory service on many UNIX systems. Queries take the format firstname_lastname (e.g., jane_doe) or for more complete information,=firstname.lastname (e.g.,=jane_doe).
firewall:
A method of preventing unauthorized access to or from a particular network; firewalls can be implemented in both hardware and software, or both.
FireWire:
A way to connect different pieces of equipment so they can quickly and easily share information. FireWire (also referred to as IEEE1394 High

Performance Serial Bus) is very similar to USB. It preceded the development of USB when it was originally created in 1995 by Apple. FireWire devices are hot pluggable, which means they can be connected and disconnected any time, even with the power on. When a new FireWire device is connected to a computer, the operating system automatically detects it and prompts for the driver disk (thus the reference "plug-and play").

flash drive:

A small device that plugs into computer's USB port and functions as a portable hard drive.

flash memory:

A type of memory that retains information even after power is turned off; commonly used in memory cards and USB flash drives for storage and transfer of data between computers and other digital products.

folder:

An area on a hard disk that contains a related set of files or alternatively, the icon that represents a directory or subdirectory.

font:

A complete assortment of letters, numbers, and symbols of a specific size and design. There are hundreds of different fonts ranging from businesslike type styles to fonts composed only of special characters such as math symbols or miniature graphics.

frames:

A feature of some web browsers that enables a page to be displayed in separate scrollable windows. Frames can be difficult to translate for text-only viewing via ADA guidelines, so their use is increasingly being discouraged.

freeware:

Copyrighted software available for downloading without charge; unlimited personal usage is permitted, but you cannot do anything else without express permission of the author. Contrast to shareware; copyrighted software which requires you to register and pay a small fee to the author if you decide to continue using a program you download.

fragmentation:

The scattering of parts of the same disk file over different areas of a disk; fragmentation occurs as files are deleted and new ones are added.

FTP:
File Transfer Protocol; a method of exchanging files between computers via the Internet. A program like WS_FTP for IBM PC or compatibles or Fetch for Macintosh is required. Files can contain documents or programs and can be ASCII text or binary data.

G

GIF:
Graphics Interchange Format; a format for a file that contains a graphic or a picture. Files of this type usually have the suffix ".gif" as part of their name. Many images seen on web pages are GIF files.

gigabyte (Gig or GB):
1024 x 1024 x 1024 (2 to the 30th power) bytes; it's usually sufficient to think of a gigabyte as approximately one billion bytes or 1000 megabytes.

GPS:
Global Positioning System; a collection of Earth-orbiting satellites. In a more common context, GPS actually refers to a GPS receiver which uses a mathematical principle called "trilateration" that can tell you exactly where you are on Earth at any moment.

greyware
Greyware (or grayware) refers to a malicious software or code that is considered to fall in the "grey area" between normal software and a virus. Greyware is a term for which all other malicious or annoying software such as adware, spyware, trackware, and other malicious code and malicious shareware fall under.

GUI:
Graphical user interface; a mouse-based system that contains icons, drop-down menus, and windows where you point and click to indicate what you want to do. All new Windows and Macintosh computers currently being sold utilize this technology.

H

handshaking:
The initial negotiation period immediately after a connection is established between two modems. This is when the modems agree about how the data will be transmitted (e.g., error correction, packet size, etc.). The set of rules they agree on is called the protocol.
hard disk:
A storage device that holds large amounts of data, usually in the range of hundreds to thousands of megabytes. Although usually internal to the computer, some types of hard disk devices are attached separately for use as supplemental disk space. "Hard disk" and "hard drive" often are used interchangeably but technically, hard drive refers to the mechanism that reads data from the disk.
hardware:
The physical components of a computer including the keyboard, monitor, disk drive, and internal chips and wiring. Hardware is the counterpart of software.
header:
The portion of an e-mail message or a network newsgroup posting that precedes the body of the message; it contains information like who the message is from, its subject, and the date. A header also is the portion of a packet that precedes the actual data and contains additional information the receiver will need.
help desk:
A help desk is an information and assistance resource that troubleshoots problems with computers or similar products. Corporations often provide help desk support their employees and to their customers via a toll-free number, website and/or e-mail. Dataprise offers 3 types of help desk service: 24 x 7 Support365™, Oustsourced and private labeled.
helper application:
A program used for viewing multimedia files that your web browser

cannot handle internally; files using a helper application must be moved to your computer before being shown or played. Contrast to a plug-in which enables you to view the file over the Internet without first downloading it.

home page:

A document you access using a web browser like Firefox or Internet Explorer. It usually refers to the first page of a particular web site; it also is the page that automatically loads each time you start your browser.

host:

A computer accessed by a user working at a remote location. Also refers to a specific computer connected to a TCP/IP network like the Internet.

HTML:

HyperText Markup Language; a language used for creating web pages. Various instructions and sets of tags are used to define how the document will look.

HTTP:

HyperText Transfer Protocol; a set of instructions that defines how a web server and a browser should interact. Example: When you open a location (e.g., enter a URL) in your browser, what actually happens is an HTTP command is sent to the web server directing it to fetch and return the requested web page.

hyperlink:

Connects one piece of information (anchor) to a related piece of information (anchor) in an electronic document. Clicking on a hyperlink takes you to directly to the linked destination which can be within the same document or in an entirely different document. Hyperlinks are commonly found on web pages, word documents and PDF files.

hypertext:

Data that contains one or more links to other data; commonly seen in web pages and in online help files. Key words usually are underlined or highlighted. Example: If you look for information about "Cats" in a reference book and see a note that says "Refer also to Mammals" the two topics are considered to be linked. In a hypertext file, you click on a link to go directly to the related information.

Hypervisor:

A hypervisor, also called virtual machine manager (VMM), is one of many hardware virtualization techniques that allow multiple operating systems,

termed guests, to run concurrently on a host computer. It is so named because it is conceptually one level higher than a supervisory program. The hypervisor presents to the guest operating systems a virtual operating platform and manages the execution of the guest operating systems. Multiple instances of a variety of operating systems may share the virtualized hardware resources. Hypervisors are installed on server hardware whose only task is to run guest operating systems. Non-hypervisor virtualization systems are used for similar tasks on dedicated server hardware, but also commonly on desktop, portable and even handheld computers.

IaaS:

Infrastructure as a Service; In the most basic cloud-service model, providers of IaaS offer computers - physical or (more often) virtual machines - and other resources.To learn more please click **here**.

icon:

On a system like Windows or Macintosh that uses a graphical user interface (GUI), a small picture or symbol that represents some object or function. Examples: a file folder for a directory; a rectangle with a bent corner for a file; or a miniature illustration for a program.

ICS:

Internet Connection Sharing; a feature in Windows that when enabled, allows you to connect computer on your home network to the Internet via one computer.

IEEE 1394 port:

An interface for attaching high-speed serial devices to your computer; IEEE 1394 connectors support plug and play.

image map:

A graphic overlay that contains more than one area (or hot spot) which is clickable and links to another web page or anchor. Image maps provide an alternative to text links for directing the user to additional information.

IMAP:
Internet Message Access Protcol. A method of accessing e-mail messages on a server without downloading them to your local hard drive; it is the main difference between IMAP and POP3 which requires messages to be downloaded to a user's hard drive before the message can be read.
Internet:
A worldwide network based on the TCP/IP protocol that can connect almost any make or model of popular computers from micros to supercomputers. Special programs called "clients" enable users with a network connection to do things like process e-mail or browse web sites using the familiar interface of a desktop computer.
Internet Domain Management:
For a comprehensive overview of Dataprise's cloud and Internet Domain Management services.
Internet Explorer:
A client program from Microsoft that comes pre installed on most new PC or compatible computers; enables you to browse the World Wide Web.
Internet radio:
An audio broadcasting service transmitted via the Internet; broadcasts consist of a continuous stream. A drawback is the inability to control selection as you can when listening to traditional radio broadcasting.
IP address:
Internet Protocol address; every computer connected to the Internet has a unique identifying number. Example: 192.168.100.2.
IRC:
Internet Relay Chat; a system that enables two or more Internet users to conduct online discussions in real time.
IRQ:
Interrupt request; refers to a number associated with a serial port on an PC or compatible computer. It usually can be changed by flipping a dip switch. Occasionally, when you're using a modem connect to the Internet, you may need to adjust the IRQ number assigned to the serial port which connects the modem to avoid conflicts with another device like your mouse.

ISP:
Internet Service Provider; an organization or company that provides Internet connectivity.

IT Assessment:
An IT Assessment is the practice of gathering information on part or whole of a IT network infrastructure, and then presented in a detailed report. This report typically analyzes the current state or health of technology or services and identifies areas needing improvement or prepare for a some type of system or application upgrade. An IT Assessment can be performed in-house or outsourced to an IT vendor. Dataprise has developed a comprehensive assessment process that includes conducting thorough, in-depth reviews all of your critical technology areas, evaluating them against best practices and then providing you with a roadmap to better leverage your IT as a competitive advantage.
Independent Verification and Validation (IV&V) is the process of checking that a project, service, or system meets specifications and that it fulfills its intended purpose. If you've recently implemented a new technology solution, you may want an independent party to assess the quality of the work.

J

Java:
A general purpose programming language commonly used in conjunction with web pages that feature animation. Small Java applications are called Java applets; many can be downloaded and run on your computer by a Java-compatible browser like Firefox or Internet Explorer.

JavaScript:
A publicly available scripting language that shares many of the features of Java; it is used to add dynamic content (various types of interactivity) to web pages.

JPEG:
Joint Photographic Experts Group; a graphics format which compresses an image to save space. Most images imbedded in web pages are GIFs, but sometimes the JPEG format is used (especially for detailed graphics or photographs). In some cases, you can click on the image to display a larger version with better resolution.
justified:
A word processing format in which text is formatted flush with both the left and right margins. Other options include left justified (text is lined up against the left margin) and right justified (text is lined up against the right margin).

K

K:
An abbreviation for kilobyte; it contains 1,024 bytes; in turn 1,024 kilobytes is equal to one megabyte.
Kbps:
Kilobits per second; a measure of data transfer speed; one Kbps is 1,000 bits per second. Example: a 28.8 Kbps modem.
Kerberos:
An authentication system developed at the Massachusetts Institute of Technology (MIT); it enables the exchange of private information across an open network by assigning a unique key called a "ticket" to a user requesting access to secure information.
kerning:
The amount of space between characters in a word; in desktop publishing, it is typically performed on pairs of letters or on a short range of text to fine-tune the character spacing.
keyword:
Most often refers to a feature of text editing and database management systems; a keyword is an index entry that correlates with a specific record or document.

kilobyte (K, KB, or Kb):
1,024 (2 to the 10th power) bytes; often used to represent one thousand bytes. Example: a 720K diskette can hold approximately 720,000 bytes (or characters).
knowledge base:
A database where information common to a particular topic is stored online for easy reference; for example, a frequently-asked questions (FAQ) list may provide links to a knowledge base.

L

LAN:
Local area network; a network that extends over a small area (usually within a square mile or less). Connects a group of computers for the purpose of sharing resources such as programs, documents, or printers. Shared files often are stored on a central file server.
laser printer:
A type of printer that produces exceptionally high quality copies. It works on the same principle as a photocopier, placing a black powder onto paper by using static charge on a rolling drum.
leading:
The vertical space between lines of text on a page; in desktop publishing, you can adjust the leading to make text easier to read.
learning management system (LMS):
Software used for developing, using, and storing course content of all types. Information within a learning management system often takes the form of learning objects (see "learning object" below).
learning object:
A chunk of course content that can be reused and independently maintained. Although each chunk is unique in its content and function, it must be able to communicate with learning systems using a standardized method not dependent on the system. Each chunk requires a description to facilitate search and retrieval.

link:
Another name for a hyperlink.
LINUX:
An open-source operating system that runs on a number of hardware platforms including PCs and Macintoshes. Linux is freely available over the Internet.
ListProcessor:
A program that manages electronic mailing lists; OIT is responsible for the ListProcessor software and also handles requests from the OSU community or new mailing lists.
LISTSERV, Listserver:
An electronic mailing list; it provides a simple way of communicating with a large number of people very quickly by automating the distribution of electronic mail. At OSU, mailing lists are used not only for scholarly communication and collaboration, but also as a means of facilitating and enhancing classroom education.
log in, log on:
The process of entering your username and password to gain access to a particular computer; e.g., a mainframe, a network or secure server, or another system capable of resource sharing.

M

MaaS:
Metal-as-a-Service; The dynamic provisioning and deployment of whole physical servers, as opposed to the provisioning of virtual machines.
MAC:
Media Access Control; The hardware address of a device connected to a shared network.
Macintosh:
A personal computer introduced in the mid-1980s as an alternative to the IBM PC. Macintoshes popularized the graphical user interface and the 3 1/2 inch diskette drive.

mail server:
A networked computer dedicated to supporting electronic mail. You use a client program like Microsoft Outlook for retrieving new mail from the server and for composing and sending messages.
mailing list:
A collection of e-mail addresses identified by a single name; mailing lists provide a simple way of corresponding with a group of people with a common interest or bond. There are two main types of lists: 1) one you create within an e-mail program like Outlook that contains addresses for two or more individuals you frequently send the same message; and 2) a Listserve type that requires participants to be subscribed (e.g., a group of collaborators, a class of students, or often just individuals interested in discussing a particular topic).
main memory:
The amount of memory physically installed in your computer. Also referred to as "RAM".
mainframe:
A very large computer capable of supporting hundreds of users running a variety of different programs simultaneously. Often the distinction between small mainframes and minicomputers is vague and may depend on how the machine is marketed.
male connector:
A cable connector that has pins and plugs into a port or interface to connect one device to another.
malware:
Software programs designed to damage or do other unwanted actions on a computer; common examples of malware include viruses, worms, trojan horses, and spyware.
Managed Workstations:
A Managed Workstation reduces downtime, improves maintenance, increases productivity and data security through an effective blend of Help Desk and on-site support and centralized deployment of software patches and virus protection updates. Dataprise can deliver expert support at the workstation level for all of your users, at any location. Using our DesktopStreaming™ live online support technology, our highly qualified certified technical staff, working remotely, are able to see exactly what is

happening on a user's computer screen — allowing us to quickly isolate issues and begin remediation.

MAPI:

Messaging Application Programming Interface; a system built into Microsoft Windows that enables different e-mail programs to interface to distribute e-mail. When both programs are MAPI-enabled, they can share messages.

MDM:

Mobile Device Management; Any routine or tool intended to distribute applications, data, and configuration settings to mobile communications devices. The intent of MDM is to optimize the functionality and security of a mobile communications network. MDM must be part of a coherent BYOD strategy.

megabyte (Meg or MB):

1,024 x 1,024 (2 to the 20th power) bytes; it's usually sufficient to think of a megabytes as one million bytes.

MHz or mHz:

Megahertz; a measurement of a microprocessor's speed; one MHz represents one million cycles per second. The speed determines how many instructions per second a microprocessor can execute. The higher the megahertz, the faster the computer.

menu:

In a graphical user interface, a bar containing a set of titles that appears at the top of a window. Once you display the contents of a menu by clicking on its title, you can select any active command (e.g., one that appears in bold type and not in a lighter, gray type).

Microsoft Exchange:

Microsoft Exchange Server is the server side of a client–server, collaborative application product developed by Microsoft. It is part of the Microsoft Servers line of server products and is used by enterprises using Microsoft infrastructure products. Exchange's major features consist of electronic mail, calendaring, contacts and tasks; support for mobile and web-based access to information; and support for data storage. Dataprise has a 100% hosted Exchange solution that includes clustered and redundant Microsoft Exchange servers that provide more then enough

horsepower to support all of your organization’s messaging needs. And we handle the entire set-up and configuration for you.

Microsoft Windows:

A group of operating systems for PC or compatible computers; Windows provides a graphical user interface so you can point and click to indicate what you want to do.

MIME:

Multipurpose Internet Mail Extensions; a protocol that enables you to include various types of files (text, audio, video, images, etc.) as an attachment to an e-mail message.

modem:

A device that enables a computer to send and receive information over a normal telephone line. Modems can either be external (a separate device) or internal (a board located inside the computer's case) and are available with a variety of features such as error correction and data compression.

moderator:

A person who reviews and has the authority to block messages posted to a supervised or "moderated" network newsgroup or online community.

monitor:

The part of a computer that contains the screen where messages to and from the central processing unit (CPU) are displayed. Monitors come in a variety of sizes and resolutions. The higher the number of pixels a screen is capable of displaying, the better the resolution. Sometimes may be referred to as a CRT.

mouse:

A handheld device used with a graphical user interface system. Common mouse actions include: 1) clicking the mouse button to select an object or to place the cursor at a certain point within a document; 2) double-clicking the mouse button to start a program or open a folder; and 3) dragging (holding down) the mouse button and moving the mouse to highlight a menu command or a selected bit of text.

MPEG:

Motion Picture Experts Group; a high quality video format commonly used for files found on the Internet. Usually a special helper application is required to view MPEG files.

MRB:
Managed Remote Back Up; a service that provides users with a system for the backup, storage, and recovery of data using cloud computing.
MSP:
Managed Service Provider; A business model for providing information-technology services.To learn more please click **here**.
multimedia:
The delivery of information, usually to a personal computer, in a combination of different formats including text, graphics, animation, audio, and video.
multitasking:
The ability of a CPU to perform more than one operation at the same time; Windows and Macintosh computers are multitasking in that each program that is running uses the CPU only for as long as needed and then control switches to the next task.

N

NaaS:
Network as a Service; a category of cloud services that provides users with the capability of where the capability provided to the cloud service user is to usinge network/transport connectivity services and/or inter-cloud network connectivity services.
nameserver:
A computer that runs a program for converting Internet domain names into the corresponding IP addresses and vice versa.
NAT:
Network Address Translation; a standard that enables a LAN to use a set of IP addresses for internal traffic and a single IP address for communications with the Internet.
network:
A group of interconnected computers capable of exchanging information.

A network can be as few as several personal computers on a LAN or as large as the Internet, a worldwide network of computers.

network adapter:

A device that connects your computer to a network; also called an adapter card or network interface card.

network hub:

A common connection point for devices on a network.

NNTP:

Network News Transport Protocol; the protocol used for posting, distributing, and retrieving network news messages.

network monitoring:

Dataprise Cloud-based Network Monitoring service, can configure and remotely monitor all of your important network systems (e-mail, servers, routers, available disk space, backup applications, critical virus detection, and more). If our system detects a problem, it alerts the Dataprise Technical Support Center, so we can take corrective action. Depending on prearranged instructions from your own network engineers, we'll correct the problem immediately, wait until the next business day or simply notify you of the issue.

network security:

Network security consists of the provisions and policies adopted by a network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and network-accessible resources. Network Security is the authorization of access to data in a network, which is controlled by a network administrator. Dataprise uses state-of-the-art network security techniques while providing authorized personnel access to important files and applications. Every organization's needs are different and hackers are always adapting their techniques, so we are extremely serious about staying up to date with the latest network security tools, threats and industry developments.

O

OCR:
Optical character recognition; the act of using a visual scanning device to read text from hard copy and translate it into a format a computer can access (e.g., an ASCII file). OCR systems include an optical scanner for reading text and sophisticated software for analyzing images.
on-Cloud:
Dataprise realizes that businesses are moving more and more of their critical infrastructure to Cloud-based providers. 'On-Cloud' is currently our own term coined for providing management and support for your Cloud-based systems and processes.
on-site:
At-place-of-work-or-business support, typically provided by a technically qualified individual.
online:
A term that has commonly come to mean "connected to the Internet". It also is used to refer to materials stored on a computer (e.g., an online newsletter) or to a device like a printer that is ready to accept commands from a computer.
OpenType
OpenType is a format for scalable computer fonts. It was built on its predecessor TrueType, retaining TrueType's basic structure and adding many intricate data structures for prescribing typographic behavior. OpenType is a registered trademark of Microsoft Corporation.

P

PaaS:
Platform as a Service, in the PaaS model, cloud providers deliver a computing platform that typically including an operating system, programming language execution environment, database, and web server.
packet:
A unit of transmission in data communications. The TCP/IP protocol

breaks large data files into smaller chunks for sending over a network so that less data will have to be re-transmitted if errors occur.

palette:

The range of colors a computer or an application is able to display. Most newer computers can display as many as 16 million colors, but a given program may use only 256 of them. Also refers to a display box containing a set of related tools within a desktop publishing or graphics design program.

page:

Refers to an HTML document on the World Wide Web or to a particular web site; usually pages contain links to related documents (or pages).

parallel port:

An interface on a computer that supports transmission of multiple bits at the same time; almost exclusively used for connecting a printer. On IBM or compatible computers, the parallel port uses a 25-pin connector. Macintoshes have an SCSI port that is parallel, but more flexible in the type of devices it can support.

password:

A secret combination of characters used to access a secured resource such as a computer, a program, a directory, or a file; often used in conjunction with a username.

PC:

Usually refers to an IBM PC or compatible, or when used generically, to a "personal computer". In a different context, PC also is an abbreviation for "politically correct."

PDA:

Personal Digital Assistant; a small hand-held computer that in the most basic form, allows you to store names and addresses, prepare to-do lists, schedule appointments, keep track of projects, track expenditures, take notes, and do calculations. Depending on the model, you also may be able to send or receive e-mail; do word processing; play MP3 music files; get news, entertainment and stock quotes from the Internet; play video games; and have an integrated digital camera or GPS receiver.

PDF:

Portable Document Format; a type of formatting that enables files to be viewed on a variety computers regardless of the program originally used to

create them. PDF files retain the "look and feel" of the original document with special formatting, graphics, and color intact. You use a special program or print driver (Adobe Distiller or PDF Writer) to convert a file into PDF format.

peer-to-peer:

A type of connection between two computers; both perform computations, store data, and make requests from each other (unlike a client-server connection where one computer makes a request and the other computer responds with information).

Perl:

Practical Extraction and Report Language; a programming language that is commonly used for writing CGI scripts used by most servers to process data received from a client browser.

personality:

A method of setting up a computer or a program for multiple users. Example: In Windows, each user is given a separate "personality" and set of relevant files.

PGP:

Pretty good privacy; a technique for encrypting e-mail messages. PGP uses a public key to give to anyone who sends you messages and a private key you keep to decrypt messages you receive.

Ph:

A type of directory service often referred to as a "phone book". When accessing this type of directory service, follow the directions from the particular site for looking up information.

phishing:

A con that scammers use to electronically collect personal information from unsuspecting users. Phishers send e-mails that appear to come from legitimate websites such as eBay, PayPal, or other banking institutions asking you to click on a link included in the email and then update or validate your information by entering your username and password and often even more information, such as your full name, address, phone number, social security number, and credit card number.

PING:

Packet Internet Groper; a utility used to determine whether a particular

computer is currently connected to the Internet. It works by sending a packet to the specified IP address and waiting for a reply.

pixel:
Stands for one picture element (one dot on a computer monitor); commonly used as a unit of measurement.

plug-in:
A program used for viewing multimedia files that your web browser cannot handle internally; files using a plug-in do not need to be moved to your computer before being shown or played. Contrast to a helper application which requires the file to first be moved to your computer. Examples of plug-ins: Adobe Flash Player (for video and animation) and Quicktime (for streamed files over the Internet).

plug and play:
A set of specifications that allows a computer to automatically detect and configure a device and install the appropriate device drivers.

POP:
Post Office Protocol; a method of handling incoming electronic mail. Example: E-mail programs may use this protocol for storing your incoming messages on a special cluster of servers called pop.service.ohio-state.edu and delivering them when requested.

pop-up blocker:
Any application that disables the pop-up, pop-over, or pop-under ad windows that appear when you use a web browser.

post:
The act of sending a message to a particular network newsgroup.

PostScript:
A page description language primarily used for printing documents on laser printers; it is the standard for desktop publishing because it takes advantage of high resolution output devices. Example: A graphic design saved in PostScript format looks much better when printed on a 600 dpi printer than on a 300 dpi printer.

PostScript fonts:
Called outline or scalable fonts; with a single typeface definition, a PostScript printer can produce many other fonts. Contrast to non-PostScript printers that represent fonts with bitmaps and require a complete set for each font size.

PPP:
Point-to-Point Protocol; a type of connection over telephone lines that gives you the functionality of a direct ethernet connection.
program:
A set of instructions that tells a computer how to perform a specific task.
private cloud
Private cloud (also called internal cloud or corporate cloud) is a term for a proprietary computing architecture that provides hosted services to a limited number of users behind a secure and robust infrastructure. A Dataprise private cloud solution is designed to offer the same features and benefits of shared cloud systems, but removes a number of objections to the cloud computing model including control over enterprise and customer data, worries about security, and issues connected to regulatory compliance. Dataprise Private clouds" are designed to facilitate organizations that needs or wants more control over their data than they can get by using a third-party shared cloud service.
protocol:
A set of rules that regulate how computers exchange information. Example: error checking for file transfers or POP for handling electronic mail.
proxy:
Refers to a special kind of server that functions as an intermediate link between a client application (like a web browser) and a real server. The proxy server intercepts requests for information from the real server and whenever possible, fills the request. When it is unable to do so, the request is forwarded to the real server.
public domain software:
Any non-copyrighted program; this software is free and can be used without restriction. Often confused with "freeware" (free software that is copyrighted by the author).
pull:
Frequently used to describe data sent over the Internet; the act of requesting data from another computer. Example: using your web browser to access a specific page. Contrast to "push" technology when data is sent to you without a specific request being made.

push:
Frequently used to describe data sent over the Internet; the act of sending data to a client computer without the client requesting it. Example: a subscriptions service that delivers customized news to your desktop. Contrast to browsing the World Wide Web which is based on "pull" technology; you must request a web page before it is sent to your computer.

Q

QoS:
Quality of service; is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Quality of service guarantees are important if the network capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP, online games and IP-TV, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication.
QuickTime:
A video format developed by Apple Computer commonly used for files found on the Internet; an alternative to MPEG. A special viewer program available for both IBM PC and compatibles and Macintosh computers is required for playback.

R

RAM:
Random Access Memory; the amount of memory available for use by programs on a computer. Also referred to as "main memory". Example: A computer with 8 MB RAM has approximately 8 million bytes of memory available. Contrast to ROM (read-only memory) that is used to store programs that start your computer and do diagnostics.
record:
A set of fields that contain related information; in database type systems, groups of similar records are stored in files. Example: a personnel file that contains employment information.
registry:
A database used by Windows for storing configuration information. Most 32-bit Windows applications write data to the registry. Although you can edit the registry, this is not recommended unless absolutely necessary because errors could disable your computer.
remote backup:
A remote, online, or managed backup service is a service that provides users with a system for the backup and storage of computer files. Dataprise remote backup solution incorporates automatic data compression and secure data encryption. This means that your critical system data backs up safely and efficiently. For additional peace of mind, our backup service features proprietary dual tapeless backup protection, including fast incremental backup to a secure on-site hard drive and a second backup to our carrier-grade data center. Our remote backup service is completely automated and immensely secure. You'll never have to think about the safety of your data again..
remote desktop:
A Windows feature that allows you to have access to a Windows session from another computer in a different location (XP and later).
remote login:
An interactive connection from your desktop computer over a network or telephone lines to a computer in another location (remote site).
remote monitoring:
See: "network monitoring"
remote support:
RGB:

Red, green, and blue; the primary colors that are mixed to display the color of pixels on a computer monitor. Every color of emitted light can be created by combining these three colors in varying levels.

RJ-45 connector:

An eight-wire connector used for connecting a computer to a local-area network. May also be referred to as an Ethernet connector.

ROM:

Read Only Memory; a special type of memory used to store programs that start a computer and do diagnostics. Data stored in ROM can only be read and cannot be removed even when your computer is turned off. Most personal computers have only a few thousand bytes of ROM. Contrast to RAM (random access or main memory) which is the amount of memory available for use by programs on your computer.

router:

A device used for connecting two Local Area Networks (LANs); routers can filter packets and forward them according to a specified set of criteria.

RTF:

Rich Text Format; a type of document formatting that enables special characteristics like fonts and margins to be included within an ASCII file. May be used when a document must be shared among users with different kinds of computers (e.g., IBM PC or compatibles and Macintoshes).

S

SaaS:

Software as a Service; a software delivery model in which software and associated data are centrally hosted on the cloud. SaaS is typically accessed by users using a thin client via a web browser.

safe mode:

A way of starting your Windows computer that can help you diagnose problems; access is provided only to basic files and drivers.

SAN:

A storage area network (SAN) is a dedicated storage network that provides

access to consolidated, block level storage. SANs primarily are used to make storage devices (such as disk arrays, tape libraries, and optical jukeboxes) accessible to servers so that the devices appear as locally attached to the operating system. A SAN typically has its own network of storage devices that are generally not accessible through the regular network by regular devices.

SATA:

Serial Advanced Technology Attachment or Serial ATA. An interface used to connect ATA hard drives to a computer's motherboard that provides a better, more efficient interface; Serial ATA is likely to replace the previous standard, Parallel ATA (PATA), which has become dated.

satellite transmission:

A method of data transmission; the sender beams data up to an orbiting satellite and the satellite beams the data back down to the receiver.

screen reader:

A software program that translates text on a Web page into audio output; typically used by individuals with vision impairment.

scroll bar:

In a graphical user interface system, the narrow rectangular bar at the far right of windows or dialog boxes. Clicking on the up or down arrow enables you to move up and down through a document; a movable square indicates your location in the document. Certain applications also feature a scroll bar along the bottom of a window that can be used to move from side-to-side.

search engine:

A tool that searches documents by keyword and returns a list of possible matches; most often used in reference to programs such as Google that are used by your web browser to search the Internet for a particular topic.

secure server:

A special type of file server that requires authentication (e.g., entry a valid username and password) before access is granted.

security token:

A small device used to provide an additional level of authorization to access a particular network service; the token itself may be embedded in some type of object like a key fob or on a smart card. Also referred to as an authentication token.

Section 508:
A 1998 amendment to the Workforce Rehabilitation Act of 1973; it states after June 25, 2001, all electronic and information technology developed, purchased, or used by the federal government must be accessible to those with disabilities. Refer to the Section 508 website for more information.
self-extracting file:
A type of compressed file that you can execute (e.g., double-click on the filename) to begin the decompression process; no other decompression utility is required. Example: on IBM PC or compatibles, certain files with an ".exe" extension and on Macintoshes, all files with a ".sea" extension.
serial port:
An interface on a computer that supports transmission of a single bit at a time; can be used for connecting almost any type of external device including a mouse, a modem, or a printer.
server:
A computer that is responsible for responding to requests made by a client program (e.g., a web browser or an e-mail program) or computer. Also referred to as a "file server".
shareware:
Copyrighted software available for downloading on a free, limited trial basis; if you decide to use the software, you're expected to register and pay a small fee. By doing this, you become eligible for assistance and updates from the author. Contrast to public domain software which is not copyrighted or to freeware which is copyrighted but requires no usage fee.
signature:
A file containing a bit of personal information that you can set to be automatically appended to your outgoing e-mail messages; many network newsreaders also have this capability. Large signatures over five lines generally are frowned upon.
SIMM:
Single In-line Memory Module; a small circuit board that can hold a group of memory chips; used to increase your computer's RAM in increments of 1,2, 4, or 16 MB.
SMTP:
Simple Mail Transfer Protocol; a method of handling outgoing electronic mail.

software:
Any program that performs a specific function. Examples: word processing, spreadsheet calculations, or electronic mail.
spam:
Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of spam that involves nearly identical messages sent to numerous recipients by email. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk. Spammers collect email addresses from chatrooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. They also use a practice known as “email appending” or "ending" in which they use known information about their target (such as a postal address) to search for the target's email address. Also see "Anti-Spam".
SSID:
Service Set Identifier; a name that identifies a wireless network.
streaming (streaming media):
A technique for transferring data over the Internet so that a client browser or plug-in can start displaying it before the entire file has been received; used in conjunction with sound and pictures. Example: The Flash Player plug-in from Adobe Systems gives your computer the capability for streaming audio; RealPlayer is used for viewing sound and video.
spyware:
Any software that covertly gathers user information, usually for advertising purposes, through the user's Internet connection.
subdirectory:
An area on a hard disk that contains a related set of files; on IBM PC or compatibles, a level below another directory. On Macintoshes, subdirectories are referred to as folders.

***Support*365®**
Dataprise's unique Support365™ plans offer the best solution for organizations that need comprehensive IT support, that either don’t have the time, skill-set or simply just don’t want the burden of managing an IT person, department, or in some situations – an entire IT division. By choosing Support365™ We make it easier than ever for you to understand, budget, and manage your monthly IT requirements. It's basically like having your own virtual IT department. To learn more click **here**.

SVGA:
Super VGA (Video Graphics Array); a set of graphics standards for a computer monitor that offers greater resolution than VGA. There are several different levels including 800 x 600 pixels, 1024 by 768 pixels, 1280 by 1024 pixels; and 1600 by 1200 pixels. Although each supports a palette of 16 million colors, the number of simultaneous colors is dependent on the amount of video memory installed in the computer.

T

T-1 carrier:
A dedicated phone connection supporting data rates of 1.544Mbits per second; T-1 lines are a popular leased line option for businesses connecting to the Internet and for Internet Service Providers connecting to the Internet backbone. Sometimes referred to as a DS1 line.
T-3 carrier:
A dedicated phone connection supporting data rates of about 43 Mbps; T-3 lines are used mainly by Internet Service Providers connecting to the Internet backbone and for the backbone itself. Sometimes referred to as a DS3 line.
10Base-T:
An adaptation of the Ethernet standard for Local Area Networks that refers to running Ethernet over twisted pair wires. Students planning on using ResNet from a residence hall must be certain to use an Ethernet adapter that is 10Base-T compatible and not BNC (used with 10Base-2 Ethernet systems).
table:
With reference to web design, a method for formatting information on a page. Use of tables and the cells within also provide a way to create columns of text. Use of tables vs frames is recommended for helping to make your web site ADA-compliant.
TCP/IP:
Transmission Control Protocol/Internet Protocol; an agreed upon set of

rules that tells computers how to exchange information over the Internet. Other Internet protocols like FTP, Gopher, and HTTP sit on top of TCP/IP.

Telephony:

Telephony encompasses the general use of equipment to provide voice communication over distances, specifically by connecting telephones to each other. Dataprise's expert team of telecommunication consultants can design and implement a system that is feature rich, simple to use and integrates seamlessly with your existing business applications.

telnet:

A generic term that refers to the process of opening a remote interactive login session regardless of the type of computer you're connecting to.

terminal emulation:

The act of using your desktop computer to communicate with another computer like a UNIX or IBM mainframe exactly as if you were sitting in front of a terminal directly connected to the system. Also refers to the software used for terminal emulation. Examples: the Telnet program for VT100 emulation and QWS3270 (Windows) and TN3270 (Macintosh) for IBM3270 fullscreen emulation.

TIFF:

Tag Image File Format; a popular file format for storing bit-mapped graphic images on desktop computers. The graphic can be any resolution and can be black and white, gray-scale, or color. Files of this type usually have the suffix ".tif" as part of their name.

token:

A group of bits transferred between computers on a token-ring network. Whichever computer has the token can send data to the other systems on the network which ensures only one computer can send data at a time. A token may also refer to a network security card, also known as a hard token.

tool bar:

On a graphical user interface system, a bar near the top of an application window that provides easy access to frequently used options.

Trojan horse:
A harmless-looking program designed to trick you into thinking it is something you want, but which performs harmful acts when it runs.
TrueType:
A technology for outline fonts that is built into all Windows and Macintosh operating systems. Outline fonts are scalable enabling a display device to generate a character at any size based on a geometrical description.
tweet:
An update of 140 characters or less published by a Twitter user meant to answer the question, "What are you doing?" which provides other users with information about you.
Twitter:
A service that allows users to stay connected with each other by posting updates, or "tweets," using a computer or cell phone or by viewing updates posted by other users.
twisted pair cable:
A type of cable that is typically found in telephone jacks; two wires are independently insulated and are twisted around each other. The cable is thinner and more flexible than the coaxial cable used in conjunction with 10Base-2 or 10Base-5 standards. Most Ohio State UNITS telephone jacks have three pairs of wires; one is used for the telephone and the other two can be used for 10Base-T Ethernet connections.
two-factor authentication:
An extra level of security achieved using a security token device; users have a personal identification number (PIN) that identifies them as the owner of a particular token. The token displays a number which is entered following the PIN number to uniquely identify the owner to a particular network service. The identification number for each user is changed frequently, usually every few minutes.

U

UNIX:
A popular multitasking computer system often used as a server for electronic mail or for a web site. UNIX also is the leading operating system for workstations, although increasingly there is competition from Windows NT which offers many of the same features while running on an PC or compatible computer.
upload:
The process of transferring one or more files from your local computer to a remote computer. The opposite action is download.
USB:
Universal Serial Bus; a connector on the back of almost any new computer that allows you to quickly and easily attach external devices such as mice, joysticks or flight yokes, printers, scanners, modems, speakers, digital cameras or webcams, or external storage devices. Current operating systems for Windows and Macintosh computers support USB, so it's simple to install the device drivers. When a new device is connected, the operating system automatically activates it and begins communicating. USB devices can be connected or disconnected at any time.
username:
A name used in conjunction with a password to gain access to a computer system or a network service.
URL:
Uniform Resource Locator; a means of identifying resources on the Internet. A full URL consists of three parts: the protocol (e.g., FTP, gopher, http, nntp, telnet); the server name and address; and the item's path. The protocol describes the type of item and is always followed by a colon (:). The server name and address identifies the computer where the information is stored and is preceded by two slashes (//). The path shows where an item is stored on the server and what the file is called; each segment of the location s preceded by a single slash (/). Examples: The URL for the Dataprise home page is http://www.dataprise.com.
USB port:
An interface used for connecting a Universal Serial Bus (USB) device to computer; these ports support plug and play.
utility:
Commonly refers to a program used for managing system resources such

as disk drives, printers, and other devices; utilities sometimes are installed as memory-resident programs. Example: the suite of programs called Norton Utilities for disk copying, backups, etc.

uuencode:
A method of converting files into an ASCII format that can be transmitted over the Internet; it is a universal protocol for transferring files between different platforms like UNIX, Windows, and Macintosh and is especially popular for sending e-mail attachments.

V

VDI:
Virtual Desktop Infrastructure or "VDI," is a desktop-centric service that hosts users' desktop environments on remote servers and/or blade PCs, which are accessed over a network using a remote display protocol.

virtual classroom:
An online environment where students can have access to learning tools any time. Interaction between the instructor and the class participants can be via e-mail, chat, discussion group, etc.

virtualization:
Virtualization is the creation of a virtual (rather than actual) version of something, such as a hardware platform, operating system, a storage device or network resources. In hardware virtualization, the term host machine refers to the actual machine on which the virtualization takes place; the term guest machine, however, refers to the virtual machine. Likewise, the adjectives host and guest are used to help distinguish the software that runs on the actual machine from the software that runs on the virtual machine. The software or firmware that creates a virtual machine on the host hardware is called a hypervisor or Virtual Machine Monitor.

virtual classroom:
An online environment where students can have access to learning tools any time. Interaction between the instructor and the class participants can be via e-mail, chat, discussion group, etc.

virtual hosting:
Virtual hosting is a method for hosting multiple domain names on a computer using a single IP address. This allows one machine to share its resources, such as memory and processor cycles, to use its resources more efficiently. Dataprise Virtual Hosting provides a high-performance hosting platform for your organization's online presence. Maintained by our specialist support staff and 24x7 active monitoring systems, we work hard to meet all of your hosted Web server needs.

virtual memory:
A technique that enables a certain portion of hard disk space to be used as auxiliary memory so that your computer can access larger amounts of data than its main memory can hold at one time.

virtual reality:
An artificial environment created with computer hardware and software to simulate the look and feel of a real environment. A user wears earphones, a special pair of gloves, and goggles that create a 3D display. Examples: manipulating imaginary 3D objects by "grabbing" them, taking a tour of a "virtual" building, or playing an interactive game.

virus:
A program intended to alter data on a computer in an invisible fashion, usually for mischievous or destructive purposes. Viruses are often transferred across the Internet as well as by infected diskettes and can affect almost every type of computer. Special antivirus programs are used to detect and eliminate them.

VoIP:
Voice over Internet Protocol; a means of using the Internet as the transmission medium for phone calls. An advantage is you do not incur any additional surcharges beyond the cost of your Internet access.

VPN:
Virtual Private Networking; a means of securely accessing resources on a network by connecting to a remote access server through the Internet or other network.

VT100:
A type of terminal emulation required when you open an interactive network connection (telnet) to a UNIX system from your desktop computer.

W

WAIS:
Wide Area Information Server; a program for finding documents on the Internet. Usually found on gopher servers to enable searching text-based documents for a particular keyword.
WAN:
Wide Area Network; a group of networked computers covering a large geographical area (e.g., the Internet).
WAP:
Wireless Application Protocol; a set of communication protocols for enabling wireless access to the Internet.
WEP:
Wired Equivalent Privacy; a security protocol for wireless local area networks defined in the 802.11b standard. WEP provides the same level of security as that of a wired LAN.
wi-fi:
Wireless Fidelity; A generic term from the Wi-Fi Alliance that refers to of any type of 802.11 network (e.g., 802.11b, 802.11a, dual-band, etc.). Products approved as "Wi-Fi Certified" (a registered trademark) are certified as interoperable with each other for wireless communications.
wild card:
A special character provided by an operating system or a particular program that is used to identify a group of files or directories with a similar characteristic. Useful if you want to perform the same operation simultaneously on more than one file. Example: the asterisk (*) that can be used in DOS to specify a groups of files such as *.txt.
window:
On a graphical user interface system, a rectangular area on a display screen. Windows are particularly useful on multitasking systems which allow you to perform a number of different tasks simultaneously. Each task has its own window which you can click on to make it the current process.

Contrast to a "dialog box" which is used to respond to prompts for input from an application.

Windows:

A casual way of referring to the Microsoft Windows operating systems.

wireless (networking)

The ability to access the Internet without a physical network connection. Devices such as cell phones and PDAs that allow you to send and receive e-mail use a wireless Internet connection based on a protocol called WAP (Wireless Application Protocol). At this point, web sites that contain wireless Internet content are limited, but will multiply as the use of devices relying on WAP increases.

wizard:

A special utility within some applications that is designed to help you perform a particular task. Example: the wizard in Microsoft Word that can guide you through creating a new document.

WLAN:

Wireless Local Area Network; the computers and devices that make up a wireless network.

workstation:

A graphical user interface (GUI) computer with computing power somewhere between a personal computer and a minicomputer (although sometimes the distinction is rather fuzzy). Workstations are useful for development and for applications that require a moderate amount of computing power and relatively high quality graphics capabilities.

World Wide Web:

A hypertext-based system of servers on the Internet. Hypertext is data that contains one or more links to other data; a link can point to many different types of resources including text, graphics, sound, animated files, a network newsgroup, a telnet session, an FTP session, or another web server. You use a special program called a "browser" (e.g., Firefox or Internet Explorer) for viewing World Wide Web pages. Also referred to as "WWW" or "the web".

worm:

A program that makes copies of itself and can spread outside your operating system worms can damage computer data and security in much the same way as viruses.

WPA:
Wi-Fi Protected Access; a standard designed to improve on the security features of WEP.
WWW:
An abbreviation for World Wide Web.
WYSIWYG:
What You See Is What You Get; a kind of word processor that does formatting so that printed output looks identical to what appears on your **screen.**

X

X2:
A technology that enables data transmission speeds up to 56 Kbps using regular telephone service that is connected to switching stations by high-speed digital lines. This technology affects only transmissions coming into your computer, not to data you send out. In addition, your ISP must have a modem at the other end that supports X2.
XHTML:
Extensible Hypertext Markup Language. A spinoff of the hypertext markup language (HTML) used for creating Web pages. It is based on the HTML 4.0 syntax, but has been modified to follow the guidelines of XML and is sometimes referred to as HTML 5.0.
XML:
Extensible Markup Language; A markup language for coding web documents that allows designers to create their own customized tags for structuring a page.

Z

zero-day:
zero-day (or zero-hour or day zero) attack, threat or virus is a computer threat that tries to exploit computer application vulnerabilities that are unknown to others or the software developer, also called zero-day vulnerabilities. Zero-day exploits (actual software that uses a security hole to carry out an attack) are used or shared by attackers before the developer of the target software knows about the vulnerability.

zip:
A common file compression format for PC or compatibles; the utility WinZip or Winrar is used for compressing and decompressing files. Zipped files usually end with a ".zip" file extension. A special kind of zipped file is self-extracting and ends with a ".exe" extension. Macintosh OSX also supports the .zip format and has tools that can compress and decompress zip files.

Zip drive:
A high capacity floppy disk drive from Iomega Corporation; the disks it uses are a little bit larger than a conventional diskette and are capable of holding 100 MB or 250 MB of data.

zoom:
The act of enlarging a portion of an onscreen image for fine detail work; most graphics programs have this capability.

Information Technology Glossary

Access

Ability to use, create, modify, view, or otherwise manipulate information on a system.

Access control

Access control is the means by which the ability to use, create, modify, view, etc., is explicitly enabled or restricted in some way (usually through physical and system-based controls).

Account

The combination of user name and password that provides an individual, group, or service with access to a computer system or computer network.

Administrative/special access account

Privileged account that impacts access to an information system or that allows circumvention of controls in order to administer the information system.

Anti-malware software

Any software package that detects and/or removes malicious code. This can include anti-virus software and spyware protection.

Areas containing critical infrastructure

Facilities that contain information systems. The ASU Data Center is an example.

ASU in security policy

The population of a group whose membership is determined by each individual's responsibility to fulfill university policy and state regulatory

requirements. If the requirement must be fulfilled by all employees, then that is the membership of “ASU” for that requirement.

Authentication

The process of confirming a claimed identity. All forms of authentication are based on something you know, something you have, or something you are.

- ‘Something you know’ is some form of information that you can recognize and keep to yourself, such as a personal identification number (PIN) or password.
- ‘Something you have’ is a physical item you possess, such as a photo ID or a security token.
- ‘Something you are’ is a human characteristic considered to be unique, such as a fingerprint, voice tone, or retinal pattern.

Authorization

The act of granting permission for someone or something to conduct an act. Even when identity and authentication have indicated who someone is, authorization may be needed to establish what actions are permitted.

Availability

The requirement that an asset or resource be accessible to authorized persons, entities, or devices.

Backup

Copy of files and applications made to avoid loss of data and facilitate recovery in the event of a system failure.

Biometrics

Methods for differentiating humans based upon one or more intrinsic physical or behavioral traits such as fingerprints or facial geometry.

Biometric authentication

Using biometrics to verify or authenticate the identity of a person.

Business continuity plan (BCP)

The documentation of a predetermined set of instructions or procedures that describe how an organization's critical business functions will be sustained during and after a significant disruption.

Category-1 data

Information whose confidentiality is protected by law or contract. For a full definition see the Data Classification Standard.

Category-1a data

Information whose confidentiality is protected by law or contract, but for which there are no specifically proscribed penalties. For a full definition see the Data Classification Standard.

Category-2 data

University information usually restricted to university employees, but which are releasable in accordance with the Texas Public Information Act. For a full definition see the Data Classification Standard.

Category-3 data

University information that is generally publicly available. For a full definition see the Data Classification Standard.

Centralized storage

Storage on a central server made available over a network to users.

Change

Any implementation of new functionality, interruption of service, repair of existing functionality, and/or removal of existing functionality to an information system.

Change management

The process of controlling modifications to hardware, software, firmware, and documentation to ensure that information systems are protected against improper modification before, during, and after system implementation.

Computer security incident

See Security incident.

Confidential

The classification of data of which unauthorized disclosure/use could cause serious damage to an organization or individual.

Confidentiality

Characteristic of information indicating it is intended to be known by a limited set of people.

Confidential information

Information maintained by the university that is exempt from disclosure under the provisions of the Public Records Act or other applicable state and federal laws. The controlling factor for confidential information is dissemination.

Control

Method used to reduce the probability of occurrence or the negative impact of the realization of a risk.

Custodian

Custodians ensure the effective and secure operation of the information owner's systems. See the OP 44 series of operating policies for more information.

Data

See the Data Classification Standard.

Data loss prevention

Prevention of unnecessary exposure of protected information.

Data owner

See information owner.

Digital certificate

An electronic document which uses a digital signature to bind specially derived numerical information with an identity - such as the name of a person or an organization. Most often encountered on web sites using encryption (SSL/https).

Digital signature

Method of adding specially derived numerical information to a file or message (most often used as part of a digital certificate infrastructure).

Digital data

The subset of Data (as defined above) that is transmitted by, maintained, or made available in electronic media.

Disclosure

The act, intentional or otherwise, of revealing information that is otherwise held as confidential or protected.

Disaster Recovery Plan (DRP)

A written plan for processing critical IT applications in the event of a major hardware or software failure or destruction of facilities. Such plans are

designed to restore operability of the target system, application, or computer facility.

DMZ

A DMZ, or demilitarized zone, is a physical or logical network that contains and exposes external-facing services to the Internet. Systems that need to be made available to the Internet, such as the ASU website, are located in a DMZ.

DNS

The Domain Name System (DNS) is a naming system for computers, services, or other resources connected to a network that associates a name with an IP address.

Electronic mail (email)

Any message, image, form, attachment, data, or other communication sent, received, or stored within an electronic mail system.

Electronic mail system

Any computer software application that allows electronic mail to be communicated from one computing system to another.

Electronic media

Any of the following: a) Electronic storage media including storage devices in computers (hard drives, memory) and any removable/transportable digital storage medium, such as magnetic tape or disk, optical disk, or digital memory card; or b) Transmission media used to exchange information already in electronic storage media. Transmission media include, for example, the internet (wide-open), extranet (using internet technology to link a business with information accessible only to collaborating parties), leased lines, dial-up lines, private networks, intranet, and the physical movement of removable/transportable electronic storage media.

Emergency change

When an unauthorized immediate response to imminent critical system failure is needed to prevent widespread service disruption.

Encrypted data

Data rendered unreadable to anyone without the appropriate cryptographic key and algorithm.

Encryption

Process of numerically changing data to enhance confidentiality. Data is obscured using a specific algorithm and key both of which are required to interpret the encrypted data.

End user

A person given authorization to access information on a system.

Escrow

Data decryption keys or passwords held in trust by a third party to be turned over to the user only upon fulfillment of specific authentication conditions.

Exposure

State during which a system’s controls do not adequately reduce risk that the information could be stolen or exploited by an unauthorized person.

Firewall

An access control mechanism that acts as a barrier between two or more segments of a computer network or overall client/server architecture, used to protect internal networks or network segments from unauthorized users or processes. Such devices include hardware that is placed in the network to create separate security zones, provide NAT, and create a point of access control.

Hardening

The process of making computer and network systems more resistant to tampering or malicious software.

Incident

Any set of circumstances in which the anticipated and configured delivery of a service is interrupted, delayed, or otherwise unavailable.

Incident management

Process of returning service as quickly and effectively as possible.

Information custodian or custodian

See Custodian.

Information owner

Responsible for specified information and establishing the controls for its collection, creation, processing, access, dissemination, and disposal. The owner of a collection of information is the person responsible for the business results of that system or the business use of the information. Where appropriate, ownership may be shared by managers of different departments. See OP 44.00 for a list of duties and responsibilities.

Information system

The equipment and software such as files, computers, tablets, servers, hard drives, removable thumb drives, cloud storage, etc. used to collect, record, process, display, and transmit information.

Information Resources Manager (IRM)

Authorized and accountable to the State of Texas for management of the university's information systems to implement security policies, procedures, and guidelines to protect the information systems of the university. The

Associate Vice President of Information Technology/CIO is designated as the university's IRM. The IRM will:

- Maintain information as a strategic asset of the university.
- Provide the resources to enable employees to carry out their responsibilities for securing information and information systems.
- Review and approve information owners and associated responsibilities.

Information Security Council

Body assembled by the CIO that contains at least the CIO and Information Security Officer. Provides direction and management of the information security program and information technology risk management program.

Information Security Officer (ISO)

Responsible for administering the information security functions within the university. The ISO is the university's internal and external point of contact and internal resource for all information security matters. The ISO will:

- Develop, coordinate and administer the ASU Information Security Program and periodically assess whether the program is implemented in accordance with ASU IT Security policies.
- Provide consultation on balancing effective IT security with business needs.
- Develop and maintain an information security awareness program.
- Provide solutions, guidance, and expertise in IT security.
- Maintain written IT security policies, standards and procedures as appropriate.
- Collecting data relative to the state of IT security at ASU and communicating as needed.
- Provide guidance on the information security requirements of federal, state and local privacy regulations.

Information security program

The elements, structure, objectives, and resources that establish an information system's security function within the university.

Integrity

The accuracy and completeness of information and assets and the authenticity of transactions.

Intellectual property

Ideas for which property rights are recognized under patent, trademark, or copyright law. Usually a work originating from thought or an idea that is distinct, separate, clearly definable, and novel.

Internet

A global system interconnecting computers and computer networks. The computers and networks are owned separately by a host of organizations, government agencies, companies, and colleges.

Intrusion detection system (IDS)

Hardware or a software application that can be installed on network devices or host operating systems to monitor network traffic and host log entries for signs of known and likely methods of intruder activity and attacks. Suspicious activities trigger administrator alarms and other configurable responses.

Lawful intercept

The interception of data on the university network by ISO and IT Networking and Telecommunications staff, in accordance with local law and after following due process and receiving proper authorization from the appropriate authorities.

Local account

Account that allows access only to a local system and uses that systems local authentication service.

Local area network (LAN)

A data communications network spanning a limited geographical area. It provides communication between computers and peripherals at relatively high data rates and relatively low error rates.

Local storage

Storage that is physically local to the workstation or server.

Malicious code

Software designed to infiltrate or damage a computer system without the owner's informed consent. The expression is a general term used by computer professionals to mean a variety of forms of hostile, intrusive, or annoying software or program code including spyware, Trojan horses, viruses, and worms.

Mission critical information system

Information system defined to be essential to the university's function and which, if made unavailable, will inflict substantial harm to the university and the university's ability to meet its instructional, research, patient care, or public service missions. Mission critical information systems include those systems containing sensitive information.

Network

All associated equipment and media creating electronic transmission between any information system(s), such as wired, optical, wireless, IP, synchronous serial, telephony, etc.

Offsite storage

Based on data criticality, offsite storage should be in a geographically different location from the campus that does not share the same disaster threat event. Based on an assessment of the data backed up, removing the backup media from the building and storing it in another secured location on the campus may be appropriate.

Owner

See information owner.

Password

A string of characters used to verify or “authenticate” a person’s identity.

Password complexity

The characteristic of a password typically described by the number of characters, the size of the character set used, and the randomness with which those characters were chosen.

Password strength

Description of a password’s ability to resist being guessed or otherwise mathematically or cryptographically discovered.

Patch

A fix or update for a software program usually related to a security issue.

Penetration Test

A controlled attempt to circumvent the security of a network or computer system to test its ability to resit hacking.

Perimeter security control

The first layer of defense against malicious traffic that filters information between university internal networks and the internet.

Personally identifiable information

Any information that alone or in conjunction with other information identifies an individual, including Social Security numbers, driver’s license numbers, military ID numbers, passport ID numbers, passwords/PINs, personal accounts, credit card numbers, protected health information, financial

information, criminal history records, unpublished home addresses or phone numbers, biometric data, and any other information that is deemed confidential by law or university policy.

Physical security

Area of knowledge concerned with creating and enhancing the safety and security of a physical space and the physical assets contained therein.

Physical security control

Devices and means to control physical access to sensitive information and to protect the availability of the information. Examples are physical access systems (fences, mantraps, guards); physical intrusion detection systems (motion detector, alarm system); and physical protection systems (sprinklers, backup generator).

PIN

Personal identification number - typically associated with systems using a physical security card (ATMs) together with a short number to authenticate an individual.

Plaintext data

Data in a form readable by anyone having access to the system on which it is stored or to the network over which it is transmitted.

Portable computing device

Any handheld portable device capable of performing basic computer tasks such as chat, email, web browsing, and storing information - smart phones, tablet computers (iPads), and PDAs all fall into this definition.

Production system

Any University system, software, or application that is used in the daily operations of the University.

Program

Set of instructions written in a computer programming language that performs a specific set of related functions (Microsoft Word, et.al.).

Protected information

Any information provided protection by law, regulation, or other legal means which mandates the methods, controls, processes, and/or procedures to afford such protection. This includes Personally Identifiable Information (PII).

Removable media

Any storage device built and intended to be easily connected to and removed from a computer system - examples include memory sticks, pen drives, external hard drives, and CD/DVDs.

Resolution

Returning service through the implementation of a permanent solution or a workaround.

Risk

Potential that a given set of circumstances and actions will lead to an undesirable outcome - in terms of information this means loss of one or more of (confidentiality, availability, and integrity).

Residual risk

Any risk remaining once controls have been applied. The amount of residual risk allowed will be determined by the organization's tolerance for risk.

Risk assessment

The process of identifying, evaluating, and documenting the level of impact that may result from the operation of an information system on an organization's mission, functions, image, reputation, assets, or individuals.

Risk assessment incorporates threat and vulnerability analyses and considers mitigations provided by planned or current security controls.

Risk management

Decisions to accept risk exposures or to reduce vulnerabilities and to align information system risk exposure with the organization's risk tolerance.

Root access

Most privileged access to a computer system allowing the use, change, and deletion of any and all configuration information, system software, and data.

Scheduled change

Formal notification received, reviewed, and approved by the review process in advance of the change being made.

Scheduled outage

Any previously agreed upon period in which a system is not available for normal use. This typically requires specific methods of discussion, approval and scheduling (Change Management).

Security administrator

The person charged with monitoring and implementing security controls and procedures for a system. Whereas each university will have one information security officer, technical management may designate a number of security administrators.

Security incident

Any incident in which the secure configuration of a system has been compromised.

Security incident management

Area of incident management focused on controlling and correcting vulnerabilities, exposures, and compromise of the secure configuration of any system.

Sensitive information

Information maintained by the university that requires special precautions to protect it from unauthorized modification or deletion. Sensitive information may be either public or confidential. It is information that requires a higher than normal assurance of accuracy and completeness. The controlling factor for sensitive information is that of integrity.

Server

Any computer providing a service over the network. Services include, but are not limited to: website publishing, SSH, chat, printing, wireless access, and file sharing.

Single sign-on

Ability for a user to sign in once and have that sign-in allow access to multiple information systems without the need for providing a username and password for each separate system.

Spyware

Software that is installed surreptitiously on a computer to intercept or take partial control over the user's interaction with the computer, without the user's informed consent. While the term suggests software that secretly monitors the user's behavior, the functions of spyware include collecting various types of personal information, interfere with control of the computer, changing computer settings, and redirecting web browser activity.

Strong password

A strong password is constructed so that it cannot be easily guessed by another user or a "hacker" program. It is typically a minimum number of

positions in length and contains a combination of alphabetic, numeric, or special characters and should not be linked to any personal information such as a birth date, Social Security number, and so on.

System administrator

Person responsible for the effective operation and maintenance of information systems, including implementation of standard procedures and controls, to enforce a university's security policy.

Synchronization

Process whereby information on two systems is shared so that each system's copy is identical to the other.

System

In the context of IT, any device capable of performing complex functions to provide services by use of hardware, firmware, software, or other programming. Systems may include workstations, desktops, laptops, servers, routers, and switches.

System hardening

Process of enhancing the configuration of a system so that there is greater assurance the system can be used only by authorized users for authorized purposes.

Test and development systems

Systems used exclusively for testing or development of software and not used to directly support university operations.

Trojan

Destructive programs-usually viruses or worms-that are hidden in an attractive or innocent-looking piece of software, such as a game or graphics program. Victims may receive a Trojan horse program by email or on portable media,

often from another unknowing victim, or may be urged to download a file from a website.

Unauthorized disclosure

The intentional or unintentional revealing of restricted information to people who do not have a legitimate need to access that information.

Unscheduled change

Failure to present notification to the formal process in advance of the change being made. Unscheduled changes will only be acceptable in the event of a system failure or the discovery of a security vulnerability.

Unscheduled outage

Any period in which a system is not available for normal use and that lack of availability was not previously discussed, approved, and scheduled.

UPS

An uninterruptible power supply. An electrical apparatus that provides emergency power to a load when the input power source (usually commercial power) fails.

Use/Using

An umbrella term that includes the terms store, process, change, delete, read, and access (and their progressive forms).

User

An individual that is authorized by the information owner to access the resource, in accordance with the information owner's procedures and rules. The user is any person who has been authorized by the information owner to read, enter, or update that information. The user is the single most effective control for providing adequate security. See OP 44.00 and OP 44.01 for a list of duties and responsibilities.

Username

A pseudonym used by a user to access a computer system - typically based on the user's legal name or some derivative thereof.

Virtual private network (VPN)

Encrypted connections over a larger network, typically over the Internet, which simulates the behavior of direct, local connections.

Virus

A computer virus refers to a program that enters your computer—often through email or Internet downloads—and makes copies of itself, spreading throughout your computer and files. There is a wide range of computer viruses out there. They can be anything from merely annoying to horribly damaging—deleting files or making your computer inoperable. Viruses attach themselves to an application on a computer and aren't actually executed until that application is accessed or run.

Vulnerability

Any exploitable aspect of a system or process.

ºWeb page

A document on the World Wide Web. Every web page is identified by a unique URL (uniform resource locator).

Web server

A computer that delivers (serves up) web pages.

Website

A location on the World Wide Web, accessed by typing its address (URL) into a web browser. A website always includes a home page and may contain additional documents or pages.

Wireless networking

Transmission of computer-based information over short to medium distances using radio frequencies.

Wireless adhoc networking

Wireless networking in which centralized authorization and infrastructure are not used - this is an unauthorized method of connecting systems to the university network.

World Wide Web

Also referred to as “the Web.” A system of Internet hosts that supports documents formatted in HTML (hypertext markup language), which contain links to other documents (hyperlinks) and to audio, video, and graphic images. Users can access the Web with special applications called browsers, such as Firefox and Microsoft Internet Explorer.

Worm

A program that makes copies of itself elsewhere in a computing system. These copies may be created on the same computer or may be sent over networks to other computers. The first use of the term described a program that copied itself benignly around a network, using otherwise-unused resources on networked machines to perform distributed computation. Some worms are security threats, using networks to spread themselves against the wishes of the system owners and disrupting networks by overloading them. A worm is similar to a virus in that it makes copies of itself, but different in that it need not attach to particular files or sectors at all.

The A-Z of Technology Terms

Don't know your bandwidth from your back-end? Here's a handy glossary of tech terms.

Adobe Acrobat Reader
Acrobat Reader is software that allows you to view a PDF document (a document that can be seen but not changed). It can be downloaded free of charge from Adobe.

ADSL
Asymmetric digital subscriber line (ADSL) is a type of digital subscriber line (DSL) broadband technology that is used to connect to the Internet. It uses standard telephone lines to deliver high-speed data communications (up to 24 megabytes per second).

Analogue
Analogue is a conventional method of transmitting data. Standard landline telephones use analogue technology. It is distinct from digital technology, which provides for greater quality and speed of data transmission.

Assistive technology
Assistive technology refers to any software or hardware that acts to assist and improve the functional capabilities of people with disabilities. Examples include wheelchairs, prosthetics, voice-to-text technology and text-to-speech technology.

Attachment
An attachment is a document sent with an email message. Many types of files can be sent this way (e.g. Word documents, PDFs, Excel files, JPEGs). Be wary of attaching large files because these can take a lot of time for the recipient to download. If you have a large file, it is considered good practice to compress the file using software such as Winzip before attaching it.

Back-end
Back-end refers to the part of an application that performs an essential task not apparent to the user.

Backward compatible
If software is backward compatible, it is compatible with earlier (superseded) versions of the same software. For example, the Microsoft word-processing program Word 2010 can read files created in the 2003 version of the same program, so it is backward compatible.

Bandwidth

Bandwidth refers to the maximum amount of data that can travel a communications path in a given time, usually measured in seconds.

Bit
A bit (short for binary digit) is the smallest unit of measurement in computing. 8 bits make up 1 byte.

Bluetooth
Bluetooth is a wireless communications technology intended to replace cables. It allows short-range connections between two or more Bluetooth-compatible devices such as mobile phones, tablets, headsets or medical equipment.

Bookmark
A bookmark is a saved link to a particular Web page. Microsoft Internet Explorer denotes bookmarks as "favourites."

Boolean operators
Most search engines (e.g. Google) allow you to limit your search or make it more specific by using words such as "and", "or" and "not". These words are known as boolean operators because of their origin as terms in logic.

Boot (re-boot)
To boot (or re-boot) is to load and initialise the operating system on a computer. Think of it as starting up your computer. In Windows you can use the key combination CTRL and ALT and DEL as a "soft" boot. This means restarting the computer rather than turning it completely off and on again, which could cause damage to your computer's hard disk under some circumstances.

Bounce back
An email message that cannot be delivered and returns an error notification to the sender is said to "bounce back". If you receive such an error notification, check that you have typed the address correctly.

Broadband
Broadband is a type of communications technology whereby a single wire can carry more than one type of signal at once; for example, audio and video. Cable TV is one technology that uses broadband data transmission.

Browser
A software program that allows you to surf the web. Popular web browsers include Google Chrome, Mozilla Firefox, Microsoft Edge and Internet Explorer.

Cache
When you download (read) a web page, the data is "cached," meaning it is temporarily stored on your computer. The next time you want that page, instead of

requesting the file from the web server, your web browser just accesses it from the cache, so the page loads quickly. The downside to this is that if the cached web page is often updated, you may miss the latest version. If you suspect that the web page you're seeing is not the latest version, use the "refresh" button on your browser.

CAD
Computer-aided design (CAD) is a type of software that allows users to create 2D and 3D design and modelling. CAD is used by architects, engineers, artists and other professionals to create precise technical drawings.

Chip
A chip is a microprocessor that performs many functions and calculations that make your computer run. Your computer's chip is also referred to as the CPU (Central Processing Unit) or the processor.

Cloud computing
Cloud computing refers to the storing and accessing of data and programs over the Internet instead of on another type of hard drive. Examples of Cloud services include iCloud, Google Cloud and Dropbox.

Compression
Compression is the reduction of the size of a file. Compressed files take up less memory and can be downloaded or sent over the Internet more quickly.

Content
Content refers to a website's text and information, as opposed to its design and structure.

Cookie
A piece of code or data created by a web server and stored on a user's computer. It is used to keep track of the user's usage patterns and preferences.

CPU
The central processing unit (CPU) is the brains behind your computer. The CPU is

responsible for performing calculations and tasks that make programs work. The higher the speed of a CPU, the faster the CPU undertakes the calculations and tasks.

Cybercrime
Cybercrime is any type of illegal activity that is undertaken (or relies heavily) on a computer. There are thousands of types of cybercrime, including network intrusions, identity theft and the spreading of computer viruses.

Cybersecurity
Cybersecurity refers to measures designed to protect your computer, device or network from cybercrime. This involves preventing unintended and unauthorised access, change and damage.

Device driver
A device driver is a small program that allows a peripheral device such as a printer or scanner to connect to your PC.

Domain
A domain is a set of computers on a network that are managed as a unit.

Download
Downloading is the method by which users access and save or "pull down" software or other files to their own computers from a remote computer via the Internet.

DV
DV stands for digital video.

Email
Email or electronic mail is a way of sending messages over the internet. Popular email programs include Outlook, Mozilla Thunderbird, Gmail and Yahoo Mail.

Encryption
Encryption is the process of converting electronic data to an unrecognisable or encrypted form, one that cannot be easily understood by unauthorised parties.

Ethernet
Ethernet is the most common way of connecting computers on a network with a wired connection. It is a type of local area network (LAN) technology, providing a simple interface for connecting multiple devices.

Firewall
A firewall is a barrier that acts as a security system to protect trusted computer systems and networks from outside connections and untrusted networks, such as the Internet.

FTP
File transfer protocol (FTP) is a common method of transferring files via the internet from one host to another host.

Gateway
A point within a network that interconnects with other networks.

GIF
Graphics interchange format (GIF) is a graphics file format. Because GIF files are compressed, they can be quickly and easily transmitted over a network. GIF is one of the main graphics formats on the Internet.

Hard disk
The physical place where a computer stores information - applications and files - is known as its hard disk drive (HDD). The bigger the HDD, the more data it can store.

Home page
The page that an Internet browser first opens up to. It is usually the starting point of an organisation's or individual's website.

HTML
Hyper-text markup language (HTML) is a set of symbols inserted into files intended for display on the world wide web. The symbols tell web browsers how to display words and images - e.g. which colour, font and type size to use - and they direct it to link to other pages on the world wide web via hyperlinks.

Internet
A set of interconnected networks that allow computers in different locations to exchange information. The Internet includes services such as the world wide web, electronic mail, file transfer protocol (FTP), chat and remote access to networks and computers.

ISP
An internet service provider (ISP) is a company that provides access to the Internet. In Australia, widely used ISPs include Bigpond, iinet and Dodo.

Intranet
An intranet is basically a private, internal internet specific to an organisation or group.

Java
Java is a programming language that is commonly used in the development of client-server web applications.

JPEG
JPEG stands for Joint Photographic Experts Group, which was the committee that

created the file format known as JPEG. The format is commonlyl used for photos displayed on the world wide web.

LAN
A local area network (LAN) is a system that connects computers and other devices that share a common communications line and wireless link, generally within a limited geographical area such as a home or office building.

Malware
"Malware" is short for malicious software. It refers to a software program that has been developed to do harm to other computers. Types of malware include viruses, worms and spyware.

Megabyte
A measure of computer processor storage and real and virtual memory. A megabyte (Mb) is 2 to the 20th power bytes, or 1,048,576 bytes in decimal notation.

Megahertz
Megahertz is the unit used to measure the speed of a computer's processor (e.g. 2.8Ghz)

Modem
A modem is a device that allows computers to transmit information to each other via ordinary telephone lines.

Online
If a computer (or computer user) is online, it is currently connected to a network or to the Internet. Online also refers to resources and services available on the Internet - e.g. online banking, online dictionary.

Operating system
An operating system (OS) is the software that manages all of a computer's processes and allows programs and applications to run. The most prominent operating system is Microsoft Windows. Others include Mac OS X and Linux.

PDF
Portable document format (PDF) is a file type created by Adobe Systems Inc. PDFs can be read using free software called Adobe Acrobat Reader or another PDF reader.

Phishing
Phishing is a type of email fraud in which the perpetrator sends out emails that appear to come from a legitimate service or reputable company, such as a bank or an email service provider. These emails aim to lure recipients to reveal confidential information that the perpetrator can use for their financial advantage - for example, online banking

log-in details and passwords.

Plug-in
A software plug-in is a component that adds to a software program's functionality.

POP
A Post office protocol (POP) is an Internet protocol used by your Internet service provider (ISP) to handle email. A POP account is an email account.

PPM
Pages per minute (PPM) generally refers to the speed of a printer.

Processor
The processor is the brains of your computer. It is responsible for performing calculations and tasks that make programs work. The faster the processor, the faster the computer works.

Protocol
A protocol is a standard or set of rules that computers and other devices use when communicating with one another.

RAM
Random access memory (RAM) is usually referred to as a computer's "memory" - it stores information used by programs. Generally, the larger your computer's RAM, the more programs it can run at once without slowing down.

Read-only
A read-only file cannot be edited, modified or deleted.

Resolution
Resolution refers to the number of distinct pixels that make up the display on a computer monitor. It is denoted in DPI (dots per inch). The higher the resolution, the finer and smoother the images appear when displayed at a given size.

ROM
ROM stands for read-only memory. It is the part of a computer's memory that cannot be changed by a user. The contents of ROM remain even when the computer is turned off.

SAAS
SAAS stands for software as a service. It is a software distribution model whereby software applications are centrally hosted and licensed on a subscription basis.

Search engine
A search engine enables a computer user to search information on the Internet. It is a

type of software that creates indexes of databases or Internet sites based on the titles of files, keywords, or the full text of files. The most popular search engines are Google.com.au, Yahoo.com.au and Bing.com.au.

SSL
SSL, or secure sockets layer, is a protocol that allows Internet users to send encrypted messages across the Internet. It is generally used when transmitting confidential information (e.g. personal data or credit card details). A web address that begins with "https" indicates that an SSL connection is in use.

SEO
SEO, or search engine optimisation, is the practice of making adjustments to certain aspects of a website in an effort to improve its ranking on search engines.

Server
A server is a computer that handles requests for data, email, file transfers, and other network services from other computers.

Spam
Spam refers to unsolicited email messages sent for marketing purposes.

Unzip
To unzip a zip file is to extract and decompress compressed files from it. If you are sent a zip file via email, you will need to unzip it before you can access the files inside it.

URL
A URL (unique resource locator) or web address is the string of characters you type into a browser to access a particular website or other resource on the Internet. (eg. http://www.ourcommunity.com.au)

Viral
If an online video, photo or article "goes viral", it experiences a sudden spike in popularity in a short period of time.

Virus
A virus is a piece of programming code inserted into other programming to cause damage. Viruses can be sent in many forms but are often transmitted via email messages that, when opened, may erase data or cause damage to your hard disk. Some viruses are able to enter your email system and send themselves to other people in your list of contacts.

WEP
Wired equivalent privacy (WEP) is a security protocol used in wi-fi networks. It is designed to provide a wireless local area network (LAN) with a level of security

similar to that of a regular wired LAN. WEP-secured networks are usually protected by passwords. (See also WAP.)

Wi-Fi
Wi-Fi is a technology that allows computers and other devices to communicate via a wireless signal. Essentially, it means you can browse the internet without tripping over phone cords.

WPA
Wi-Fi protected access (WPA) is a security protocol used in wi-fi networks. It is an improvement on WEP because it offers greater protection through more sophisticated data encryption.

Zip
To zip files is to archive and compress them into one file of smaller size using a program such as WinZip. It's a handy way to make files smaller before sending them via email.

TECHNOLOGY TERMS

The IT industry is chock full of its own lingo. Below are definitions for some of the most of the common words and terms in today's technology vocabulary.

A

Access point: A device that allows wireless-equipped computers and other devices to communicate with a wired network.

Alias: A short, easy to remember name created for use in place of a longer, more complicated name; commonly used in e-mail applications. Also referred to as a “nickname”.

Authentication: The process of identifying yourself and the verification that you’re who you say you are. Computers where restricted information is stored may require you to enter your username and password to gain access.

B

Bandwidth: A measurement of the amount of data that can be transmitted over a network at any given time. The higher the network’s bandwidth, the greater the volume of data that can be transmitted. Network bandwidth is usually expressed in bits per second (bps); modern networks typically have speeds measured in the millions of bits

per second (megabits per second, or Mbps) or billions of bits per second (gigabits per second, or Gbps).

Business Continuity: Activity performed by an organization to ensure that critical business functions will be available to customers, suppliers, regulators, and other entities that must have access to those functions. These activities include many daily chores such as project management, system backups, change control, and help desk. Business Continuity is not something implemented at the time of a disaster; Business Continuity refers to those activities performed daily to maintain service, consistency, and recoverability.

BYOD: Bring Your Own Device or "BYOD" is a business and technology policy that allows employees to bring in personal mobile devices and use these devices to access company data, email, etc.

C

Cache: A set of files saved on your hard disk that help your browser display pages you have already visited more quickly. It displays the files from your hard disk instead of the web.

Cloud: A common shorthand for a provided cloud computing services (or even an aggregation of all existing cloud services) is "The Cloud". The cloud, simply, refers to software and services that run on the Internet instead of your computer. Apple iCloud, Dropbox, Netflix, Amazon Cloud Drive, Flickr, Google Drive, Microsoft Office 365, Yahoo Mail — those are all cloud services.

Content Management: Content management is a collection of tools, software and processes that allow you to collect, manage and publish information on any medium. Content management also ensures that whatever content you have is indexable or searchable, making it easy for users to find that information. It also encompasses deletion of content.

Cookie: A small piece of information you may be asked to accept when connecting to certain servers via a web browser. It is used throughout your session as a means of identifying you. A cookie is specific to, and sent only to the server that generated it.

CPU: Central processing unit; the part of a computer that oversees all operations and calculations.

Cross-Platform: Cross-platform (a.k.a. agnostic, agnostic application and agnostic platform) refers to something that is interoperable among different operating systems, systems and platforms. Cross-platform, however, is more focused on software that can run on any operating system and on any processor architecture.

Crowdsourcing: Crowdsourcing means the outsourcing of your tasks to, well, a crowd of people. That is, you will have a loosely defined group, which will do the work for you. Crowdsourcing takes away the task that you have to do in-house or by using an employee and letting a bigger group of people do it. In IT, crowdsourcing usually involves the general public.

CSP: Cloud Service Provider; a business model for providing cloud services.

CSS: Cascading Style Sheet; A set of rules that define how web pages are displayed using CSS, designers can create rules that define how page.

Customer Relationship Management: Customer relationship management refers to a set of processes, tools, technologies and software that help businesses manage their relationships with their customers. For instance, a customer relationship management software would be able to give businesses enough details about their customers from the level of communication that has been done by each customer and when. This allows businesses to personalize and customize their contacts with the customer to gain more sales.

D

Database: A collection of information organized so that a computer application can quickly access selected information; it can be thought of as an electronic filing system. Traditional databases are organized by fields, records (a complete set of fields), and files (a collection of records). Alternatively, in a Hypertext database, any object (e.g., text, a picture, or a film) can be linked to any other object.

Data center: Facility used to house computer systems and associated components, such as telecommunications and storage systems. Generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and security devices.

Desktop: On computers like IBM PC or compatibles and Macintoshes, the backdrop where windows and icons for disks and applications reside.

Devices: Input devices include things like microphones, keyboards, mouse, touchpads, wheels, joysticks, etc. Output devices include printers, monitors, projectors and speakers.

DHCP: Dynamic Host Configuration Protocol; a protocol that lets a server on a local network assign temporary IP addresses to a computer or other network devices.

Dialog box: Sometimes referred to as a window; on a graphical user interface system, an enclosed area displayed by a program or process to prompt a user for entry of information in one or more boxes (fields).

Disaster recovery: Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of technology infrastructure critical to an organization after a natural or human-induced disaster. Disaster recovery is a subset of business continuity. While business continuity involves planning for keeping all

aspects of a business functioning in the midst of disruptive events, disaster recovery focuses on the IT or technology systems that support business functions.

Disruptive Technology: New technology that surprisingly displaces an already established one.

DNS: The domain name system is how computers convert human-readable domain names and hostnames to numerical IP addresses. When you type howtogeek.com into your web browser's address bar, your computer contacts its DNS server and the DNS server replies with the numerical IP address of How-To Geek's server, which is what your computer connects to.

DHCP: The dynamic host configuration protocol allows computers to automatically request and be assigned IP addresses and other network settings. For example, when you connect your laptop or smartphone to your Wi-Fi network, your device asks the router for an IP address using DHCP and the router assigns an IP address. This simplifies things — you don't have to set up static IP addresses manually.

Domain Name: Domain names are the base part of website names like howtogeek.com or google.com. Note that domain names are just another type of hostname.

Domain: Part of an Internet address. The network hierarchy consists of domains and subdomains. At the top are a number of major categories (e.g., com, edu, gov); next are domains within these categories (e.g., ohio-state); and then there are subdomains. The computer name is at the lowest level of the hierarchy.

Download: The process of transferring one or more files from a remote computer to your local computer. The opposite action is upload.

E

EHR/EMR/PHR (Electronic Health Record/Electronic Health Record/Personal Health Record) these three are together because people often use them interchangeably. EMR stands for Electronic Medical Record, EHR stands for Electronic Health Record and PHR stands for Personal Health Record.

Encryption: The manipulation of data to prevent accurate interpretation by all but those for whom the data is intended.

Ethernet: Ethernet is the standard wired network technology in use almost everywhere today. If your computer is connected to a network via a cable, it's likely using an Ethernet cable. That cable plugs into an Ethernet port on your computer.

Ethernet card: An adapter card that fits into a computer and connects to Ethernet cabling; different types of adaptor cards fit specific computers.

F

Female connector: A cable connector that has holes and plugs into a port or interface to connect one device to another.

Filter: Refers to: 1) a program that has the function of translating data into a different format (e.g., a program used to import or export data or a particular file); 2) a pattern that prevents non-matching data from passing through (e.g., email filters); and 3) in paint programs and image editors, a special effect that can be applied to a bit map.

Firewall: A firewall is a piece of software or hardware that blocks certain types of traffic. For example, a firewall could block incoming traffic on a certain port or block all incoming traffic except traffic coming from a specific IP address.

Flash drive: A small device that plugs into computer's USB port and functions as a portable hard drive.

FTP: File Transfer Protocol; a method of exchanging files between computers via the Internet. A program like WS FTP for IBM PC or compatibles or Fetch for Macintosh is required. Files can contain documents or programs and can be ASCII text or binary data.

G

Gateway: A gateway is a device that routes traffic between networks. For example, at home, your router is your gateway. It provides a "gateway" between your LAN and WAN.

Gigabyte (Gig or GB): 1024 x 1024 x 1024 (2 to the 30th power) bytes; it's usually sufficient to think of a gigabyte as approximately one billion bytes or 1000 megabytes.

H

Help desk: A help desk is an information and assistance resource that troubleshoots problems with computers or similar products.

Hostnames: A hostname is a human-readable label that points to a device connected to a network. For example, on your home network, your Windows computer's hostname may be WINDOWSPC. Your other devices can connect to WINDOWSPC and will be pointed at that computer's local IP address.

HTML: HyperText Markup Language; a language used for creating web pages. Various instructions and sets of tags are used to define how the document will look.

HTTP: The hypertext transfer protocol is the standard protocol modern web browsers and the web itself uses. FTP and BitTorrent are examples of alternative protocols.

I

IaaS: Infrastructure as a Service; In the most basic cloud-service model, providers of IaaS offer computers – physical or (more often) virtual machines – and other resources.

IP Address: An Internet Protocol address, or IP address, is a numerical address that corresponds to your computer on a network. When a computer wants to connect to another computer, it connects to that computer's IP address.

ISP: Your Internet service provider is the company that provides you with your Internet connection. For example, your ISP may be Comcast, Time Warner, or whatever other company you're paying each month.

ITIL®: The Information Technology Infrastructure Library (ITIL®) sets forth the best practices and standards for providing IT services.

L

LAN: A local area network is a small network that's confined to a local area. For example, your home network or an office network is a LAN.

Connects a group of computers for the purpose of sharing resources such as programs, documents, or printers. Shared files often are stored on a central file server.

Localhost: The hostname “localhost” always corresponds to the device you’re using. This uses the loopback network interface — a network interface implemented in software — to connect directly to your own PC.

M

MAC: Media Access Control; The hardware address of a device connected to a shared network.

MAC Address: Each network interface has a media access control address, or MAC address — also known as a physical address. This is a unique identifier designed to identify different computers on a network. MAC addresses are usually assigned when a manufacturer creates a network device.

Mail server: A networked computer dedicated to supporting electronic mail. You use a client program like Microsoft Outlook for retrieving new mail from the server and for composing and sending messages.

Male connector: A cable connector that has pins and plugs into a port or interface to connect one device to another.

MDM: Mobile Device Management; Any routine or tool intended to distribute applications, data, and configuration settings to mobile communications devices. The intent of MDM is to optimize the functionality and security of a mobile communications network. MDM must be part of a coherent BYOD strategy.

MDR: Managed detection and response (MDR) is an outsourced service that provides organizations with threat hunting services and responds to threats once they are discovered. It also involves a human element:

Security providers provide their MDR customers access to their pool of security researchers and engineers, who are responsible for monitoring networks, analyzing incidents, and responding to security cases.

Megabyte (Meg or MB): 1,024 x 1,024 (2 to the 20th power) bytes; it's usually sufficient to think of a megabytes as one million bytes.

Microsoft Exchange: Microsoft Exchange Server is the server side of a client–server, collaborative application product developed by Microsoft. It is part of the Microsoft Servers line of server products and is used by enterprises using Microsoft infrastructure products. Exchange's major features consist of electronic mail, calendaring, contacts and tasks; support for mobile and web-based access to information; and support for data storage

Microsoft Windows: A group of operating systems for PC or compatible computers; Windows provides a graphical user interface so you can point and click to indicate what you want to do.

MIME: Multipurpose Internet Mail Extensions; a protocol that enables you to include various types of files (text, audio, video, images, etc.) as an attachment to an e-mail message.

Monitor: The part of a computer that contains the screen where messages to and from the central processing unit (CPU) are displayed. Monitors come in a variety of sizes and resolutions. The higher the number of pixels a screen is capable of displaying, the better the resolution. Sometimes may be referred to as a CRT.

MSP: Managed Service Provider; A business model for providing information-technology services.

Multi-Factor Authentication: An authentication method in which a computer user is granted access only after successfully presenting two or more pieces of evidence to an authentication mechanism: knowledge, possession, and inherence. Two-factor authentication is a type, or subset, of multi-factor authentication.

N

Nameserver: A computer that runs a program for converting Internet domain names into the corresponding IP addresses and vice versa.

NAT: Network Address Translation; a standard that enables a LAN to use a set of IP addresses for internal traffic and a single IP address for communications with the Internet.

Network: A group of interconnected computers capable of exchanging information. A network can be as few as several personal computers on a LAN or as large as the Internet, a worldwide network of computers.

Network adapter: A device that connects your computer to a network; also called an adapter card or network interface card.

Network hub: A common connection point for devices on a network.

Network Interface / Network Adapter: Your computer's wired Ethernet connection and Wi-Fi connection are basically both network interfaces. If your laptop was connected to both a wired connection and a Wi-Fi network, each network interface would have its own IP address. Each is a different connection.

NNTP: Network News Transport Protocol; the protocol used for posting, distributing, and retrieving network news messages.

O

On-site: At-place-of-work-or-business support, typically provided by a technically qualified individual.

P

Patch: Piece of software designed to update a computer program or its supporting data, to fix or improve it. This includes fixing security vulnerabilities and other bugs, and improving the usability or performance.

PC: Usually refers to an IBM PC or compatible, or when used generically, to a “personal computer”. In a different context, PC also is an abbreviation for “politically correct.”

Port: When an application wants to send or receive traffic, it has to use a numbered port between 1 to 65535. This is how you can have multiple applications on a computer using the network and each application knows which traffic is for it.

Protocol – TCP, UDP, ICMP, etc.: Protocols are different ways of communicating over the Internet. TCP and UDP are the most common protocols. The ICMP protocol is also used, but primarily so network devices can check each other’s status. Different protocols are ideal for different types of communication.

R

RAM: Random Access Memory provides space for your computer to read and write data so that the CPU can find it quickly and easily. When people refer to memory upgrades they are usually talking about RAM.

Remote desktop: A Windows feature that allows you to have access to a Windows session from another computer in a different location

Remote login: An interactive connection from your desktop computer over a network or telephone lines to a computer in another location (remote site).

Router: A device used for connecting two Local Area Networks (LANs); a device that passes traffic back and forth. You likely have a home router. It's that router's job to pass outgoing traffic from your local devices to the Internet, and to pass incoming traffic from the Internet to your devices.

S

SAN: A storage area network (SAN) is a dedicated storage network that provides access to consolidated, block level storage. SANs primarily are used to make storage devices (such as disk arrays, tape libraries, and optical jukeboxes) accessible to servers so that the devices appear as locally attached to the operating system. A SAN typically has its own network of storage devices that are generally not accessible through the regular network by regular devices.

SD-WAN: SD-WAN is an acronym for software-defined networking in a wide area network. SD-WAN simplifies the management and operation of a WAN by decoupling the networking hardware from its control mechanism.

Security Operations Center (SOC): A centralized unit that deals with security issues on an organizational and technical level. A SOC within a building or facility is a central location from where staff supervises the site, using data processing technology.Typically, a SOC is equipped for access monitoring, and controlling of lighting, alarms, and vehicle barriers.

SOC-as-a-service: A subscription- or software-based service that manages and monitors your logs, devices, clouds, network and assets

for internal IT teams. The service provides companies with the knowledge and skills necessary to combat cyber-security threats.

Serial port: An interface on a computer that supports transmission of a single bit at a time; can be used for connecting almost any type of external device including a mouse, a modem, or a printer.

Server: A computer that is responsible for responding to requests made by a client program (e.g., a web browser or an e-mail program) or computer. Also referred to as a “file server”.

SID: Service Set Identifier; a name that identifies a wireless network.

SIEM: Security information and event management (SIEM), software products and services combine security information management (SIM) and security event management (SEM). They provide real-time analysis of security alerts generated by applications and network hardware.

SMPT: Simple Mail Transfer Protocol is an Internet standard for e-mail transmission.

SSL: Small data files that digitally bind a cryptographic key to an organization’s details. When installed on a web server, it activates the padlock and the https protocol (over port 443) and allows secure connections from a web server to a browser. Typically, SSL is used to secure credit card transactions, data transfer and logins, and more recently is becoming the norm when securing browsing of social media sites. SSL Certificates bind together:

A domain name, server name or hostname.

An organizational identity (i.e. company name) and location.

An organization needs to install the SSL Certificate onto its web server to initiate secure sessions with browsers. Depending on the type of SSL Certificate applied for, the organization will need to go through differing levels of vetting. Once installed, it is possible to connect to the website over https://www.domain.com, as this tells the server to establish a secure connection with the browser. Once a secure connection is

established, all web traffic between the web server and the web browser will be secure. To view the details of an SSL Certificate, go to a secure site, click on the padlock and select “View Certificate”. All browsers are slightly different, but the Certificate always contains the same information.

SMB: The Server Message Block Protocol (SMB protocol) provides a method for client applications in a computer to read and write to files on and to request services from server programs in a computer network.

SMTP: Simple Mail Transfer Protocol; a method of handling outgoing electronic mail.

Software: Any program that performs a specific function. Examples: word processing, spreadsheet calculations, or electronic mail.

Spam: Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of spam that involves nearly identical messages sent to numerous recipients by email. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk. Spammers collect email addresses from chatrooms, websites, customer lists, newsgroups, and viruses which harvest users’ address books, and are sold to other spammers.

Switch: A switch serves as a controller, enabling networked devices to talk to each other efficiently. Through information sharing and resource allocation, switches save businesses money and increase employee productivity.

What is a Network Switch: Unmanaged Switches: An unmanaged switch works right out of the box. It’s not designed to be configured, so you don’t have to worry about installing or setting it up correctly. Unmanaged switches have less network capacity than managed switches. You’ll usually find unmanaged switches in home networking equipment.

What is a Network Switch: Managed Switches: A managed network switch is configurable, offering greater flexibility and capacity than an unmanaged switch. You can monitor and adjust a managed switch locally or remotely, to give you greater network control.

T

TCP/IP: Transmission Control Protocol/Internet Protocol; an agreed upon set of rules that tells computers how to exchange information over the Internet. Other Internet protocols like FTP, Gopher, and HTTP sit on top of TCP/IP.

Two-factor authentication: An extra level of security achieved using a security token device; users have a personal identification number (PIN) that identifies them as the owner of a particular token. The token displays a number which is entered following the PIN number to uniquely identify the owner to a particular network service. The identification number for each user is changed frequently, usually every few minutes.

U

USB: Universal Serial Bus; a connector on the back of almost any new computer that allows you to quickly and easily attach external devices such as mice, joysticks or flight yokes, printers, scanners, modems, speakers, digital cameras or webcams, or external storage devices.

Username: A name used in conjunction with a password to gain access to a computer system or a network service.

URL: A uniform resource locator, or URL, is also known as a web address. The current URL is displayed in your web browser’s address bar.

USB port: An interface used for connecting a Universal Serial Bus (USB) device to computer; these ports support plug and play.

V

Virtualization: Virtualization is the creation of a virtual (rather than actual) version of something, such as a hardware platform, operating system, a storage device or network resources. In hardware virtualization, the term host machine refers to the actual machine on which the virtualization takes place; the term guest machine, however, refers to the virtual machine.

Virus: A program intended to alter data on a computer in an invisible fashion, usually for mischievous or destructive purposes. Viruses are often transferred across the Internet as well as by infected diskettes and can affect almost every type of computer. Special antivirus programs are used to detect and eliminate them.

VoIP: Voice over Internet Protocol; a means of using the Internet as the transmission medium for phone calls. An advantage is you do not incur any additional surcharges beyond the cost of your Internet access.

VPN: Virtual Private Networking; a means of securely accessing resources on a network by connecting to a remote access server through the Internet or other network.

W

WAN: A wide area network is a larger network that covers a wider area. Your ISP provides you with a connection to their own WAN, which connects to the Internet.

WAP: Wireless Application Protocol; a set of communication protocols for enabling wireless access to the Internet.

WEP: Wired Equivalent Privacy; a security protocol for wireless local area networks defined in the 802.11b standard. WEP provides the same level of security as that of a wired LAN.

Wi-Fi: Wireless Fidelity; A generic term from the Wi-Fi Alliance that refers to of any type of 802.11 network (e.g., 802.11b, 802.11a, dual-band, etc.). Products approved as “Wi-Fi Certified” (a registered trademark) are certified as inter-operable with each other for wireless communications.

Window: On a graphical user interface system, a rectangular area on a display screen. Windows are particularly useful on multitasking systems which allow you to perform a number of different tasks simultaneously. Each task has its own window which you can click on to make it the current process. Contrast to a “dialog box” which is used to respond to prompts for input from an application.

Windows: A casual way of referring to the Microsoft Windows operating systems.

Wireless (networking): The ability to access the Internet without a physical network connection. Devices such as cell phones and PDAs that allow you to send and receive email use a wireless Internet connection based on a protocol called WAP (Wireless Application Protocol). At this point, web sites that contain wireless Internet content are limited, but will multiply as the use of devices relying on WAP increases.

WLAN: Wireless Local Area Network; the computers and devices that make up a wireless network.

WPA: Wi-Fi Protected Access; a standard designed to improve on the security features of WEP.

Confused About All These Tech

Terms: Information Technology Glossary

Glossary

Artificial Intelligence-The development of computer systems that can perform tasks that would normally require a human. They include speech recognition, language translation, visual perception, and decision making skills.

BDR – This abbreviation stands for "backup and disaster recover." This is a plan where all hardware and software is regularly saved in both onsite and offsite locations. This can prevent data from actually being lost. If a data breach occurs, an IT specialist can restore the whole system from backed-up copies. This helps a business to get back up and running very quickly even after hurricanes or other disasters.

Big data-This term refers to extremely large data sets used for the purposes of analysis. This complex data processing is often completed by larger companies in order to reveal trends and patterns that can help them with upcoming marketing campaigns.

Breach (cyber) defense-Any method of protecting your networking and computing resources from intruders and hackers. This can include protection for your IT infrastructure along with antivirus software and firewalls. All known threats are analyzed, then strategies are formulated to stop these attacks before they occur. Most breach defense mechanisms include a planned response to any type of cyber-attack. A good breach defense will include strong measures to safeguard company assets.

Business Continuity– The ability for a business to continue its operations even when there is a cyber breach or other disaster that would normally cripple a company.

Cyber or security breach– An internet security breach where cyber thieves hack into your computer systems and steal data or plant malware. These breaches can cripple your organization and damage your data including customer records. Cyber breaches usually cost a business over one million dollars for each event and they damage the company's reputation, harming the brand.

Dark Web– This is the Back Alley of the Internet. It is where illegal transactions often take place. Drugs and other contraband

are sold on the Dark Web. Entry into this area requires a special browser known as Tor (The Onion) and special software to protect the user from the many dangerous threats that loom on the Dark Web.

Deep Web– Any content that is not indexed by Google bots including educational sites, email programs, company intranet, etc.

Hybrid Cloud-A cloud computing environment where a mixture of public and private cloud services are created to lower operating costs and gain access to a wider range of computing resources.

IT Infrastructure-This term includes all networking, servers, computers, software, hardware, and other technology used to manage and support all information technology resources.

Machine learning-This is a subset of artificial intelligence (AI) where a computer system automatically learns as it processes data. The computer doesn't need to be explicitly programmed to achieve desired results. This science develops computer programs that can access data and learn while they're working so that they evolve into smarter machines over time.

Malware– A combination of the words "malicious" and "software". This term has come to refer to any type of software that was built for the specific intent of disrupting a company's computer network and damaging computer equipment. It can include spyware, viruses, Trojan horses, worms, adware, and many other types of malicious software.

MSP– An IT Managed Service Provider (MSP) is a company that provides a full range of services to help businesses manage all their computer and network resources.

Outsourcing-The process of hiring an outside company to manage all aspects of networking, IT, communications, computers, servers, and other information technology.

Private Cloud-A private cloud refers to a cloud computing environment where a private IT infrastructure is created for one specific organization. It is usually managed with internal resources.

RPO– Recovery Point Objective is the maximum amount of time that data should be lost due to a natural or manmade disaster.

RTO– Recovery Time Objective is the maximum amount of time a business can be without its data (due to a natural or manmade disaster) before the business is at risk.

Ransomware-A type of malicious software where a cyber thief blocks the user's access to his or her own computers, network, and data. The thief demands a ransom in order to restore full access to all computer systems. Money is extorted from business owners usually in the form of a cryptocurrency like Bitcoin which is untraceable. Ransomware attacks have escalated due to their overwhelming success. Normally, a Trojan disguised as a legitimate file, is accidentally downloaded. These can come as email attachments or links. One of the most dangerous Ransomware programs is called "WannaCry worm" and it can travel between computers with no user interaction.

Resolution Time-The amount of time it takes to resolve an issue, most often a security, network, or IT problem.

Response Time-The amount of time it takes for a service provider to respond to a call for service from one of its clients. Most IT providers guarantee specific response times as a perk to selecting their service over others.

Scalability-A company's ability and flexibility to scale up or down as business needs arise.

SLA- A service-level agreement (SLA) is a contract between a customer and a service provider. It outlines the duties and responsibilities of the service provider and the terms of the contract. It includes the signatures of those in management from both entities.

Software bug-An error, fault, or flaw in a computer program that produces an unintended effect. Bugs are usually the result of mistakes by programmers when developing the source code. The term "debugging" is often used to describe the process of fixing software bugs.

Virtualization-Virtual Machines (VM) are created that look and behave exactly like the real thing. This can include servers, networks, operating systems, or storage devices. This allows a company to have a much more sophisticated IT infrastructure at lower costs. It also enables a business to enjoy the benefits of scalability. With virtualization, you can easily scale up or down as needed, so you only pay for services and equipment that you need at that time.

VPN– Virtual private networks (VPN) are built over public infrastructures to provide a higher level of security to the data transmitted. Usually, encryption is used to protect apps and data from intruders as the data is processed across the internet. There are numerous types of VPN, including Advanced VPN, PureVPN, Hotspot Shield, and ExpressVPN.

Zero Day Threats– This is a threat exploiting vulnerabilities within computer security systems. The term is derived from the word "zeroth", which is the time of the exploitation. It occurs either before or on the first day the developer becomes aware of it.

Therefore, there is no security fix for the threat due to the fact that developers are completely unaware of its existence. A number of different vectors can be used in a zero day attack. These attacks are usually instigated by well-known hacker groups who take pride in being able to outsmart developers.

Information Security Glossary of Terms

A

- **Access Control** - Access Control ensures that resources are only granted to those users who are entitled to them.
- **Assurance (or Information Assurance)** - Measure of confidence that the security features, practices, procedures, and architecture of an information system accurately mediate and enforces the security policy.
- **Auditing** - Auditing is the information gathering and analysis of assets to ensure such things as policy compliance and security from vulnerabilities
- **Authentication** - Authentication is the process of confirming the correctness of the claimed identity.
- **Authenticity** - Authenticity is the validity and conformance of the original informatio
- **Authorization** - Authorization is the approval, permission, or empowerment for someone or something to do something.
- **Authorization to Operate (ATO)** - The official management decision given by a senior organizational official to authorize operation of an information system and to explicitly accept the risk to organizational operations (including mission, functions, image, or reputation), organizational assets, individuals, other organizations, and the Nation based on the implementation of an agreed-upon set of security controls. An ATO must be issued to a research organization before it can begin working with federal data associated with a grant or contract.
- **Authorizing Official (AO)** - A senior (federal) official or executive with the authority to formally assume responsibility for operating an information system at an acceptable level of risk to organizational operations (including mission, functions, image, or reputation), organizational assets, individuals, other organizations, and the Nation.
- **Availability** - Availability is the need to ensure that the business purpose of the system can be met and that it is accessible to those who need to use it

- # B
- **Basic Authentication** - Basic Authentication is the simplest web-based authentication scheme that works by sending the username and password with each reques
- **Botnet** - A botnet is a large number of compromised computers that are used to create and send spam or viruses or flood a network with messages as a denial of service attac
- **Business Continuity Plan (BCP)** - A Business Continuity Plan is the plan for emergency response, backup operations, and post-disaster recovery steps that will ensure the availability of critical resources and facilitate the continuity of operations in an emergency situation
- **Business Impact Analysis (BIA)** - A Business Impact Analysis determines what levels of impact to a system are tolerable.
- # C
- **Checksum** - A value that is computed by a function that is dependent on the contents of a data object and is stored or transmitted together with the object, for the purpose of detecting changes in the dat
- **Cipher** - A cryptographic algorithm for encryption and decryption.
- **Ciphertext** - Ciphertext is the encrypted form of the message being sen
- **Confidentiality** - Confidentiality is the need to ensure that information is disclosed only to those who are authorized to view it.
- **Configuration Control Board (CCB)** - A group of qualified people with responsibility for the process of regulating and approving changes to hardware, firmware, software, and documentation throughout the development and operational life cycle of an information system.
- **Configuration Management** - Establish a known baseline condition and manage it.
- **Cookie** - Data exchanged between an HTTP server and a browser (a client of the server) to store state information on the client side and retrieve it later for server use. An HTTP server, when sending data to a client, may send along a cookie, which the client retains after the HTTP connection closes. A server can use this mechanism to maintain persistent client-side state information for HTTP-based applications, retrieving the state information in later connections.
- # D

- **Data Aggregation** - Data Aggregation is the ability to get a more complete picture of the information by analyzing several different types of records at once.
- **Data Custodian** - A Data Custodian is the entity currently using or manipulating the data, and therefore, temporarily taking responsibility for the data
- **Data Steward (Owner)** - A Data Owner is the entity having responsibility and authority for the dat
- **Data User** - A person, organization entity, or automated process that accesses a system, whether authorized to do so or not.
- **Data Warehousing** - Data Warehousing is the consolidation of several previously independent databases into one location.
- **Day Zero** - The "Day Zero" or "Zero Day" is the day a new vulnerability is made known. In some cases, a "zero day" exploit is referred to an exploit for which no patch is available yet. ("day one"-> day at which the patch is made available).
- **Denial of Service** - The prevention of authorized access to a system resource or the delaying of system operations and functions.
- **Dictionary Attack** - An attack that tries all of the phrases or words in a dictionary, trying to crack a password or key. A dictionary attack uses a predefined list of words compared to a brute force attack that tries all possible combination
- **Disaster Recovery Plan (DRP)** - A Disaster Recovery Plan is the process of recovery of IT systems in the event of a disruption or disaster.
- **Due Care** - Due care ensures that a minimal level of protection is in place in accordance with the best practice in the industry.
- **Due Diligence** - Due diligence is the requirement that organizations must develop and deploy a protection plan to prevent fraud, abuse, and additional deploy a means to detect them if they occu
- **Dumpster Diving** - Dumpster Diving is obtaining passwords and corporate directories by searching through discarded media

- # E

- **Encryption** - Cryptographic transformation of data (called "plaintext") into a form (called "cipher text") that conceals the data's original meaning to prevent it from being known or use
- **Event** - An event is an observable occurrence in a system or network
- **Exposure** - A threat action whereby sensitive data is directly released to an unauthorized entit

- # F
- **FERPA**- The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99) is a Federal law that protects the privacy of student education records.
- **File Transfer Protocol (FTP)** - A TCP/IP protocol specifying the transfer of text or binary files across the network.
- **FIPS** - Federal Information Processing Standard
- **Firewall** - A logical or physical discontinuity in a network to prevent unauthorized access to data or resources.
- **FISMA** - The Federal Information Security Management Act (FISMA) is United States legislation that defines a comprehensive framework to provide information security for the information and information systems that support the operations and assets of the agency, including those provided or managed by another agency, contractor, or other source.
- **Flooding** - An attack that attempts to cause a failure in (especially, in the security of) a computer system or other data processing entity by providing more input than the entity can process properl
- # G
- **Gateway** - A network point that acts as an entrance to another network
- **GLBA** - The Gramm-Leach-Bliley Act (GLBA) requires companies that offer consumers financial products or services like loans, financial or investment advice, or insurance to explain their information-sharing practices to their customers and to safeguard sensitive data. Colleges and universities are also subject to some of the provisions of GLBA because they collect and maintain financial information about their students and others with whom they interact.
- # H
- **Hardening** - Hardening is the process of identifying and fixing vulnerabilities on a system
- **HIPAA** – The Health Insurance Portability and Accountability Act of 1996 (HIPAA) is United States legislation that provides data privacy and security provisions for safeguarding medical information.
- **HTTPS** - When used in the first part of a URL (the part that precedes the colon and specifies an access scheme or protocol), this term specifies the use of HTTP enhanced by a security mechanism, which is usually referred to as SSL. Note: Commonly called SSL, standard is TLS 1.2

- **Hyperlink** - In hypertext or hypermedia, an information object (such as a word, a phrase, or an image; usually highlighted by color or underscoring) that points (indicates how to connect) to related information that is located elsewhere and can be retrieved by activating the link.
- **Hypertext Markup Language (HTML)** - The set of markup symbols or codes inserted in a file intended for display on a World Wide Web browser page.
- **Hypertext Transfer Protocol (HTTP)** - The protocol in the Internet Protocol (IP) family used to transport hypertext documents across an internet.

- # I

- **Identity** - Identity is whom someone or what something is, for example, the name by which something is known.
- **Incident** - An incident as an adverse network event in an information system or network or the threat of the occurrence of such an event
- **Incident Handling** - Incident Handling is an action plan for dealing with intrusions, cyber-theft, denial of service, fire, floods, and other security-related events. It is comprised of a six step process: Preparation, Identification, Containment, Eradication, Recovery, and Lessons Learne
- **Information** - Any communication or representation of knowledge, such as facts, data, or opinions, in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual.
- **Information Security** - The protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability.
- **Information System** - A discrete set of information resources organized for the collection, processing, maintenance, use, sharing, dissemination, or disposition of information.
- **Information System Security Officer (ISSO)** - Individual who is assigned responsibility for maintaining the appropriate operational security posture for an information system or program.
- **Integrity** - Integrity is the need to ensure that information has not been changed accidentally or deliberately, and that it is accurate and complete
- **Internet Protocol (IP)** - The method or protocol by which data is sent from one computer to another on the Internet
- **Internet Protocol Security (IPsec)** - A developing standard for security at the network or packet processing layer of network communication

- **Internet Standard** - A specification, approved by the IESG and published as an RFC, that is stable and well-understood, is technically competent, has multiple, independent, and interoperable implementations with substantial operational experience, enjoys significant public support, and is recognizably useful in some or all parts of the Interne
- **Intrusion Detection** - A security management system for computers and networks. An IDS gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization)
- **IP Address** - A computer's inter-network address that is assigned for use by the Internet Protocol and other protocols. An IP version 4 address is written as a series of four 8-bit numbers separated by periods
- **IP Flood** - A denial of service attack that sends a host more echo request ("ping") packets than the protocol implementation can handle.
- **IP Forwarding** - IP forwarding is an Operating System option that allows a host to act as a router. A system that has more than 1 network interface card must have IP forwarding turned on in order for the system to be able to act as a router
- **IP Spoofing** - The technique of supplying a false IP address.
- **IT-SP** - Information Technology Security Plan; see System Security Plan

L

- **Least Privilege** - Least Privilege is the principle of allowing users or applications the least amount of permissions necessary to perform their intended function

M

- **MAC Address** - A physical address; a numeric value that uniquely identifies that network device from every other device on the planet.
- **Malicious Code** - Software (e.g., Trojan horse) that appears to perform a useful or desirable function, but actually gains unauthorized access to system resources or tricks a user into executing other malicious logic
- **Malware** - A generic term for a number of different types of malicious code

N

- **NAT** - Network Address Translation. It is used to share one or a small number of publicly routable IP addresses among a larger number of hosts. The hosts are assigned private IP addresses, which are then "translated" into one of the publicly

routed IP addresses. Typically home or small business networks use NAT to share a single DLS or Cable modem IP address. However, in some cases NAT is used for servers as an additional layer of protection

- **National Institute of Standards and Technology (NIST)** - National Institute of Standards and Technology, a unit of the US Commerce Department. Formerly known as the National Bureau of Standards, NIST promotes and maintains measurement standards. It also has active programs for encouraging and assisting industry and science to develop and use these standard
- **Netmask** - 32-bit number indicating the range of IP addresses residing on a single IP network/subnet/supernet. This specification displays network masks as hexadecimal numbers. For example, the network mask for a class C IP network is displayed as 0xffffff00. Such a mask is often displayed elsewhere in the literature as 255.255.255.0
- **Network Address Translation** - The translation of an Internet Protocol address used within one network to a different IP address known within another network. One network is designated the inside network and the other is the outside
- **Network Mapping** - To compile an electronic inventory of the systems and the services on your networ
- **Non-Repudiation** - Non-repudiation is the ability for a system to prove that a specific user and only that specific user sent a message and that it hasn't been modified

- # O

- **One-Way Encryption** - Irreversible transformation of plaintext to cipher text, such that the plaintext cannot be recovered from the cipher text by other than exhaustive procedures even if the cryptographic key is known.
- **One-Way Function** - A (mathematical) function, f, which is easy to compute the output based on a given input. However given only the output value it is impossible (except for a brute force attack) to figure out what the input value is.

- # P

- **Packet** - A piece of a message transmitted over a packet-switching network. One of the key features of a packet is that it contains the destination address in addition to the data. In IP networks, packets are often called datagrams
- **Packet Switched Network** - A packet switched network is where individual packets each follow their own paths through the network from one endpoint to another.
- **Partitions** - Major divisions of the total physical hard disk spac

- **Password Authentication Protocol (PAP)** - Password Authentication Protocol is a simple, weak authentication mechanism where a user enters the password and it is then sent across the network, usually in the clear.
- **Password Cracking** - Password cracking is the process of attempting to guess passwords, given the password file information
- **Password Sniffing** - Passive wiretapping, usually on a local area network, to gain knowledge of passwords.
- **Patch** - A patch is a small update released by a software manufacturer to fix bugs in existing program
- **Patching** - Patching is the process of updating software to a different version.
- **Payload** - Payload is the actual application data a packet contains.
- **Penetration** - Gaining unauthorized logical access to sensitive data by circumventing a system's protections.
- **Penetration Testing** - Penetration testing is used to test the external perimeter security of a network or facility.
- **Permutation** - Permutation keeps the same letters but changes the position within a text to scramble the message.
- **Personal Firewalls** - Personal firewalls are those firewalls that are installed and run on individual PCs
- **Pharming** - This is a more sophisticated form of MITM attack. A user's session is redirected to a masquerading website. This can be achieved by corrupting a DNS server on the Internet and pointing a URL to the masquerading website's IP. Almost all users use a URL like www.worldbank.com instead of the real IP (192.86.99.140) of the website. Changing the pointers on a DNS server, the URL can be redirected to send traffic to the IP of the pseudo website. At the pseudo website, transactions can be mimicked and information like login credentials can be gathered. With this the attacker can access the real www.worldbank.com site and conduct transactions using the credentials of a valid user on that website.
- **PHI** - Protected health information (PHI) under US law is any information about health status, provision of health care, or payment for health care that is created or collected by a "Covered Entity" (or a Business Associate of a Covered Entity), and can be linked to a specific individual.
- **Phishing** - The use of e-mails that appear to originate from a trusted source to trick a user into entering valid credentials at a fake website. Typically the e-mail and the web site looks like they are part of a bank the user is doing business with.

- **Personally Identifiable Information (PII)** - Information which can be used to distinguish or trace the identity of an individual (e.g., name, social security number, biometric records, etc.) alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual (e.g., date and place of birth, mother's maiden name, etc.).
- **Ping of Death** - An attack that sends an improperly large ICMP echo request packet (a "ping") with the intent of overflowing the input buffers of the destination machine and causing it to crash.
- **Ping Scan** - A ping scan looks for machines that are responding to ICMP Echo Requests.
- **Ping Sweep** - An attack that sends ICMP echo requests ("pings") to a range of IP addresses, with the goal of finding hosts that can be probed for vulnerabilities
- **Plaintext** - Ordinary readable text before being encrypted into ciphertext or after being decrypted.
- **Plan of Action & Milestones (POA&M)** - A document that identifies tasks needing to be accomplished. It details resources required to accomplish the elements of the plan, any milestones in meeting the tasks, and scheduled completion dates for the milestones.
- **Port** - A port is nothing more than an integer that uniquely identifies an endpoint of a communication stream. Only one process per machine can listen on the same port number.
- **Port Scan** - A port scan is a series of messages sent by someone attempting to break into a computer to learn which computer network services, each associated with a "well-known" port number, the computer provides. Port scanning, a favorite approach of computer cracker, gives the assailant an idea where to probe for weaknesses. Essentially, a port scan consists of sending a message to each port, one at a time. The kind of response received indicates whether the port is used and can therefore be probed for weakness.
- **Proprietary Information** - Proprietary information is that information unique to a company and its ability to compete, such as customer lists, technical data, product costs, and trade secrets
- **Protocol** - A formal specification for communicating; an IP address the special set of rules that end points in a telecommunication connection use when they communicate. Protocols exist at several levels in a telecommunication connection.
- **Public Key** - The publicly-disclosed component of a pair of cryptographic keys used for asymmetric cryptography.
- **Public Key Encryption** - The popular synonym for "asymmetric cryptography

- **Public Key Infrastructure (PKI)** - A PKI (public key infrastructure) enables users of a basically unsecured public network such as the Internet to securely and privately exchange data and money through the use of a public and a private cryptographic key pair that is obtained and shared through a trusted authority. The public key infrastructure provides for a digital certificate that can identify an individual or an organization and directory services that can store and, when necessary, revoke the certificate

R

- **Reconnaissance** - Reconnaissance is the phase of an attack where an attackers finds new systems, maps out networks, and probes for specific, exploitable vulnerabilities.
- **Registry** - The Registry in Windows operating systems in the central set of settings and information required to run the Windows computer.
- **Risk** - Risk is the product of the level of threat with the level of vulnerability. It establishes the likelihood of a successful attack.
- **Risk Assessment (RA)** - The process of identifying risks to organizational operations (including mission, functions, image, reputation), organizational assets, individuals, other organizations, and the Nation, resulting from the operation of an information system. Part of risk management, incorporates threat and vulnerability analyses,and considers mitigations provided by security controls planned or in place. Synonymous with risk analysis.
- **Risk Averse** - Avoiding risk even if this leads to the loss of opportunity. For example, using a (more expensive) phone call vs. sending an e-mail in order to avoid risks associated with e-mail may be considered "Risk Averse"
- **Risk Management** - The program and supporting processes to manage information security risk to organizational operations (including mission, functions, image, reputation), organizational assets, individuals, other organizations, and the Nation, and includes: (i) establishing the context for risk-related activities; (ii) assessing risk; (iii) responding to risk once determined; and (iv) monitoring risk over time.
- **Risk Management Framework (RMF)** - A six-step process created by the National Institute of Standards and Technology, detailed in NIST Special Publication 800-37: Guide for Applying the Risk Management Framework to Federal Information Systems.
- **Risk Mitigation** - Prioritizing, evaluating, and implementing the appropriate risk-reducing controls/countermeasures recommended from the risk management process.
- **Root** - Root is the name of the administrator account in Unix systems.

- **Rootkit** - A collection of tools (programs) that a hacker uses to mask intrusion and obtain administrator-level access to a computer or computer network.
- **Router** - Routers interconnect logical networks by forwarding information to other networks based upon IP addresses.

S

- **Safety** - Safety is the need to ensure that the people involved with the company, including employees, customers, and visitors, are protected from harm
- **Secure Shell (SSH)** - A program to log into another computer over a network, to execute commands in a remote machine, and to move files from one machine to another.
- **Security Assessment Report (SAR)** - This deliverable is one of three key documents in the security authorization package developed for authorizing officials. The assessment report includes information from the assessor/auditor that is necessary to determine the effectiveness of the security controls employed within or inherited by the information system based upon the assessor's findings.
- **Security Control** - A safeguard or countermeasure prescribed for an information system or an organization designed to protect the confidentiality, integrity, and availability of its information and to meet a set of defined security requirements.
- **Security Policy** - A set of rules and practices that specify or regulate how a system or organization provides security services to protect sensitive and critical system resources.
- **Sensitive Information** - Sensitive information, as defined by the federal government, is any unclassified information that, if compromised, could adversely affect the national interest or conduct of federal initiatives.
- **Separation of Duties** - Separation of duties is the principle of splitting privileges among multiple individuals or systems.
- **Server** - A system entity that provides a service in response to requests from other system entities called clients.
- **Session** - A session is a virtual connection between two hosts by which network traffic is passed.
- **Session Hijacking** - Take over a session that someone else has established.
- **Sniffer** - A sniffer is a tool that monitors network traffic as it received in a network interfac
- **Sniffing** - A synonym for "passive wiretapping."

- **Social Engineering** - A euphemism for non-technical or low-technology means - such as lies, impersonation, tricks, bribes, blackmail, and threats - used to attack information system
- **Spam** - E-mail that is unsolicited and irrelevant to University business sent out in large quantities.
- **Sub Network** - A separately identifiable part of a larger network that typically represents a certain limited number of host computers, the hosts in a building or geographic area, or the hosts on an individual local area network.
- **Subnet Mask** - A subnet mask (or number) is used to determine the number of bits used for the subnet and host portions of the address. The mask is a 32-bit value that uses one-bits for the network and subnet portions and zero-bits for the host portion
- **Switch** - A switch is a networking device that keeps track of MAC addresses attached to each of its ports so that data is only transmitted on the ports that are the intended recipient of the data.
- **System Owner (SO)** - Official responsible for the overall procurement, development, integration, modification, or operation and maintenance of an information system.
- **System Security Officer (SSO)** - A person responsible for enforcement or administration of the security policy that applies to the system.
- **System Security Plan (SSP)** - Formal document that provides an overview of the security requirements for an information system and describes the security controls in place or planned for meeting those requirements.
- **System-Specific Policy** - A System-specific policy is a policy written for a specific system or device.
- # T
- **TCP/IP** - A synonym for "Internet Protocol Suite;" in which the Transmission Control Protocol and the Internet Protocol are important parts. TCP/IP is the basic communication language or protocol of the Internet. It can also be used as a communications protocol in a private network (either an Intranet or an Extranet).
- **Threat** - A potential for violation of security, which exists when there is a circumstance, capability, action, or event that could breach security and cause harm.
- **Threat Assessment** - A threat assessment is the identification of types of threats that an organization might be exposed to.
- **Threat Model** - A threat model is used to describe a given threat and the harm it could to do a system if it has a vulnerability.

- **Threat Vector** - The method a threat uses to get to the target.
- **Topology** - The geometric arrangement of a computer system. Common topologies include a bus, star, and ring. The specific physical, i.e., real, or logical, i.e., virtual, arrangement of the elements of a network. Note 1: Two networks have the same topology if the connection configuration is the same, although the networks may differ in physical interconnections, distances between nodes, transmission rates, and/or signal types. Note 2: The common types of network topology are illustrated
- **Transport Layer Security (TLS)** - and its predecessor, Secure Sockets Layer (SSL), both frequently referred to as "SSL", are cryptographic protocols that provide communications security over a computer network
- **Trojan Horse** - A computer program that appears to have a useful function, but also has a hidden and potentially malicious function that evades security mechanisms, sometimes by exploiting legitimate authorizations of a system entity that invokes the program.
- **Tunnel** - A communication channel created in a computer network by encapsulating a communication protocol's data packets in (on top of) a second protocol that normally would be carried above, or at the same layer as, the first one. Most often, a tunnel is a logical point-to-point link - i.e., an OSI layer 2 connection - created by encapsulating the layer 2 protocol in a transport protocol (such as TCP), in a network or inter-network layer protocol (such as IP), or in another link layer protocol. Tunneling can move data between computers that use a protocol not supported by the network connecting them.
- # U
- **UDP Scan** - UDP scans perform scans to determine which UDP ports are open.
- **Uniform Resource Locator (URL)** - The global address of documents and other resources on the World Wide Web. The first part of the address indicates what protocol to use, and the second part specifies the IP address or the domain name where the resource is located. For example, http://www.pcwebopedia.com/index.html .
- **Unprotected Share** - In Windows terminology, a "share" is a mechanism that allows a user to connect to file systems and printers on other systems. An "unprotected share" is one that allows anyone to connect to it.
- **User** - A person, organization entity, or automated process that accesses a system, whether authorized to do so or not.
- # V

- **Virtual Private Network (VPN)** - A restricted-use, logical (i.e., artificial or simulated) computer network that is constructed from the system resources of a relatively public, physical (i.e., real) network (such as the Internet), often by using encryption (located at hosts or gateways), and often by tunneling links of the virtual network across the real network. For example, if a corporation has LANs at several different sites, each connected to the Internet by a firewall, the corporation could create a VPN by (a) using encrypted tunnels to connect from firewall to firewall across the Internet and (b) not allowing any other traffic through the firewalls. A VPN is generally less expensive to build and operate than a dedicated real network, because the virtual network shares the cost of system resources with other users of the real network.
- **Virus** - A hidden, self-replicating section of computer software, usually malicious logic, that propagates by infecting - i.e., inserting a copy of itself into and becoming part of - another program. A virus cannot run by itself; it requires that its host program be run to make the virus active.
- **Vulnerability** – A weakness in a system, application, or network that is subject to exploitation or misuse.
- **Vulnerability Management** - process surrounding vulnerability scanning, also taking into account other aspects such as risk acceptance, remediation etc. security risks.
- **Vulnerability scanning** - using a computer program to identify vulnerabilities in networks, computer infrastructure or applications.

- # W

- **Web of Trust** - A web of trust is the trust that naturally evolves as a user starts to trust other's signatures, and the signatures that they trust.
- **Web Server** - A software process that runs on a host computer connected to the Internet to respond to HTTP requests for documents from client web browsers.
- **Wiretapping** - Monitoring and recording data that is flowing between two points in a communication system.
- **World Wide Web ("the Web", WWW, W3)** - The global, hypermedia-based collection of information and services that is available on Internet servers and is accessed by browsers using Hypertext Transfer Protocol and other information retrieval mechanisms.
- **Worm** - A computer program that can run independently, can propagate a complete working version of itself onto other hosts on a network, and may consume computer resources destructively.

- # Z
- **Zero Day** - The "Day Zero" or "Zero Day" is the day a new vulnerability is made known. In some cases, a "zero day" exploit is referred to an exploit for which no patch is available yet. ("day one" - day at which the patch is made available).
- **Zero-day attack** - A zero-day (or zero-hour or day zero) attack or threat is a computer threat that tries to exploit computer application vulnerabilities that are unknown to others or undisclosed to the software developer. Zero-day exploits (actual code that can use a security hole to carry out an attack) are used or shared by attackers before the software developer knows about the vulnerability.

INFORMATION SECURITY GLOSSARY

A

Access Controlled Environment: A physically secured location with appropriate environmental controls accessible only to authorized personnel with a legitimate business need.

Account: That combination of username and password that provides an individual, group, or service with access to a computer system or computer network.

Affiliates: Select individuals who have been afforded contractual affiliate status by the university. *Affiliates* do not receive a salary from the university for the duties and services they perform.

Antivirus Software: Software specifically designed for the detection and prevention of known computer viruses. See also *Antivirus Updates*.

Antivirus Updates: Frequently released definitions that identify new computer viruses. These definitions are used to keep *antivirus software* effective.

Application Administration Account: An account for the administration of an application (e.g., Oracle database administrator, MS-SQL SA administrator).

Associates: Individuals such as unpaid faculty, principal investigators, visiting scholars, dissertation special members and others who are regularly engaged in activities that directly support the teaching and research mission of the university, but who are not compensated by the university by salary.

Attack: An attempt to gain unauthorized access or deny authorized access to a *university information resource*.

Attacker: An entity that attempts to gain unauthorized access or deny authorized access to a u*niversity information resource*.

Authentication: The process of confirming a claimed identity. All forms of authentication are based on something you know, something you have, or something you are:

> "Something you know" is some form of information that you can recognize and keep to yourself, such as a personal identification number (PIN) or password.
> "Something you have" is a physical item you possess, such as a photo ID or a security token.
> "Something you are" is a human characteristic considered to be unique, such as a fingerprint, voice tone, or retinal pattern.

Authorization: The act of granting permission for someone or something to conduct an act. Even when identity and authentication have indicated who someone is, authorization may be needed to establish what actions are permitted.

Authorized User: An individual that is not a UNC Charlotte faculty, staff or student who has been granted permission to access university server, workstation, networked device, or application resources.

Availability: The degree to which information and vital services are accessible for use when required.

B

Business Continuity: The ability to carry out vital business services in a timely manner despite loss of or damage to *university information resources*.

C

CIO: The University's Chief Information Officer.

CISO: The University's Chief Information Security Officer

Compromise: An unauthorized intrusion into a u*niversity information resource* where unauthorized disclosure, modification or destruction of *confidential university data* may have occurred.

Confidentiality: The degree to which *confidential university data* are protected from unauthorized disclosure.

Confidential University Data: *Personally identifiable information*, *proprietary information*, *confidential non-personally identifiable information*, and any other data the disclosure of which could cause significant harm to the University or its constituents.

> **Confidential Non-Personally Identifiable Information** - Information about university-related activities not available to the public by the operation of law. The protection of *confidential non-personally identifiable information* is governed by the university's own policies. Examples may include detailed information about some university buildings, activities, or events, information about future university development plans, and research information.
>
> **Personally Identifiable Information** - Information relating to an individual that reasonably identifies the individual, except where such information is public by operation of university policy or applicable law (e.g., past or present employees' names, titles, positions, salaries or other information designated as public records under the North Carolina State Personnel Act; student names, local addresses and telephone numbers, email addresses and other "directory" information

under FERPA, unless such student has requested nondisclosure consistent with FERPA and University Policy 402, Student Records (FERPA)). Examples may include, but are not limited to: Social Security numbers, payment card numbers, financial account information, NC driver license number, NC non-operating identification license number (State ID card), student grades or disciplinary information, all FERPA non-directory information about students and former students, including citizenship, income tax withholdings, personnel records, relatives' names and addresses, student and employee identification numbers, donations, patient health information, human subject data, information the university has agreed to keep confidential, and account passwords or encryption keys used to protect access to *confidential university data*. Confidentiality of *personally identifiable information* is largely governed by law or contract (e.g., HIPAA, FERPA, GLBA, PCI DSS, and laws governing human subject data).

Proprietary Information - Data, information, or intellectual property in which the university has an exclusive legal interest or ownership right, which, if compromised could cause significant harm to the university. Examples may include, but are not limited to, business planning, financial information, trade secret, copyrighted material, and software or comparable material from a third party when the university has agreed to keep such information confidential.

Contractors: A person who contracts with the university to furnish supplies or perform work at a certain price or rate not paid through Payroll.

Custodian: Guardian or caretaker; the holder of data, the agent charged with implementing the controls specified by the owner. The *custodian* is responsible for the processing and storage of information. The *custodians* of information resources, including

entities providing outsourced information resources services to the university, must:

Implement the controls specified by the owner(s).
Provide physical and procedural safeguards for the information resources.
Assist owners in evaluating the cost-effectiveness of controls and monitoring.
Implement the monitoring techniques and procedures for detecting, reporting, and investigating *incidents*.

D

Data: Information that has been translated into a form that is more convenient to move or process.

Data Classification: Data classification is the conscious decision to assign a level of sensitivity to data as it is being created, enhanced, stored, or transmitted. The classification of the data should determine the extent to which the data needs to be controlled and secured.

Data Facilities: Controlled facilities with a primary focus of housing servers, networking equipment and other *devices*.

Data Owner: Person responsible for managing institutional data owned by the university. The *data owner* has the responsibility for classifying their data in order to ensure that the appropriate steps are taken to protect the *data* and that respective standards and guidelines are being properly implemented.

Data Security Officer: An individual designated by Vice Chancellors, Deans, Directors, or other department heads to serve as the primary contact for ensuring secure management of data in his or her functional area. The DSO assists faculty and staff with Data Security Plans and ensures that each Data Security Plan complies with all university policies and procedures.

Devices: Any apparatus used to access, store, transmit or interface with a *university information resource*. This includes but is

not limited to computers (servers, workstations and laptops), PDAs, printers, network appliances, devices situated behind firewalls, *Network Address Translation* devices, or use of *Virtual Private Networks*.

Disaster Recovery: The ability to restore lost or damaged *data* or systems in a timely manner.

E

Electronic Communication: Transmitting *data* electronically with or without human interaction (i.e., email, web, instant messaging, etc.).

Encrypted: Transformed using an algorithm to make information unreadable to anyone other than those with special knowledge, usually referred to as a key.

Encryption: The process of transforming information using an algorithm to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key.

F

Firewall: A hardware or software device that limits access to a computer or network to help prevent unauthorized access. Also see *Firewall Appliance* and *Host-based Firewall Software*.

Firewall Appliance: A physical device that provides firewall protection for a network. Also see *Firewall* and *Host-based Firewall Software*.

H

Host-based Firewall Software: A software program that provides firewall protection solely for the system on which it is running. Also see *Firewall* and *Firewall* ***Appliance*****.**

I

Information Security Liaison (ISL): Area associate who serves as the intermediary between his/her respective unit and ITS and

assists with implementing information security policy, standards and guidelines.

Incident: An attempted or successful unauthorized access, use, disclosure, modification or destruction of information; interference with information technology operation; or violation of explicit or implied acceptable use policy.

Integrity: The degree to which the accuracy and completeness of information and computer software is safeguarded to protect the business process for the university.

ITS: Information and Technology Services.

L

Local Support Provider: A person with principal responsibility for the installation, configuration, security, and ongoing maintenance of a *device* (e.g., system administrator or network administrator).

Log: Electronic information about activity recorded by a computer during the course of operation.

M

Merchant: Unit that accepts payment cards in payment for goods, services or gifts.

Merchant Account: The payment card account number assigned by the university's office of Financial Services to permit payment card payment processing.

N

Network: A logical collection of devices and communication paths.

Networked Device: Any equipment that resides on a network.

Network Manager: See *System Administrator*.

Non-Compliance: Failure to meet or exceed standards or recommendations set by the University or by individual units.

O

Offsite: Located in a university-approved secure location other than the building in which backups are performed.

P

Password: A string of characters that serves as authentication of a person's identity, that may be used to grant, or deny, access to private or shared data.

Patches: Updates to operating systems and application software that enhance security and/or operability.

Personally Identifiable Information: See *Confidential University Data*.

Personal Information: A person's first name or first initial and last name in combination with any one or more of the following data elements:

- The person's Social Security number
- The person's NC driver license number or non-operating identification license
- The person's financial account number or credit or debit card number in combination with any required security code, access code or password that would permit access to the person's financial account

Protected Health Information (PHI): *Personally Identifiable Information* consisting of health information that can be linked to a particular person and that is protected under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The university's use and protection of PHI is governed by University Policy 605.2, Privacy and Confidentiality of Individually Identifiable Health Care Information under HIPAA.

Physically Secured Location: An access controlled environment that has any or all of the following control measures in place:

Alarm system installed to detect and report break-ins
Networking equipment (switches, cabling, Internet connection) kept in a locked room and/or cabinet with restricted access
Servers kept in a locked room and/or cabinet with restricted access
Laptops secured with cable locks
Sensitive printed materials stored in locked file cabinets
Use of visitor logs, visitor escorts, visitor badges, entrance controls and employee badges to restrict access

S

Scan: A series of messages or transmissions attempting to access a device to learn what network services and information the device provides in order to identify potential weaknesses.

Security Breach: See *Compromise*.

Sensitive Information: Information that must be protected from unauthorized access or disclosure because of laws, regulations, university policy, or by agreement, whether the information is in physical or electronic format.

Server: A system that provides services to others outside their local network.

Site-licensed: Licensed for use by the University at low or no cost to the user.

Special Account: An account that is permitted privileges above and beyond those of normal users. Examples of this type of account include: root, super user, administrator, etc.

Strong Password: A strong password is a password that is designed to be hard for a person, program or automated process to discover. It is normally constructed of a sequence of characters, numbers, and special characters, depending on the capabilities of the operation system. Typically, the longer the password the stronger it is. Passwords should never include a proper name,

dictionary word in any language, or be linked to any personal information such as birthdate, social security number, etc.

System Administrator: Person responsible for the effective operation and maintenance of u*niversity information resources*, including implementation of standard procedures and controls to enforce University Policy 311, Information Security. Students, faculty, staff members may be the system administrators for their own machines.

T

Technical Support Person: See *System Administrator*.

U

UNC Charlotte Community: Any staff, faculty, students, associates, affiliates, contractors, volunteers or visitors who use UNC Charlotte facilities and resources.

University Network: The collection of central and outlying data, voice, and other networks that provides direct access to u*niversity information resources*.

University Information Resource: Data in any form and recorded in any manner and computer-related resources operated, owned or leased by the university, including but not limited to:

Networks and network appliances
Computers (servers, workstations and laptops)
Printers
Software and applications
Thumb drives, paper, etc.
Any other computer-related equipment, device or hardware used to access, store, transmit or interface with another u*niversity information resource*

University-Related Persons: University students and applicants for admission, university employees and applicants for employment, *Affiliates*, *Associates*, volunteers, alumni, temporary employees of agencies who are assigned to work for the university,

and third party contractors engaged by the University and their agents and employees.

V

VPN or Virtual Private Network: An encrypted communication channel between two computers or networks which is intended to prevent eavesdropping between the endpoints. The university offers a free site-licensed VPN to its employees.

Vulnerability: Any flaw in the software, hardware, or configuration of a computing device that can be used to compromise the security of a *university information resource*.

Vulnerability Assessment: An audit by a responsible party that is intended to identify potential vulnerabilities in a computer system or network.

Glossary

Health IT Terms

Glossary of Selected Terms Related to Health IT

Computerized Provider Order Entry (CPOE) – A computer application that allows a physician's orders for diagnostic and treatment services (such as medications, laboratory, and other tests) to be entered electronically instead of being recorded on order sheets or prescription pads. The computer compares the order against standards for dosing, checks for allergies or interactions with other medications, and warns the physician about potential problems.

Consolidated Health Informatics (CHI) Initiative – One of the 24 Presidential eGovernment initiatives with the goal of adopting vocabulary and messaging standards to facilitate communication of clinical information across the federal health enterprise. CHI now falls under FHA.

Decision-Support System (DSS) - Computer tools or applications to assist physicians in clinical decisions by providing evidence-based knowledge in the context of patient specific data. Examples include drug interaction alerts at the time medication is prescribed and reminders for specific guideline-based interventions during the care of patients with chronic disease. Information should be presented in a patient-centric view of individual care and also in a population or aggregate view to support population management and quality improvement.

Electronic Health Record (EHR) – A real-time patient health record with access to evidence-based decision support tools that can be used to aid clinicians in decision making. The EHR can automate and streamline a clinician's workflow, ensuring that all clinical information is communicated. It can also prevent delays in response that result in gaps in care. The EHR can also support the collection of data for uses other than clinical care, such as billing, quality management, outcome reporting, and public health disease surveillance and reporting.

Electronic Prescribing (eRx) – A type of computer technology whereby physicians use handheld or personal computer devices to review drug and formulary coverage and to transmit prescriptions to a printer or to a local pharmacy. E-prescribing software can be integrated into existing clinical

information systems to allow physician access to patient specific information to screen for drug interactions and allergies.

Enterprise Architecture – A strategic resource that aligns business and technology, leverages shared assets, builds internal and external partnerships, and optimizes the value of information technology services.

Federal Health Architecture (FHA) – A collaborative body composed of several federal departments and agencies, including the Department of Health and Human Services (HHS), the Department of Homeland Security (DHS), the Department of Veterans Affairs (VA), the Environmental Protection Agency (EPA), the United States Department of Agriculture (USDA), the Department of Defense (DoD), and the Department of Energy (DOE). FHA provides a framework for linking health business processes to technology solutions and standards and for demonstrating how these solutions achieve improved health performance outcomes.

Health Information Technology (HIT) – The application of information processing involving both computer hardware and software that deals with the storage, retrieval, sharing, and use of health care information, data, and knowledge for communication and decision making.

Personal Health Record (PHR) – An electronic application through which individuals can maintain and manage their health information (and that of others for whom they are authorized) in a private, secure, and confidential environment.

802.11b: Standard for high-speed, wireless LANs that can transmit up to 11 Mbps within a 30-50-meter area, providing a low-cost flexible technology for connecting workgroups and providing mobile Internet access.

A

acceptance testing: Provides the final certification that the system is ready to be used in a production setting.

access control: Policies and procedures to prevent improper access to systems.

access point: Box consisting of a radio receiver/transmitter and antennae that link to a wired network, router, or hub.

accountability: The mechanisms for assessing responsibility for decisions made and actions taken.

accounting rate of return on investment (ROI): Calculation of the rate of return on an investment by adjusting cash inflows produced by the investment for depreciation. Approximates the accounting income earned by the investment.

accumulated balance digital payment systems: Systems enabling users to make micropayments and purchases on the Web, accumulating a debit balance on their credit card or telephone bills.

activity-based costing: Model for identifying all the company activities that cause costs to occur while producing a specific product or service so that

managers can see which products or services are profitable or losing money and make changes to maximize firm profitability.

administrative controls: Formalized standards, rules, procedures, and disciplines to ensure that the organization's controls are properly executed and enforced.

agency theory: Economic theory that views the firm as a nexus of contracts among self-interested individuals who must be supervised and managed.

AI shell: The programming environment of an expert system.

analog signal: A continuous waveform that passes through a communications medium; used for voice communications.

analytical CRM: Customer relationship management applications dealing with the analysis of customer data to provide information for improving business performance.

antivirus software: Software designed to detect, and often eliminate, computer viruses from an information system.

application controls: Specific controls unique to each computerized application.

application server: Software that handles all application operations between browser-based computers and a company's back-end business applications or databases.

application service provider (ASP): Company providing software that can be rented by other companies over the Web or a private network.

application software package: A set of prewritten, precoded application software programs that are commercially available for sale or lease.

application software: Programs written for a specific application to perform functions specified by end users.

arithmetic-logic unit (ALU): Component of the CPU that performs the computer's principal logic and arithmetic operations.

artificial intelligence (AI): The effort to develop computer-based systems that can behave like humans, with the ability to learn languages, accomplish

physical tasks, use a perceptual apparatus, and emulate human expertise and decision making.

asynchronous transfer mode (ATM): A networking technology that parcels information into 8-byte cells, allowing data to be transmitted between computers from different vendors at any speed.

attribute: A piece of information describing a particular entity.

authentication: The ability of each party in a transaction to ascertain the identity of the other party.

automation: Using the computer to speed up the performance of existing tasks.

B

backbone: Part of a network handling the major traffic and providing the primary path for traffic flowing to or from other networks.

backward chaining: A strategy for searching the rule base in an expert system that acts like a problem solver by beginning with a hypothesis and seeking out more information until the hypothesis is either proved or disproved.

balanced scorecard: Model for analyzing firm performance which supplements traditional financial measures with measurements from additional business perspectives, such as customers, internal business processes, and learning and growth.

bandwidth: The capacity of a communications channel as measured by the difference between the highest and lowest frequencies that can be transmitted by that channel.

banner ad: A graphic display on a Web page used for advertising. The banner is linked to the advertiser's Web site so that a person clicking on it will be transported to the advertiser's Web site.

batch processing: A method of collecting and processing data in which transactions are accumulated and stored until a specified time when it is convenient or necessary to process them as a group.

baud: A change in signal from positive to negative or vice versa that is used as a measure of transmission speed.

behavioral models: Descriptions of management based on behavioral scientists' observations of what managers actually do in their jobs.

benchmarking: Setting strict standards for products, services, or activities and measuring organizational performance against those standards.

best practices: The most successful solutions or problem-solving methods that have been developed by a specific organization or industry.

biometric authentication: Technology for authenticating system users that compares a person's unique characteristics such as fingerprints, face, or retinal image, against a stored set profile of these characteristics.

bit: A binary digit representing the smallest unit of data in a computer system. It can only have one of two states, representing 0 or 1.

bluetooth: Standard for wireless personal area networks that can transmit up to 722 Kbps within a 10-meter area.

broadband: High-speed transmission technology. Also designates a single communications medium that can transmit multiple channels of data simultaneously.

bugs: Program code defects or errors.

bullwhip effect: Distortion of information about the demand for a product as it passes from one entity to the next across the supply chain.

bundling: Cross-selling in which a combination of products is sold as a bundle at a price lower than the total cost of the individual products.

bureaucracy: Formal organization with a clear-cut division of labor, abstract rules and procedures, and impartial decision making that uses technical qualifications and professionalism as a basis for promoting employees.

bureaucratic models of decision making: Models of decision making where decisions are shaped by the organization's standard operating procedures (SOPs).

bus network: Network topology linking a number of computers by a single circuit with all messages broadcast to the entire network.

business continuity planning: Planning that focuses on how the company can restore business operations after a disaster strikes.

business functions: Specialized tasks performed in a business organization, including manufacturing and production, sales and marketing, finance and accounting, and human resources.

business intelligence: Applications and technologies that focus on gathering, storing, analyzing, and providing access to data from many different sources to help users make better business decisions.

business model: An abstraction of what an enterprise is and how the enterprise delivers a product or service, showing how the enterprise creates wealth.

business process management: Methodology for revising the organization's business processes to use business processes as fundamental building blocks of corporate information systems.

business process reengineering: The radical redesign of business processes, combining steps to cut waste and eliminating repetitive, paper-intensive tasks in order to improve cost, quality, and service, and to maximize the benefits of information technology.

business processes: The unique ways in which organizations coordinate and organize work activities, information, and knowledge to produce a product or service.

business-to-business (B2B) electronic commerce: Electronic sales of goods and services among businesses.

business-to-consumer (B2C) electronic commerce: Electronic retailing of products and services directly to individual consumers.

byte: A string of bits, usually eight, used to store one number or character in a computer system.

C

C: A powerful programming language with tight control and efficiency of execution; is portable across different microprocessors and is used primarily with PCs.

C ++: Object-oriented version of the C programming language.

cable modem: Modem designed to operate over cable TV lines to provide high-speed access to the Web or corporate intranets.

call center: An organizational department responsible for handling customer service issues by telephone and other channels.

capacity planning: The process of predicting when a computer hardware system becomes saturated to ensure that adequate computing resources are available for work of different priorities and that the firm has enough computing power for its current and future needs.

capital budgeting: The process of analyzing and selecting various proposals for capital expenditures.

carpal tunnel syndrome (CTS): Type of RSI in which pressure on the median nerve through the wrist's bony carpal tunnel structure produces pain.

case-based reasoning (CBR): Artificial intelligence technology that represents knowledge as a database of cases and solutions.

CD-ROM (compact disk read-only memory): Read-only optical disk storage used for imaging, reference, and database applications with massive amounts of unchanging data and for multimedia.

CD-RW (CD-ReWritable): Optical disk storage that can be rewritten many times by users.

cellular telephone: A device that transmits voice or data, using radio waves to communicate with radio antennas placed within adjacent geographic areas called cells.

central processing unit (CPU): Area of the computer system that manipulates symbols, numbers, and letters, and controls the other parts of the computer system.

centralized processing: Processing that is accomplished by one large central computer.

change agent: In the context of implementation, the individual acting as the catalyst during the change process to ensure successful organizational adaptation to a new system or innovation.

channel conflict: Competition between two or more different distribution chains used to sell the products or services of the same company.

channel: The link by which data or voice are transmitted between sending and receiving devices in a network.

chatting: Live, interactive conversations over a public network.

chief information officer (CIO): Senior manager in charge of the information systems function in the firm.

chief knowledge officer (CKO): Senior executive in charge of the organization's knowledge management program.

choice: Simon's third stage of decision making, when the individual selects among the various solution alternatives.

churn rate: Measurement of the number of customers who stop using or purchasing products or services from a company. Used as an indicator of the growth or decline of a firm’s customer base.

classical model of management: Traditional description of management that focused on its formal functions of planning, organizing, coordinating, deciding, and controlling.

clicks-and-mortar: Business model where the Web site is an extension of a traditional bricks-and-mortar business.

clickstream tracking: Tracking data about customer activities at Web sites and storing them in a log.

client: The user point-of-entry for the required function in client/server computing. Normally a desktop computer, workstation, or laptop computer.

client/server computing: A model for computing that splits processing between clients and servers on a network, assigning functions to the machine most able to perform the function.

clustering: Linking two computers together so that the second computer can act as a backup to the primary computer or speed up processing.

coaxial cable: A transmission medium consisting of thickly insulated copper wire; can transmit large volumes of data quickly.

COBOL (COmmon Business Oriented Language): Major programming language for business applications because it can process large data files with alphanumeric characters.

cognitive style: Underlying personality dispositions toward the treatment of information, selection of alternatives, and evaluation of consequences.

collaborative commerce: The use of digital technologies to enable multiple organizations to collaboratively design, develop, build and manage products through their life cycles.

collaborative filtering: Tracking users' movements on a Web site, comparing the information gleaned about a user's behavior against data about other customers with similar interests to predict what the user would like to see next.

collaborative planning, forecasting, and replenishment (CPFR): Firms collaborating with their suppliers and buyers to formulate demand forecasts, develop production plans, and coordinate shipping, warehousing, and stocking activities.

communications technology: Physical devices and software that link various computer hardware components and transfer data from one physical location to another.

competitive forces model: Model used to describe the interaction of external influences, specifically threats and opportunities, that affect an organization's strategy and ability to compete.

compiler: Special system software that translates a high-level language into machine language for execution by the computer.

complementary assets: Additional assets required to derive value from a primary investment.

component-based development: Building large software systems by combining pre-existing software components.

computer: Physical device that takes data as an input, transforms the data by executing stored instructions, and outputs information to a number of devices.

computer abuse: The commission of acts involving a computer that may not be illegal but are considered unethical.

computer crime: The commission of illegal acts through the use of a computer or against a computer system.

computer forensics: The scientific collection, examination, authentication, preservation, and analysis of data held on or retrieved from computer storage media in such a way that the information can be used as evidence in a court of law.

computer hardware: Physical equipment used for input, processing, and output activities in an information system.

computer literacy: Knowledge about information technology, focusing on understanding of how computer-based technologies work.

computer software: Detailed, preprogrammed instructions that control and coordinate the work of computer hardware components in an information system.

computer virus: Rogue software programs that are difficult to detect which spread rapidly through computer systems, destroying data or disrupting processing and memory systems.

computer vision syndrome (CVS): Eyestrain condition related to computer display screen use; symptoms include headaches, blurred vision, and dry and irritated eyes.

computer-aided design (CAD): Information system that automates the creation and revision of designs using sophisticated graphics software.

computer-aided software engineering (CASE): Automation of step-by-step methodologies for software and systems development to reduce the amounts of repetitive work the developer needs to do.

computer-based information systems (CBIS): Information systems that rely on computer hardware and software for processing and disseminating information.

conceptual schema: The logical description of the entire database showing all the data elements and relationships among them.

connectivity: The ability of computers and computer-based devices to communicate with each other and share information in a meaningful way without human intervention..

consumer-to-consumer (C2C): electronic commerce Consumers selling goods and services electronically to other consumers.

control unit: Component of the CPU that controls and coordinates the other parts of the computer system.

controls: All of the methods, policies, and procedures that ensure protection of the organization's assets, accuracy and reliability of its records, and operational adherence to management standards.

converged network: Network with technology to enable voice and data to run over a single network.

conversion: The process of changing from the old system to the new system.

cookies: Tiny file deposited on a computer hard drive when an individual visits certain Web sites. Used to identify the visitor and track visits to the Web site.

cooptation: Bringing the opposition into the process of designing and implementing a solution without giving up control of the direction and nature of the change.

copyright: A statutory grant that protects creators of intellectual property against copying by others for any purpose for a minimum of 70 years.

core competency: Activity at which a firm excels as a world-class leader.

core systems: Systems that support functions that are absolutely critical to the organization.

cost-benefit ratio: A method for calculating the returns from a capital expenditure by dividing total benefits by total costs.

counterimplementation: A deliberate strategy to thwart the implementation of an information system or an innovation in an organization.

critical success factors (CSFs): A small number of easily identifiable operational goals shaped by the industry, the firm, the manager, and the broader environment that are believed to assure the success of an organization. Used to determine the information requirements of an organization.

cross-selling: Marketing complementary products to customers.

customer-decision-support system (CDSS): System to support the decision-making process of an existing or potential customer.

customer lifetime value (CLTV): Difference between revenues produced by a specific customer and the expenses for acquiring and servicing that customer minus the cost of promotional marketing over the lifetime of the customer relationship, expressed in today's dollars.

customer relationship management (CRM): Business and technology discipline that uses information systems to coordinate all of the business processes surrounding the firm's interactions with its customers in sales, marketing, and service.

customer relationship management systems: Information systems that track all the ways in which a company interacts with its customers and analyze these interactions to optimize revenue, profitability, customer satisfaction, and customer retention.

customization: The modification of a software package to meet an organization's unique requirements without destroying the package software's integrity.

D

data: Streams of raw facts representing events occurring in organizations or the physical environment before they have been organized and arranged into a form that people can understand and use.

data administration: A special organizational function for managing the organization's data resources, concerned with information policy, data planning, maintenance of data dictionaries, and data quality standards.

data definition language: The component of a database management system that defines each data element as it appears in the database.

data dictionary: An automated or manual tool for storing and organizing information about the data maintained in a database.

data-driven DSS: A system that supports decision making by allowing users to extract and analyze useful information that was previously buried in large databases.

data element: A field.

data flow diagram (DFD): Primary tool for structured analysis that graphically illustrates a system's component process and the flow of data between them.

data inconsistency: The presence of different values for same attribute when the same data are stored in multiple locations.

data management software: Software used for creating and manipulating lists, creating files and databases to store data, and combining information for reports.

data manipulation language: A language associated with a database management system that end users and programmers use to manipulate data in the database.

data mart: A small data warehouse containing only a portion of the organization's data for a specified function or population of users.

data quality audit: A survey and/or sample of files to determine accuracy and completeness of data in an information system.

data redundancy: The presence of duplicate data in multiple data files.

data security controls: Controls to ensure that data files on either disk or tape are not subject to unauthorized access, change, or destruction.

data visualization: Technology for helping users see patterns and relationships in large amounts of data by presenting the data in graphical form.

data warehouse: A database, with reporting and query tools, that stores current and historical data extracted from various operational systems and consolidated for management reporting and analysis.

data workers: People such as secretaries or bookkeepers who process the organization's paperwork.

database: A group of related files.

database (rigorous definition): A collection of data organized to service many applications at the same time by storing and managing data so that they appear to be in one location.

database administration: Refers to the more technical and operational aspects of managing data, including physical database design and maintenance.

database management system (DBMS): Special software to create and maintain a database and enable individual business applications to extract the data they need without having to create separate files or data definitions in their computer programs.

database server: A computer in a client/server environment that is responsible for running a DBMS to process SQL statements and perform database management tasks.

dataconferencing: Teleconferencing in which two or more users are able to edit and modify data files simultaneously.

datamining: Analysis of large pools of data to find patterns and rules that can be used to guide decision making and predict future behavior.

debugging: The process of discovering and eliminating the errors and defects--bugs--in program code.

decisional roles: Mintzberg's classification for managerial roles where managers initiate activities, handle disturbances, allocate resources, and negotiate conflicts.

decision-support systems (DSS): Information systems at the organization's management level that combine data and sophisticated analytical models or data analysis tools to support semistructured and unstructured decision making.

dedicated lines: Telephone lines that are continuously available for transmission by a lessee. Typically conditioned to transmit data at high speeds for high-volume applications.

demand planning: Determining how much product a business needs to make to satisfy all its customers' demands.

denial of service (DoS) attack: Flooding a network server or Web server with false communications or requests for services in order to crash the network.

dense wave division multiplexing (DWDM): Technology for boosting transmission capacity of optical fiber by using many different wavelengths to carry separate streams of data over the same fiber strand at the same time.

Descartes' rule of change: A principle that states that if an action cannot be taken repeatedly, then it is not right to be taken at any time.

design: Simon's second stage of decision making, when the individual conceives of possible alternative solutions to a problem.

desktop publishing software: Software for producing professional-quality documents with capabilities for design, layout, and work with graphics.

development methodology: A collection of methods, one or more for every activity within every phase of a development project.

digital cash: Currency that is represented in electronic form that moves outside the normal network of money.

digital certificate: An attachment to an electronic message to verify the identity of the sender and to provide the receiver with the means to encode a reply.

digital checking: Systems that extend the functionality of existing checking accounts so they can be used for online shopping payments.

digital credit card payment system: Secure services for credit card payments on the Internet that protect information transmitted among users, merchant sites, and processing banks.

digital divide: Large disparities in access to computers and the Internet among different social groups and different locations.

digital firm: Organization where nearly all significant business processes and relationships with customers, suppliers, and employees are digitally enabled, and key corporate assets are managed through digital means.

digital market: A marketplace that is created by computer and communication technologies that link many buyers and sellers.

Digital Millennium Copyright Act (DMCA): Adjusts copyright laws to the Internet Age by making it illegal to make, distribute, or use devices that circumvent technology-based protections of copy-righted materials.

digital signal: A discrete waveform that transmits data coded into two discrete states as 1-bits and 0-bits, which are represented as on-off electrical pulses; used for data communications.

digital signature: A digital code that can be attached to an electronically transmitted message to uniquely identify its contents and the sender.

digital subscriber line (DSL): A group of technologies providing high-capacity transmission over existing copper telephone lines.

digital video disk (DVD): High-capacity optical storage medium that can store full-length videos and large amounts of data.

digital wallet: Software that stores credit card, electronic cash, owner identification, and address information and provides this data automatically during electronic commerce purchase transactions.

direct cutover: A risky conversion approach where the new system completely replaces the old one on an appointed day.

disaster recovery planning: Planning for the restoration of computing and communications services after they have been disrupted.

disintermediation: The removal of organizations or business process layers responsible for certain intermediary steps in a value chain.

distance learning: Education or training delivered over a distance to individuals in one or more locations.

distributed database: A database that is stored in more than one physical location. Parts or copies of the database are physically stored in one location, and other parts or copies are stored and maintained in other locations.

distributed processing: The distribution of computer processing work among multiple computers linked by a communications network.

documentation: Descriptions of how an information system works from either a technical or end-user standpoint.

domain name: The name identifying a unique node on the Internet.

Domain Name System (DNS): A hierarchical system of servers maintaining database enabling the conversion of domain names to their IP addresses.

domestic exporter: Form of business organization characterized by heavy centralization of corporate activities in the home county of origin.

downsizing: The process of transferring applications from large computers to smaller ones.

downtime: Period of time in which an information system is not operational.

drill down: The ability to move from summary data to lower and lower levels of detail.

DSS database: A collection of current or historical data from a number of applications or groups. Can be a small PC database or a massive data warehouse.

DSS software system: Collection of software tools that are used for data analysis, such as OLAP tools, datamining tools, or a collection of mathematical and analytical models.

due process: A process in which laws are well-known and understood and there is an ability to appeal to higher authorities to ensure that laws are applied correctly.

Dynamic page generation: Technology for storing the contents of Web pages as objects in a database where they can be accessed and assembled to create constantly changing Web pages.

dynamic pricing: Pricing of items based on real-time interactions between buyers and sellers that determine what a item is worth at any particular moment.

E

e-government: Use of the Internet and related technologies to digitally enable government and public sector agencies' relationships with citizens, businesses, and other arms of government.

e-learning: Instruction delivered through purely digital technology, such as CD-ROMs, the Internet, or private networks.

efficient customer response system: System that directly links consumer behavior back to distribution, production, and supply chains.

electronic billing and payment presentation system: Systems used for paying routine monthly bills that allow users to view their bills electronically and pay them through electronic funds transfers from banks or credit card accounts.

electronic business (e-business): The use of the Internet and digital technology to execute all the business processes in the enterprise. Includes e-commerce as well as processes for the internal management of the firm and for coordination with suppliers and other business partners.

electronic commerce server software: Software that provides functions essential for running e-commerce Web sites, such as setting up electronic catalogs and storefronts, and mechanisms for processing customer purchases.

electronic commerce: The process of buying and selling goods and services electronically involving transactions using the Internet, networks, and other digital technologies.

electronic data interchange (EDI): The direct computer-to-computer exchange between two organizations of standard business transaction documents.

electronic mail (e-mail): The computer-to-computer exchange of messages.

electronic payment system: The use of digital technologies, such as credit cards, smart cards and Internet-based payment systems, to pay for products and services electronically.

encryption: The coding and scrambling of messages to prevent their being read or accessed without authorization.

end users: Representatives of departments outside the information systems group for whom applications are developed.

end-user development: The development of information systems by end users with little or no formal assistance from technical specialists.

end-user interface: The part of an information system through which the end user interacts with the system, such as on-line screens and commands.

enterprise analysis: An analysis of organization-wide information requirements by looking at the entire organization in terms of organizational units, functions, processes, and data elements; helps identify the key entities and attributes in the organization's data.

enterprise applications: Systems that can coordinate activities, decisions, and knowledge across many different functions, levels, and business units in a firm. Include enterprise systems, supply chain management systems, and knowledge management systems.

enterprise application integration (EAI) software: Software that works with specific software platforms to tie together multiple applications to support enterprise integration.

enterprise networking: An arrangement of the organization's hardware, software, network, and data resources to put more computing power on the desktop and create a company-wide network linking many smaller networks.

enterprise portal: Web interface providing a single entry point for accessing organizational information and services, including information from various enterprise applications and in-house legacy systems so that information appears to be coming from a single source.

enterprise software: Set of integrated modules for applications such as sales and distribution, financial accounting, investment management, materials

management, production planning, plant maintenance, and human resources that allow data to be used by multiple functions and business processes.

enterprise systems: Integrated enterprise-wide information systems that coordinate key internal processes of the firm.

entity: A person, place, thing, or event about which information must be kept.

entity-relationship diagram: A methodology for documenting databases illustrating the relationship between various entities in the database.

ergonomics: The interaction of people and machines in the work environment, including the design of jobs, health issues, and the end-user interface of information systems.

ethical "no free lunch" rule: Assumption that all tangible and intangible objects are owned by someone else, unless there is a specific declaration otherwise, and that the creator wants compensation for this work.

ethics: Principles of right and wrong that can be used by individuals acting as free moral agents to make choices to guide their behavior.

exchange: Third-party Net marketplace that is primarily transaction oriented and that connects many buyers and suppliers for spot purchasing.

executive support systems (ESS): Information systems at the organization's strategic level designed to address unstructured decision making through advanced graphics and communications.

expert system: Knowledge-intensive computer program that captures the expertise of a human in limited domains of knowledge.

explicit knowledge: Knowledge that has been documented.

external integration tools: Project management technique that links the work of the implementation team to that of users at all organizational levels.

extranet: Private intranet that is accessible to authorized outsiders.

F

facsimile (fax): A machine that digitizes and transmits documents with both text and graphics over telephone lines.

Fair Information Practices (FIP): A set of principles originally set forth in 1973 that governs the collection and use of information about individuals and forms the basis of most U.S. and European privacy laws.

fault-tolerant computer systems: Systems that contain extra hardware, software, and power supply components that can back a system up and keep it running to prevent system failure.

feasibility study: As part of the systems analysis process, the way to determine whether the solution is achievable, given the organization's resources and constraints.

feedback: Output that is returned to the appropriate members of the organization to help them evaluate or correct input.

fiber-optic cable: A fast, light, and durable transmission medium consisting of thin strands of clear glass fiber bound into cables. Data are transmitted as light pulses.

field: A grouping of characters into a word, a group of words, or a complete number, such as a person's name or age.

file transfer protocol (FTP): Tool for retrieving and transferring files from a remote computer.

file: A group of records of the same type.

Finance and accounting information systems: Systems keep track of the firm's financial assets and fund flows.

firewall: Hardware and software placed between an organization's internal network and an external network to prevent outsiders from invading private networks.

floppy disk: Removable magnetic disk storage primarily used with PCs.

focused differentiation: Competitive strategy for developing new market niches for specialized products or services where a business can compete in the target area better than its competitors.

formal control tools: Project management technique that helps monitor the progress toward completion of a task and fulfillment of goals.

formal planning tools: Project management technique that structures and sequences tasks, budgeting time, money, and technical resources required to complete the tasks.

formal system: System resting on accepted and fixed definitions of data and procedures, operating with predefined rules.

forward chaining: A strategy for searching the rule base in an expert system that begins with the information entered by the user and searches the rule base to arrive at a conclusion.

fourth-generation language: A programming language that can be employed directly by end users or less-skilled programmers to develop computer applications more rapidly than conventional programming languages.

frame relay: A shared network service technology that packages data into bundles for transmission but does not use error-correction routines. Cheaper and faster than packet switching.

framing: Displaying the content of another Web site inside one's own Web site within a frame or a window.

franchiser: Form of business organization in which a product is created, designed, financed, and initially produced in the home country, but for product-specific reasons relies heavily on foreign personnel for further production, marketing, and human resources.

fuzzy logic: Rule-based AI that tolerates imprecision by using nonspecific terms called membership functions to solve problems.

G

"garbage can" model: Model of decision making that states that organizations are not rational and that decisions are solutions that become attached to problems for accidental reasons.

general controls: Overall controls that establish a framework for controlling the design, security, and use of computer programs throughout an organization.

genetic algorithms: Problem-solving methods that promote the evolution of solutions to specified problems using the model of living organisms adapting to their environment.

geographic information system (GIS): System with software that can analyze and display data using digitized maps to enhance planning and decision-making.

graphical user interface (GUI): The part of an operating system users interact with that uses graphic icons and the computer mouse to issue commands and make selections.

grid computing: Applying the resources of many computers in a network to a single problem.

group decision-support system (GDSS): An interactive computer-based system to facilitate the solution to unstructured problems by a set of decision makers working together as a group.

groupware: Software that provides functions and services that support the collaborative activities of work groups.

H

hacker: A person who gains unauthorized access to a computer network for profit, criminal mischief, or personal pleasure.

hard disk: Magnetic disk resembling a thin steel platter with a metallic coating; used in large computer systems and in most PCs.

hierarchical DBMS: One type of logical database model that organizes data in a treelike structure. A record is subdivided into segments that are connected to each other in one-to-many parent-child relationships.

high-availability computing: Tools and technologies ,including backup hardware resources, to enable a system to recover quickly from a crash.

hit: An entry into a Web server's log file generated by each request to the server for a file.

home page: A World Wide Web text and graphical screen display that welcomes the user and explains the organization that has established the page.

hot spot: A specific geographic location in which an access point provides public Wi-Fi network service.

human resources information systems: Systems that maintain employee records, track employee skills, job performance and training, and support planning for employee compensation and career development.

hybrid AI systems: Integration of multiple AI technologies into a single application to take advantage of the best features of these technologies.

hypermedia database: An approach to data management that organizes data as a network of nodes linked in any pattern the user specifies; the nodes can contain text, graphics, sound, full-motion video, or executable programs.

hypertext markup language (HTML): Page description language for creating Web pages and other hypermedia documents.

hypertext transport protocol: The communications standard used to transfer pages on the Web. Defines how messages are formatted and transmitted.

I

identity theft: Theft of key pieces of personal information, such as credit card or Social Security numbers, in order to obtain merchandise and services in the name of the victim or to obtain false credentials.

Immanuel Kant's Categorical Imperative: A principle that states that if an action is not right for everyone to take it is not right for anyone.

I-mode: Standard developed by Japan's NTT DoCoMo mobile phone network for enabling cell phones to received Web-based content and services.

implementation: Simon's final stage of decision-making, when the individual puts the decision into effect and reports on the progress of the solution.

industry structure: The nature of participants in an industry and their relative bargaining power. Derives from the competitive forces and establishes the general business environment in an industry and the overall profitability of doing business in that environment.

inference engine: The strategy used to search through the rule base in an expert system; can be forward or backward chaining.

information: Data that have been shaped into a form that is meaningful and useful to human beings.

information appliance: Device that has been customized to perform a few specialized computing tasks well with minimal user effort.

information architecture: The particular design that information technology takes in a specific organization to achieve selected goals or functions.

information asymmetry: Situation where the relative bargaining power of two parties in a transaction is determined by one party in the transaction possessing more information essential to the transaction than the other party.

information center: A special facility within an organization that provides training and support for end-user computing.

information partnership: Cooperative alliance formed between two or more corporations for the purpose of sharing information to gain strategic advantage.

information policy: Formal rules governing the maintenance, distribution, and use of information in an organization.

information requirements: A detailed statement of the information needs that a new system must satisfy; identifies who needs what information, and when, where, and how the information is needed.

information rights: The rights that individuals and organizations have with respect to information that pertains to themselves.

information system: Interrelated components working together to collect, process, store, and disseminate information to support decision making, coordination, control, analysis, and visualization in an organization.

information systems department: The formal organizational unit that is responsible for the information systems function in the organization.

information systems literacy: Broad-based understanding of information systems that includes behavioral knowledge about organizations and individuals using information systems as well as technical knowledge about computers.

information systems managers: Leaders of the various specialists in the information systems department.

information systems plan: A road map indicating the direction of systems development: the rationale, the current situation, the management strategy, the implementation plan, and the budget.

information technology (IT) infrastructure: Computer hardware, software, data, storage technology, and networks providing a portfolio of shared IT resources for the organization.

informational roles: Mintzberg's classification for managerial roles where managers act as the nerve centers of their organizations, receiving and disseminating critical information.

informed consent: Consent given with knowledge of all the facts needed to make a rational decision.

input: The capture or collection of raw data from within the organization or from its external environment for processing in an information system.

input controls: The procedures to check data for accuracy and completeness when they enter the system.

instant messaging: Chat service that allows participants to create their own private chat channels so that a person can be alerted whenever someone on his or her private list is on-line to initiate a chat session with that particular individual.

intangible benefits: Benefits that are not easily quantified; they include more efficient customer service or enhanced decision making.

Integrated Services Digital Network (ISDN): International standard for transmitting voice, video, image, and data to support a wide range of service over the public telephone lines.

integrated software package: A software package that provides two or more applications, such as word processing and spreadsheets, providing for easy transfer of data between them.

intellectual property: Intangible property created by individuals or corporations that is subject to protections under trade secret, copyright, and patent law.

intelligence: The first of Simon's four stages of decision making, when the individual collects information to identify problems occurring in the organization.

intelligent agent: Software program that uses a built-in or learned knowledge base to carry out specific, repetitive, and predictable tasks for an individual user, business process, or software application.

internal integration tools: Project management technique that ensures that the implementation team operates as a cohesive unit.

Internal Rate of Return (IRR): The rate of return or profit that an investment is expected to earn.

Internet Protocol (IP) address: Four-part numeric address indicating a unique computer location on the Internet.

Internet Service Provider (ISP): A commercial organization with a permanent connection to the Internet that sells temporary connections to subscribers.

Internet telephony: Technologies that use the Internet Protocol's packet-switched connections for voice service.

Internet: International network of networks that is a collection of hundreds of thousands of private and public networks.

Internet2: Research network with new protocols and transmission speeds that provides an infrastructure for supporting high-bandwidth Internet applications.

internetworking: The linking of separate networks, each of which retains its own identity, into an interconnected network.

interorganizational systems: Information systems that automate the flow of information across organizational boundaries and link a company to its customers, distributors, or suppliers.

interpersonal roles: Mintzberg's classification for managerial roles where managers act as figureheads and leaders for the organization.

intranet: An internal network based on Internet and World Wide Web technology and standards.

intrusion detection system: Tools to monitor the most vulnerable points in a network to detect and deter unauthorized intruders.

intuitive decision makers: Cognitive style that describes people who approach a problem with multiple methods in an unstructured manner, using trial and error to find a solution.

investment workstation: Powerful desktop computer for financial specialists, which is optimized to access and manipulate massive amounts of financial data.

iteration construct: The logic pattern in programming where certain actions are repeated while a specified condition occurs or until a certain condition is met.

iterative: A process of repeating over and over again the steps to build a system.

J

Java: Programming language that can deliver only the software functionality needed for a particular task, such as a small applet downloaded from a network; can run on any computer and operating system.

Joint Application Design (JAD): Process to accelerate the generation of information requirements by having end users and information systems specialists work together in intensive interactive design sessions.

just-in-time: Scheduling system for minimizing inventory by having components arrive exactly at the moment they are needed and finished goods shipped as soon as they leave the assembly line.

K

key field: A field in a record that uniquely identifies instances of that record so that it can be retrieved, updated, or sorted.

knowledge: Concepts, experience, and insight that provide a framework for creating, evaluating, and using information.

knowledge- and information-intense products: Products that require a great deal of learning and knowledge to produce.

knowledge base: Model of human knowledge that is used by expert systems.

knowledge discovery: Identification of novel and valuable patterns in large databases.

knowledge engineer: A specialist who elicits information and expertise from other professionals and translates it into a set of rules, or frames, for an expert system.

knowledge management: The set of processes developed in an organization to create, gather, store, maintain, and disseminate the firm's knowledge.

knowledge management systems: Systems that support the creation, capture, storage, and dissemination of firm expertise and knowledge.

knowledge network: Online directory for locating corporate experts in well-defined knowledge domains.

knowledge repository: Collection of documented internal and external knowledge in a single location for more efficient management and utilization by the organization.

knowledge workers: People such as engineers or architects who design products or services and create knowledge for the organization.

L

learning management system (LMS): Tools for the management, delivery, tracking, and assessment of various types of employee learning.

legacy system: A system that has been in existence for a long time and that continues to be used to avoid the high cost of replacing or redesigning it.

liability: The existence of laws that permit individuals to recover the damages done to them by other actors, systems, or organizations.

Linux: Reliable and compactly designed operating system that is an offshoot of UNIX and that can run on many different hardware platforms and is available free or at very low cost. Used as alternative to UNIX and Windows NT.

LISTSERV: On-line discussion groups using e-mail broadcast from mailing list servers.

load balancing: Distribution of large numbers of requests for access among multiple servers so that no single device is overwhelmed.

local area network (LAN): A telecommunications network that requires its own dedicated channels and that encompasses a limited distance, usually one building or several buildings in close proximity.

logistics: Planning and control of all factors that will have an impact on transporting a product or service.

M

machine cycle: Series of operations required to process a single machine instruction.

machine language: A programming language consisting of the 1s and 0s of binary code.

magnetic disk: A secondary storage medium in which data are stored by means of magnetized spots on a hard or floppy disk.

magnetic tape: Inexpensive, older secondary-storage medium in which large volumes of information are stored sequentially by means of magnetized and nonmagnetized spots on tape.

mainframe: Largest category of computer, used for major business processing.

maintenance: Changes in hardware, software, documentation, or procedures to a production system to correct errors, meet new requirements, or improve processing efficiency.

managed security service provider (MSSP): Company that provides security management services for subscribing clients.

management control: Monitoring how efficiently or effectively resources are utilized and how well operational units are performing.

management information systems (MIS): The study of information systems focusing on their use in business and management..

management-level systems: Information systems that support the monitoring, controlling, decision-making, and administrative activities of middle managers.

managerial roles: Expectations of the activities that managers should perform in an organization.

man-month: The traditional unit of measurement used by systems designers to estimate the length of time to complete a project. Refers to the amount of work a person can be expected to complete in a month.

Manufacturing and production information systems: Systems that deal with the planning, development, and production of products and services and with controlling the flow of production.

market segmentation: Dividing a heterogeneous market into smaller, more homogeneous subgroups where marketing efforts can be more specifically targeted and effective.

mass customization: The capacity to offer individually tailored products or services using mass production resources..

massively parallel computers: Computers that use hundreds or thousands of processing chips to attack large computing problems simultaneously.

megahertz: A measure of cycle speed, or the pacing of events in a computer; one megahertz equals one million cycles per second.

message integrity: The ability to ascertain that a transmitted message has not been copied or altered.

metric: A standard measurement of performance.

metropolitan area network (MAN): Network that spans a metropolitan area, usually a city and its major suburbs. Its geographic scope falls between a WAN and a LAN.

microbrowser: Web browser software with a small file size that can work with low-memory constraints, tiny screens of handheld wireless devices, and low bandwidth of wireless networks.

micropayment: Payment for a very small sum of money, often less than $10.

microprocessor: Very large scale integrated circuit technology that integrates the computer's memory, logic, and control on a single chip.

microwave: A high-volume, long-distance, point-to-point transmission in which high-frequency radio signals are transmitted through the atmosphere from one terrestrial transmission station to another.

middle managers: People in the middle of the organizational hierarchy who are responsible for carrying out the plans and goals of senior management.

middleware: Software that connects two disparate applications, allowing them to communicate with each other and to exchange data.

midrange computer: Middle-size computer that is capable of supporting the computing needs of smaller organizations or of managing networks of other computers.

minicomputer: Middle-range computer used in systems for universities, factories, or research laboratories.

mirroring: Duplicating all the processes and transactions of a server on a backup server to prevent any interruption in service if the primary server fails.

MIS audit: Identifies all the controls that govern individual information systems and assesses their effectiveness.

mobile commerce (m-commerce): The use of wireless devices, such as cell phones or handheld digital information appliances, to conduct both business-to-consumer and business-to-business e-commerce transactions over the Internet.

mobile computing: Wireless computing that allows Internet-enabled cell phones, PDAs, and other wireless computing devices to access digital information from the Internet and other sources from any location.

mobile data networks: Wireless networks that enable two-way transmission of data files cheaply and efficiently.

model: An abstract representation that illustrates the components or relationships of a phenomenon.

model-driven DSS: Primarily stand-alone system that uses some type of model to perform "what-if" and other kinds of analyses.

modem: A device for translating a computer's digital signals into analog form for transmission over ordinary telephone lines, or for translating analog signals back into digital form for reception by a computer.

module: A logical unit of a program that performs one or several functions.

MP3 (MPEG3): Compression standard that can compress audio files for transfer over the Internet with virtually no loss in quality.

multicasting: Transmission of data to a selected group of recipients.

multimedia: The integration of two or more types of media such as text, graphics, sound, voice, full-motion video, or animation into a computer-based application.

multinational: Form of business organization that concentrates financial management and control out of a central home base while decentralizing

multiplexing: Ability of a single communications channel to carry data transmissions from multiple sources simultaneously.

N

natural language: Nonprocedural language that enables users to communicate with the computer using conversational commands resembling human speech.

net marketplace: A single digital marketplace based on Internet technology linking many buyers to many sellers.

Net Present Value (NPV): The amount of money an investment is worth, taking into account its cost, earnings, and the time value of money.

network: The linking of two or more computers to share data or resources, such as a printer.

network-attached storage: Attaching high-speed RAID storage devices to a network so that the devices in the network can access these storage devices through a specialized server dedicated to file service and storage.

Network Computer (NC): Simplified desktop computer that does not store software programs or data permanently. Users download whatever software or data they need from a central computer over the Internet or an organization's own internal network.

network DBMS: Older logical database model that is useful for depicting many-to-many relationships.

network economics: Model of strategic systems at the industry level based on the concept of a network where adding another participant entails zero marginal costs but can create much larger marginal gains.

Network Operating system (NOS): Special software that routes and manages communications on the network and coordinates network resources.

neural network: Hardware or software that attempts to emulate the processing patterns of the biological brain.

nomadic computing: Wireless computing where users move from wireless hot spot to wireless hot spot to gain network or Internet access.

nonobvious relationship awareness (NORA): Technology that can find obscure hidden connections between people or other entities by analyzing information from many different sources to correlate relationships.

normalization: The process of creating small stable data structures from complex groups of data when designing a relational database.

O

object-oriented DBMS: An approach to data management that stores both data and the procedures acting on the data as objects that can be automatically retrieved and shared; the objects can contain multimedia.

object-oriented development: Approach to systems development that uses the object as the basic unit of systems analysis and design. The system is modeled as a collection o objects and the relationship between them.

object-oriented programming: An approach to software development that combines data and procedures into a single object.

object-relational DBMS: A database management system that combines the capabilities of a relational DBMS for storing traditional information and the capabilities of an object-oriented DBMS for storing graphics and multimedia.

Office 2000, Office XP, and Office 2003: Integrated desktop productivity software suites with capabilities for supporting collaborative work on the Web or incorporating information from the Web into documents.

office systems: Systems such as word processing, desktop publishing, e-mail, electronic scheduling, and videoconferencing, designed to increase worker productivity in the office.

on-line analytical processing (OLAP): Capability for manipulating and analyzing large volumes of data from multiple perspectives.

on-line processing: A method of collecting and processing data in which transactions are entered directly into the computer system and processed immediately.

on-line transaction processing: Transaction processing mode in which transactions entered on-line are immediately processed by the computer.

Open Systems Interconnect (OSI): Less widely used network connectivity model developed by International Standards Organization for linking different types of computers and networks.

open-source software: Software that provides free access to its program code, allowing users to modify the program code to make improvements or fix errors.

operating system: The system software that manages and controls the activities of the computer.

operational control: Deciding how to carry out specific tasks specified by upper and middle management and establishing criteria for completion and resource allocation.

operational CRM: Customer-facing applications, such as sales force automation, call center and customer service support, and marketing automation.

operational managers: People who monitor the day-to-day activities of the organization.

operational-level systems: Information systems that monitor the elementary activities and transactions of the organization.

opt-in: Model of informed consent permitting prohibiting an organization from collecting any personal information unless the individual specifically takes action to approve information collection and use.

opt-out: Model of informed consent permitting the collection of personal information until the consumer specifically requests that the data not be collected.

optical network: High-speed networking technologies for transmitting data in the form of light pulses.

organization (behavioral definition): A collection of rights, privileges, obligations, and responsibilities that are delicately balanced over a period of time through conflict and conflict resolution.

organization (technical definition): A stable, formal, social structure that takes resources from the environment and processes them to produce outputs.

organizational and management capital: Investments in organization and management such as new business processes, management behavior, organizational culture, or training.

organizational culture: The set of fundamental assumptions about what products the organization should produce, how and where it should produce them, and for whom they should be produced.

organizational impact analysis: Study of the way a proposed system will affect organizational structure, attitudes, decision making, and operations.

organizational learning: Creation of new standard operating procedures and business processes that reflect organizations' experience.

organizational memory: The stored learning from an organization's history that can be used for decision making and other purposes.

organizational models of decision making: Models of decision making that take into account the structural and political characteristics of an organization.

output controls: Measures that ensure that the results of computer processing are accurate, complete, and properly distributed.

output: The distribution of processed information to the people who will use it or to the activities for which it will be used.

outsourcing: The practice of contracting computer center operations, telecommunications networks, or applications development to external vendors.

P

P3P: Industry standard designed to give users more control over personal information gathered on Web sites they visit. Stands for Platform for Privacy Preferences Project.

packet switching: Technology that breaks messages into small, fixed bundles of data and routes them in the most economical way through any available communications channel..

paging system: A wireless transmission technology in which the pager beeps when the user receives a message; used to transmit short alphanumeric messages.

paradigm shift: Radical reconceptualization of the nature of the business and the nature of the organization.

parallel processing: Type of processing in which more than one instruction can be processed at a time by breaking down a problem into smaller parts and processing them simultaneously with multiple processors.

parallel strategy: A safe and conservative conversion approach where both the old system and its potential replacement are run together for a time until everyone is assured that the new one functions correctly.

partner relationship management (PRM): Automation of the firm's relationships with its selling partners using customer data and analytical tools to improve coordination and customer sales.

patent: A legal document that grants the owner an exclusive monopoly on the ideas behind an invention for 17 years; designed to ensure that inventors of new machines or methods are rewarded for their labor while making widespread use of their inventions.

payback method: A measure of the time required to pay back the initial investment on a project.

peer-to-peer computing: Form of distributed processing that links computers via the Internet or private networks so that they can share processing tasks.

peer-to-peer payment system: Electronic payment system for people who want to send money to vendors or individuals who are not set up to accept credit card payments.

peer-to-peer: Network architecture that gives equal power to all computers on the network; used primarily in small networks.

Personal Communication Services (PCS): A wireless cellular technology that uses lower power, higher frequency radio waves than does cellular technology and so can be used with smaller size telephones.

Personal Computer (PC): Small desktop or portable computer.

Personal Digital Assistants (PDA): Small, pen-based, handheld computers with built-in wireless telecommunications capable of entirely digital communications transmission.

phased approach: Introduces the new system in stages either by functions or by organizational units.

pilot study: A strategy to introduce the new system to a limited area of the organization until it is proven to be fully functional; only then can the conversion to the new system across the entire organization take place.

political models of decision making: Models of decision making where decisions result from competition and bargaining among the organization's interest groups and key leaders.

pop-up ad: Ad that opens automatically and does not disappear until the user clicks on it.

portal: Web site or other service that provides an initial point of entry to the Web or to internal company data.

portfolio analysis: An analysis of the portfolio of potential applications within a firm to determine the risks and benefits, and to select among alternatives for information systems.

post-implementation audit: Formal review process conducted after a system has been placed in production to determine how well the system has met its original objectives.

predictive analysis: Use of datamining techniques, historical data, and assumptions about future conditions to predict outcomes of events.

present value: The value, in current dollars, of a payment or stream of payments to be received in the future.

presentation graphics: Software to create professional-quality graphics presentations that can incorporate charts, sound, animation, photos, and video clips.

primary activities: Activities most directly related to the production and distribution of a firm's products or services.

primary storage: Part of the computer that temporarily stores program instructions and data being used by the instructions.

privacy: The claim of individuals to be left alone, free from surveillance or interference from other individuals, organizations, or the state.

private exchange: Another term for a private industrial network.

private industrial networks: Web-enabled networks linking systems of multiple firms in an industry for the coordination of trans-organizational business processes.

process specifications: Describe the logic of the processes occurring within the lowest levels of a data flow diagram.

processing controls: The routines for establishing that data are complete and accurate during updating.

processing: The conversion, manipulation, and analysis of raw input into a form that is more meaningful to humans.

procurement: Sourcing goods and materials, negotiating with suppliers, paying for goods, and making delivery arrangements.

product differentiation: Competitive strategy for creating brand loyalty by developing new and unique products and services that are not easily duplicated by competitors.

production or service workers: People who actually produce the products or services of the organization.

production: The stage after the new system is installed and the conversion is complete; during this time the system is reviewed by users and technical specialists to determine how well it has met its original goals.

profiling: The use of computers to combine data from multiple sources and create electronic dossiers of detailed information on individuals.

profitability index: Used to compare the profitability of alternative investments; it is calculated by dividing the present value of the total cash inflow from an investment by the initial cost of the investment.

program: A series of statements or instructions to the computer.

program-data dependence: The close relationship between data stored in files and the software programs that update and maintain those files. Any change in data organization or format requires a change in all the programs associated with those files.

programmers: Highly trained technical specialists who write computer software instructions.

programming: The process of translating the system specifications prepared during the design stage into program code.

protocol: A set of rules and procedures that govern transmission between the components in a network.

prototype: The preliminary working version of an information system for demonstration and evaluation purposes.

prototyping: The process of building an experimental system quickly and inexpensively for demonstration and evaluation so that users can better determine information requirements.

public key infrastructure: System for creating public and private keys using a certificate authority (CA) and digital certificates for authentication.

pull-based model: Supply chain driven by actual customer orders or purchases so that members of the supply chain produce and deliver only what customers have ordered.

pure-play: Business models based purely on the Internet.

push-based model: Supply chain driven by production master schedules based on forecasts or best guesses of demand for products, and products are “pushed” to customers.

"push" technology: Method of obtaining relevant information on networks by having a computer broadcast information directly to the user based on prespecified interests.

Q/R

query language: Software tool that provides immediate online answers to requests for information that are not predefined.

radio-frequency identification (RFID): Technology using tiny tags with embedded microchips containing data about an item and its location to transmit short-distance radio signals to special RFID readers that then pass the data on to a computer for processing.

RAID (Redundant Array of Inexpensive Disks): Disk storage technology to boost disk performance by packaging more than 100 smaller disk drives with a controller chip and specialized software in a single large unit to deliver data over multiple paths simultaneously.

RAM (Random Access Memory): Primary storage of data or program instructions that can directly access any randomly chosen location in the same amount of time.

Rapid Application Development (RAD): Process for developing systems in a very short time period by using prototyping, fourth-generation tools, and close teamwork among users and systems specialists.

rational model: Model of human behavior based on the belief that people, organizations, and nations engage in basically consistent, value-maximizing calculations.

rationalization of procedures: The streamlining of standard operating procedures, eliminating obvious bottlenecks, so that automation makes operating procedures more efficient.

reach: Measurement of how many people a business can connect with and how many products it can offer those people.

real options pricing models: Models for evaluating information technology investments with uncertain returns by using techniques for valuing financial options.

record: A group of related fields.

recovery-oriented computing: Computer systems designed to recover rapidly when mishaps occur.

Reduced Instruction Set Computing (RISC): Technology used to enhance the speed of microprocessors by embedding only the most frequently used instructions on a chip.

reintermediation: The shifting of the intermediary role in a value chain to a new source.

relational DBMS: A type of logical database model that treats data as if they were stored in two-dimensional tables. It can relate data stored in one table to data in another as long as the two tables share a common data element.

Repetitive Stress Injury (RSI): Occupational disease that occurs when muscle groups are forced through repetitive actions with high-impact loads or thousands of repetitions with low-impact loads.

Request for Proposal (RFP): A detailed list of questions submitted to vendors of software or other services to determine how well the vendor's product can meet the organization's specific requirements.

resource allocation: The determination of how costs, time, and personnel are assigned to different phases of a systems development project.

responsibility: Accepting the potential costs, duties, and obligations for the decisions one makes.

reverse logistics: The return of items from buyers to sellers in a supply chain.

richness: Measurement of the depth and detail of information that a business can supply to the customer as well as information the business collects about the customer.

ring network: A network topology in which all computers are linked by a closed loop in a manner that passes data in one direction from one computer to another.

risk assessment: Determining the potential frequency of the occurrence of a problem and the potential damage if the problem were to occur. Used to determine the cost/benefit of a control.

Risk Aversion Principle: Principle that one should take the action that produces the least harm or incurs the least cost.

ROM (Read-Only Memory): Semiconductor memory chips that contain program instructions. These chips can only be read from; they cannot be written to.

router: Specialized communications processor that forwards packets of data from one network to another network.

rule base: The collection of knowledge in an AI system that is represented in the form of IF-THEN rules.

S

safe harbor: Private self-regulating policy and enforcement mechanism that meets the objectives of government regulations but does not involve government regulation or enforcement.

Sales and marketing information systems: Systems that help the firm identify customers for the firm's products or services, develop products and

services to meet their needs, promote these products and services, sell the products and services, and provide ongoing customer support.

satellite: The transmission of data using orbiting satellites that serve as relay stations for transmitting microwave signals over very long distances.

scalability: The ability of a computer, product, or system to expand to serve a larger number of users without breaking down.

scoring model: A quick method for deciding among alternative systems based on a system of ratings for selected objectives.

search-based advertising: Payment to a search service to display a sponsored link to a company's Web site as a way of advertising that company.

search costs: The time and money spent locating a suitable product and determining the best price for that product.

search engine: A tool for locating specific sites or information on the Internet.

secondary storage: Relatively long term, nonvolatile storage of data outside the CPU and primary storage.

security: Policies, procedures, and technical measures used to prevent unauthorized access, alteration, theft, or physical damage to information systems.

selection construct: The logic pattern in programming where a stated condition determines which of two alternative actions can be taken.

Semantic web: Collaborative effort led by the World Wide Web Consortium to make Web searching more efficient by reducing the amount of human involvement in searching for and processing web information.

semistructured knowledge: Information in the form of less structured objects, such as e-mail, chat room exchanges, videos, graphics, brochures, or bulletin boards.

semistructured knowledge system: System for organizing and storing less structured information, such as e-mail, voice mail, videos, graphics, brochures, or bulleting boards. Also known as digital asset management system.

senior managers: People occupying the topmost hierarchy in an organization who are responsible for making long-range decisions.

sensitivity analysis: Models that ask "what-if" questions repeatedly to determine the impact of changes in one or more factors on the outcomes.

sequence construct: The sequential single steps or actions in the logic of a program that do not depend on the existence of any condition.

server: Computer specifically optimized to provide software and other resources to other computers over a network.

server farm: Large group of servers maintained by a commercial vendor and made available to subscribers for electronic commerce and other activities requiring heavy use of servers.

service platform: Integration of multiple applications from multiple business functions, business units, or business partners to deliver a seamless experience for the customer, employee, manager, or business partner.

shopping bot: Software with varying levels of built-in intelligence to help electronic commerce shoppers locate and evaluate products or service they might wish to purchase.

six sigma: A specific measure of quality, representing 3.4 defects per million opportunities; used to designate a set of methodologies and techniques for improving quality and reducing costs.

smart card: A credit-card-size plastic card that stores digital information and that can be used for electronic payments in place of cash.

smart phone: Wireless phone with voice, text, and Internet capabilities.

SOAP (Simple Object Access Protocol): Set of rules that allows Web services applications to pass data and instructions to one another.

social engineering: Tricking people into revealing their passwords by pretending to be legitimate users or members of a company in need of information.

sociotechnical design: Design to produce information systems that blend technical efficiency with sensitivity to organizational and human needs.

software metrics: The objective assessments of the software used in a system in the form of quantified measurements.

software package: A prewritten, precoded, commercially available set of programs that eliminates the need to write software programs for certain functions.

source code: Program instructions written in a high-level language that must be translated into machine language to be executed by the computer.

spam: Unsolicited commercial e-mail.

spreadsheet: Software displaying data in a grid of columns and rows, with the capability of easily recalculating numerical data.

spyware: Technology that aids in gathering information about a person or organization without their knowledge.

Standard Operating Procedures (SOPs): Formal rules for accomplishing tasks that have been developed to cope with expected situations.

star network: A network topology in which all computers and other devices are connected to a central host computer. All communications between network devices must pass through the host computer.

Storage Area Network (SAN): A high-speed network dedicated to storage that connects different kinds of storage devices, such as tape libraries and disk arrays so they can be shared by multiple servers.

Storage Service Provider (SSP): Third-party provider that rents out storage space to subscribers over the Web, allowing customers to store and access their data without having to purchase and maintain their own storage technology.

storage technology: Physical media and software governing the storage and organization of data for use in an information system.

stored value payment systems: Systems enabling consumers to make instant on-line payments to merchants and other individuals based on value stored in a digital account.

strategic decision making: Determining the long-term objectives, resources, and policies of an organization.

strategic information systems: Computer systems at any level of the organization that change goals, operations, products, services, or environmental relationships to help the organization gain a competitive advantage.

strategic transitions: A movement from one level of sociotechnical system to another. Often required when adopting strategic systems that demand changes in the social and technical elements of an organization.

strategic-level systems: Information systems that support the long-range planning activities of senior management.

streaming technology: Technology for transferring data so that they can be processed as a steady and continuous stream.

structure chart: System documentation showing each level of design, the relationship among the levels, and the overall place in the design structure; can document one program, one system, or part of one program.

structured: Refers to the fact that techniques are carefully drawn up, step by step, with each step building on a previous one.

structured analysis: A method for defining system inputs, processes, and outputs and for partitioning systems into subsystems or modules that show a logical graphic model of information flow.

structured decisions: Decisions that are repetitive, routine, and have a definite procedure for handling them.

structured design: Software design discipline encompassing a set of design rules and techniques for designing systems from the top down in hierarchical fashion.

structured knowledge: Knowledge in the form of structured documents and reports.

structured knowledge system: System for organizing structured knowledge in a repository where it can be accessed throughout the organization. Also known as content management system.

structured programming: Discipline for organizing and coding programs that simplifies the control paths so that the programs can be easily understood and

modified; uses the basic control structures and modules that have only one entry point and one exit point.

Structured Query Language (SQL): The standard data manipulation language for relational database management systems.

subschema: The specific set of data from the database that is required by each user or application program.

supercomputer: Highly sophisticated and powerful computer that can perform very complex computations extremely rapidly.

supply chain: Network of organizations and business processes for procuring materials, transforming raw materials into intermediate and finished products, and distributing the finished products to customers.

supply chain execution systems: Systems to manage the flow of products through distribution centers and warehouses to ensure that products are delivered to the right locations in the most efficient manner.

supply chain management: Integration of supplier, distributor, and customer logistics requirements into one cohesive process.

supply chain management systems: Information systems that automate the flow of information between a firm and its suppliers in order to optimize the planning, sourcing, manufacturing, and delivery of products and services.

supply chain planning systems: Systems that enable a firm to generate demand forecasts for a product and to develop sourcing and manufacturing plans for that product.

support activities: Activities that make the delivery of a firm's primary activities possible. Consist of the organization's infrastructure, human resources, technology, and procurement.

switched lines: Telephone lines that a person can access from a terminal to transmit data to another computer, the call being routed or switched through paths to the designated destination.

switching costs: The expense a customer or company incurs in lost time and expenditure of resources when changing from one supplier or system to a competing supplier or system.

syndicator: Business aggregating content or applications from multiple sources, packaging them for distribution, and reselling them to third-party Web sites.

system failure: An information system that either does not perform as expected, is not operational at a specified time, or cannot be used in the way it was intended.

system software: Generalized programs that manage the computer's resources, such as the central processor, communications links, and peripheral devices.

system testing: Tests the functioning of the information system as a whole in order to determine if discrete modules will function together as planned.

systematic decision makers: Cognitive style that describes people who approach a problem by structuring it in terms of some formal method.

systems analysis: The analysis of a problem that the organization will try to solve with an information system.

systems analysts: Specialists who translate business problems and requirements into information requirements and systems, acting as liaison between the information systems department and the rest of the organization.

systems design: Details how a system will meet the information requirements as determined by the systems analysis.

systems development: The activities that go into producing an information systems solution to an organizational problem or opportunity.

systems lifecycle: A traditional methodology for developing an information system that partitions the systems development process into formal stages that must be completed sequentially with a very formal division of labor between end users and information systems specialists.

T

T1 line: A dedicated telephone connection comprising 24 channels that can support a data transmission rate of 1.544 megabits per second. Each channel can be configured to carry voice or data traffic.

tacit knowledge: Expertise and experience of organizational members that has not been formally documented.

tangible benefits: Benefits that can be quantified and assigned a monetary value; they include lower operational costs and increased cash flows.

taxonomy: Method of classifying things according to a predetermined system.

teamware: Group collaboration software that is customized for teamwork.

technostress: Stress induced by computer use; symptoms include aggravation, hostility toward humans, impatience, and enervation.

telecommunications system: A collection of compatible hardware and software arranged to communicate information from one location to another.

teleconferencing: The ability to confer with a group of people simultaneously using the telephone or electronic-mail group communication software.

Telnet: Network tool that allows someone to log on to one computer system while doing work on another.

test plan: Prepared by the development team in conjunction with the users; it includes all of the preparations for the series of tests to be performed on the system.

testing: The exhaustive and thorough process that determines whether the system produces the desired results under known conditions.

touch point: Method of firm interaction with a customer, such as telephone, e-mail, customer service desk, conventional mail, or point-of-purchase.

topology: The way in which the components of a network are connected.

Total Cost of Ownership (TCO): Designates the total cost of owning technology resources, including initial purchase costs, the cost of hardware and software upgrades, maintenance, technical support, and training.

Total Quality Management (TQM): A concept that makes quality control a responsibility to be shared by all people in an organization.

trade secret: Any intellectual work or product used for a business purpose that can be classified as belonging to that business, provided it is not based on information in the public domain.

transaction cost theory: Economic theory stating that firms grow larger because they can conduct marketplace transactions internally more cheaply than they can with external firms in the marketplace.

Transaction Processing Systems (TPS): Computerized systems that perform and record the daily routine transactions necessary to conduct the business; they serve the organization's operational level.

transborder data flow: The movement of information across international boundaries in any form.

Transmission Control Protocol/Internet Protocol (TCP/IP): Dominant model for achieving connectivity among different networks. Provides a universally agree-on method for breaking up digital messages into packets, routing them to the proper addresses, and then reassembling them into coherent messages.

transnational: Truly global form of business organization with no national headquarters; value-added activities are managed from a global perspective without reference to national borders, optimizing sources of supply and demand and local competitive advantage.

Trojan horse: A software program that appears legitimate but contains a second hidden function that may cause damage.

tuple: A row or record in a relational database.

twisted wire: A transmission medium consisting of pairs of twisted copper wires; used to transmit analog phone conversations but can be used for data transmission.

U

UDDI (Universal Description, Discovery, and Integration): Allows a Web service to be listed in a directory of Web services so that it can be easily located by other organizations and systems.

unified messaging: System combining voice messages, e-mail, and fax so that they can all be obtained from a single system.

Unified Modeling Language (UML): Industry standard methodology for analysis and design of an object-oriented software system.

Uniform Resource Locator (URL): The address of a specific resource on the Internet.

unit testing: The process of testing each program separately in the system. Sometimes called program testing.

UNIX: Operating system for all types of computers, which is machine independent and supports multiuser processing, multitasking, and networking. Used in high-end workstations and servers.

unstructured decisions: Nonroutine decisions in which the decision maker must provide judgment, evaluation, and insights into the problem definition; there is no agreed-upon procedure for making such decisions.

up-selling: Marketing higher-value products or services to new or existing customers.

Usenet: Forums in which people share information and ideas on a defined topic through large electronic bulletin boards where anyone can post messages on the topic for others to see and to which others can respond.

user interface: The part of the information system through which the end user interacts with the system; type of hardware and the series of on-screen commands and responses required for a user to work with the system.

user-designer communications gap: The difference in backgrounds, interests, and priorities that impede communication and problem solving among end users and information systems specialists.

Utilitarian Principle: Principle that assumes one can put values in rank order and understand the consequences of various courses of action.

utility computing: Model of computing in which companies pay only for the information technology resources they actually use during a specified time period. Also called on-demand computing or usage-based pricing.

V

value chain model: Model that highlights the primary or support activities that add a margin of value to a firm's products or services where information systems can best be applied to achieve a competitive advantage.

value web: Customer-driven network of independent firms who use information technology to coordinate their value chains to collectively produce a product or service for a market.

Value-Added Network (VAN): Private, multipath, data-only, third-party-managed network that multiple organizations use on a subscription basis.

videoconferencing: Teleconferencing in which participants see each other over video screens.

virtual organization: Organization using networks to link people, assets and ideas to create and distribute products and services without being limited to traditional organizational boundaries or physical location.

Virtual Private Network (VPN): A secure connection between two points across the Internet to transmit corporate data. Provides a low-cost alternative to a private network.

Virtual Reality Modeling Language (VRML): A set of specifications for interactive three-dimensional modeling on the World Wide Web.

virtual reality systems: Interactive graphics software and hardware that create computer-generated simulations that provide sensations that emulate real-world activities.

Visual Basic: Widely used visual programming tool and environment for creating applications that run on Microsoft Windows.

visual programming: The construction of software programs by selecting and arranging programming objects rather than by writing program code.

voice mail: A system for digitizing a spoken message and transmitting it over a network.

Voice over IP (VoIP): Facilities for managing the delivery of voice information using the Internet Protocol (IP).

W

walkthrough: A review of a specification or design document by a small group of people carefully selected based on the skills needed for the particular objectives being tested.

Web browser: An easy-to-use software tool for accessing the World Wide Web and the Internet.

Web bugs: Tiny graphic files embedded in e-mail messages and Web pages that are designed to monitor online Internet user behavior.

Web content management tools: Software to facilitate the collection, assembly, and management of content on a Web site, intranet, or extranet.

Web hosting service: Company with large Web server computers to maintain the Web sites of fee-paying subscribers.

Web personalization: The tailoring of Web content directly to a specific user.

Web server: Software that manages requests for Web pages on the computer where they are stored and that delivers the page to the user's computer.

Web services: Set of universal standards using Internet technology for integrating different applications from different sources without time-consuming custom coding. Used for linking systems of different organizations or for linking disparate systems within the same organization.

Web site performance monitoring tools: Software tools for monitoring the time to download Web pages, perform Web transactions, identify broken links between Web pages, and pinpoint other Web site problems and bottlenecks.

Web site: All of the World Wide Web pages maintained by an organization or an individual.

Webmaster: The person in charge of an organization's Web site.

Wi-Fi: Standards for Wireless Fidelity and refers to the 802.11 family of wireless networking standards.

Wide Area Network (WAN): Telecommunications network that spans a large geographical distance. May consist of a variety of cable, satellite, and microwave technologies.

Windows 2000: Windows operating system for high-performance PCs and network servers. Supports networking, multitasking, multiprocessing, and Internet services.

Windows 2003: Most recent Windows operating system for servers.

Windows 98: Earlier version of the Windows operating system that is closely integrated with the Internet.

Windows XP: Powerful Windows operating system that provides reliability, robustness, and ease of use for both corporate and home PC users.

Wireless Application Protocol (WAP): System of protocols and technologies that lets cell phones and other wireless devices with tiny displays, low-bandwidth connections, and minimal memory access Web-based information and services.

wireless NIC: add-in-card (network interface card) that has a built-in-radio and antenna.

wisdom: The collective and individual experience of applying knowledge to the solution of problems.

WML (Wireless Markup Language): Markup language for Wireless Web sites; based on XML and optimized for tiny displays.

word processing software: Software for electronically creating, editing, formatting, and printing documents.

work-flow management: The process of streamlining business procedures so that documents can be moved easily and efficiently from one location to another.

workstation: Desktop computer with powerful graphics and mathematical capabilities and the ability to perform several complicated tasks at once.

World Wide Web: A system with universally accepted standards for storing, retrieving, formatting, and displaying information in a networked environment.

worms: Independent software programs that propagate themselves to disrupt the operation of computer networks or destroy data and other programs.

WSDL(Web Services Description Language): Common framework for describing the tasks performed by a Web service so that it can be used by other applications.

X

XHTML (Extensible Hypertext Markup Language): Hybrid of HTML and XML that provides more flexibility than HTML.

XML (eXtensible Markup Language): General-purpose language that describes the structure of a document and supports links to multiple documents, allowing data to be manipulated by the computer. Used for both Web and non-Web applications.

New Tech Terms You Should Know; To Win Your Job's Interview?

When you're new to tech, it can feel like you've stumbled into a conversation where everyone knows what they're talking about—except, of course, you. Even after you've broken in, the moment you meet with a different team or tackle a new type of project, the jargon and acronyms start shifting all over again. Well, enough with the frantic Googling every time someone says "CTR" "CTO" and "CMS."

Sometimes a short and simple explanation is all you need to get your head around a new concept. And that's exactly what we have here: a jargon-free, plain-English tech dictionary, if you will, of 99 fundamental terms you'll need if you're considering a career shift into tech. Trust us, we know—these are the terms the Skillcrush team uses all the time in our online tech courses (and our free 10-day coding bootcamp, too!).

You're sure to find more than a few terms you've already wondered about during nerve-wracking meetings or interviews—so go ahead and scratch those off your "what the heck is that?" list right away—then bookmark the rest, and come back whenever another pesky tech term that has baffles you.

Tech Terms 101: Social Media

1. CONTENT CURATION

Even if you're new to tech, chances are you're not new to social media. Content curation is the process social media sites use to gather and present content (articles, links,

videos, images, etc.) that are relevant to a specific topic or a user's area of interest. Content can be selected manually by a person designated as a curator, or it can be gathered through automated programs that track things like upvoting, likes, hashtags, or analysis of a user's previous online activity.

2. ENGAGEMENT

Engagement is the term used for likes, shares, comments, and other interactions with a business' social media presence. It's one metric companies use to measure and evaluate their social media performance. Liking posts on Facebook, tweeting on Twitter, and viewing pins on Pinterest are all examples of social media engagement.

3. EMBEDDING

Embedding is the process of putting social media content on a web page. When you see a YouTube video on a blog you're reading, or a tweet posted on a business' website you're checking out, that's an example of embedding. Embedding is done through HTML code, and most social media sites have an "Embed" option that gives you the exact code you'll need.

4. IMPRESSIONS

Impressions are the number of times an online ad, article, video, or other piece of content has been "fetched" (or requested) from it's source. Whenever social media content shows up in your web browser, it's been fetched

from its web server, and you've just helped make an impression.

5. INFLUENCER

Influencers are individuals with a presence on social media who use their platform to affect group opinions and behavior. Celebrities who use Twitter, YouTubers with legions of subscribers, and activists organizing on Facebook are all examples of social media influencers.

6. MENTIONS

Mentions are ways of referring to a person or account on social media platforms (@yourTwitterhandle on Twitter or +YourName on Google+). Mentions are another metric companies can use to measure their social media footprint.

7. MICROBLOGGING

Microblogs are a subset of traditional blogs where instead of longform content, short messages consisting of a few sentences, an image, a video, or a link are posted and shared. Twitter, Google+, Tumblr, and Facebook are all considered microblogging platforms.

8. ORGANIC

Organic content describes social media content that ranks highly because individual users have liked, reposted, or viewed it—not because companies have paid to promote the content. Articles, memes, and videos that go viral through online community sharing (as opposed to paid advertising) are considered organic.

9. REACH

Reach is a measure of your potential social media audience size—it's the number of people who have access to your content, though reach doesn't measure whether or not the content was actually viewed.

10. SOCIAL GRAPH

Social graph is a model of Internet users on a social network and how each one is related to all the others. It's the way sites like Facebook suggest other users you might know, or how online games like Words With Friends connect you with your cousin who you didn't even know had a social media account.

11. USER-GENERATED CONTENT (UGC)

User generated is social media content (posts, photos, videos) created by users about a brand or product and not by companies or organizations themselves. When a local restaurant retweets a Twitter user's positive shout out, or Coca-Cola reposts a picture of an Instagram user drinking a Coke, that's UGC.

Tech Terms 101: Digital Marketing

12. AFFILIATE MARKETING

Affiliate marketing is a strategy where businesses reward individual affiliates (people or organizations outside the business) for bringing in new customers or visitors through ads or content on the affiliate's website. Affiliates receive payments or product discounts based on the number of customers they generate. These are

exceptionally common on personal or lifestyle sites, where you might notice that the product recommended has a link with a bunch of additional stuff tracked on. That means that if you buy the product, the blog or site that lead you there will make some percentage of that sale.

13. BOUNCE RATE

Bounce rate is the percentage of visitors to a website who leave the website quickly without really looking at it—Google analytics calculates bounce rates based on website sessions where a visitor only sticks round to look at one page before bouncing. Sites aim to keep this number low—they want you spend time on their site—and so they'll try to keep content as engaging and relevant to you as possible.

14. CALL TO ACTION (CTA)

A call to action is the text, banner, form, or image on a web page (or email) asking a visitor to literally take an action—read more content, join an email list, sign up for a webinar, buy a product, etc. CTAs are a marketing tool that converts web users into leads for businesses.

15. CLICK THROUGH RATE (CTR)

The click through rate is the percentage of users who click on links in web pages or marketing emails. CTR is significant because it measures how many users are actively engaging with linked content on a site.

16. SALES FUNNEL

A sales funnel is an internet marketing method where potential customers are led through a series of events or actions that can be mapped out in the shape of funnel. The broadest level at the top of the funnel involves drawing users to your website, after which they move down the funnel where they're offered services or resources if they sign up for your email list, until eventually they move to the bottom of the funnel and become a paying customer.

17. CONVERSION RATE OPTIMIZATION (CRO)

CRO is a marketing system for raising the percentage of website visitors who convert to paying customers. CRO methods usually involve encouraging users to take specific actions on the website, such as filling out a web form, signing up for a trial, or joining an email list.

18. KEYWORDS

Keywords are words or phrases commonly used in search engines to look for online content. Keyword research is a powerful internet marketing tool, since web advertising and website search engine placement can be optimized to match high traffic keywords, making it more likely that Googlers will end up on your site.

19. MARKETING AUTOMATION

Marketing automation describes the use of software or online services (like HubSpot, MailChimp, and Act-on) to automate repetitive marketing tasks like emails, customer relationship management, social media posts, and

analytics. Marketing Automation programs allow marketers to input specific criteria for the tasks in question and that data is interpreted and executed by the program.

20. MULTICHANNEL MARKETING

Multichannel marketing uses a variety of of communication platforms (website banner ads, Facebook ads, marketing emails, a blog) to interact with potential customers. This approach allows users to choose which channel they want to use to interact with your product and increases options for converting impressions into customers.

21. SEARCH ENGINE MARKETING (SEM)

SEM describes the use of paid advertising (such as brief copy, product listings, and video clips) on search engines (Google, Yahoo, Bing) in order to drive user traffic to your website.

22. SEARCH ENGINE OPTIMIZATION (SEO)

SEO is the internet marketing practice of optimizing a website so that it's more likely to show up in unpaid search results (e.g. it's one of the first sites to appear when you Google a related topic). SEO is done through using algorithms based on specific search engine's behaviors, analyzing the specific keywords typed into search engines, and researching which search engines are popular with particular demographics. Then, you align the format of your content (including things like making sure

headings have extra searchable terms) to give yourself the best shot at moving up in search engines' rankings. Be careful, though: going too far has consequences. "Stuffing" your content with keywords will get you dinged on Google's rankings, so don't let the goal of SEO overtake producing authentic, excellent content.

23. TRAFFIC

Traffic is the total amount of users who visit a website. Overall traffic is then broken down into specific types of visits—like unique visitors and total clicks.

Tech Terms 101: Growth Hacking

24. GROWTH HACKING

Growth hacking is the creative use of technology (websites, marketing emails, apps) and analytics (data mining, A/B testing) in combination with product development to increase a company's growth. Growth hacking focuses on lower cost marketing alternatives to traditional television, newspaper, and radio ads and is often used by startup companies looking to grow their business rapidly during their launch phase.

25. A/B TESTING

A/B testing is the practice of comparing two versions of online content—websites, apps, marketing emails, etc.—in order to see which version performs better. The two versions (A and B) are presented to users at random in order to gauge reactions.

26. CONTENT MARKETING

Content marketing uses online content—e-books, videos, blog posts, podcasts etc—in order to attract an audience who might become customers. Content marketing doesn't explicitly advertise a brand, but is designed to indirectly stimulate interest in products and services.

27. CONVERSION

Conversion is the successful end result of marketing attempts like A/B Testing and Content Marketing—it's the point when a potential customer becomes a lead or an actual customer.

28. DATA MINING

Data mining is the practice of examining large amounts of data in user databases and websites to find consumer patterns, behaviors, and relationships that can be useful in marketing goods and services online.

29. EMAIL MARKETING

Email marketing is the use of direct emails (directly from a business to an individual) in order to communicate with current and potential customers. Email marketing allows companies to strike a more personal tone and establish a voice for their brand.

30. WEBSITE OPTIMIZATION

Site optimization is the process of using growth hacks to improve the site's ability to convert visitors to customers. Controlled experiments like A/B testing are carried out and the results are used to make website changes that lead

to more pageviews, product purchases, etc. This might include making it easier to get from one article to another, making sure pop ups most accurately match the interest of the user, or simply making the site as a whole easier to navigate.

Tech Terms 101: Graphic Design

31. COLOR VALUE

Color value is a color's shade. Value can be defined in HTML by name (black), hexadecimal color code (#000000), or RGB code (0, 0, 0).

32. COLOR THEORY

Color theory is the study of color mixing and the visual effects of specific color combinations (complimentary colors, color harmony, how to achieve specific tints and shades, etc).

33. GRID SYSTEM

Grids are a set of columns and rows that can be used as guidelines to arrange content on a web page. Grid systems help provide a solid base of uniformity and consistency in your design, making it more legible for the viewer.

34. PIXELS PER INCH (PPI)

PPI—also known as "density"—describes the amount of pixels (small dots of light that combine to create an image) per inch on a device display. The higher the PPI count, the sharper images and text will appear.

35. RASTER IMAGES

Raster images are computer graphics made of pixels that can be edited pixel-by-pixel with programs like Painter and Photoshop. Raster image are stored in familiar image file types like GIFs, JPEGs, and PNGs.

36. RESOLUTION

Resolution is the number of pixels that can be shown on a device's display. As with PPI, the higher the resolution on a display the more pixels can be shown, making images and texts sharper.

37. RETINA DISPLAY

Retina display is a term trademarked by Apple for an electronic device display with density so high (usually over 300 ppi) that people can't see the individual pixels. Apple uses these high density displays in Apple Watches, iPhones, iPads, MacBooks, and iMacs.

38. VECTOR IMAGES

Vector images are made up of lines that are calculated mathematically. Because of the images' line composition (instead of being composed by individual pixels), they can't be styled to the same degree as raster images, but they don't lose lose quality when they're magnified (pixel-based raster images start to look like a collection of squares when zoomed-in too closely), whereas vector images can be infinitely scaled up. You can use programs like Adobe Photoshop to work with vectors.

Tech Terms 101: User Experience

39. MOOD BOARD

Mood boards are collections of content (images, materials, pieces of text) used to represent the visual style of a website—or any creative project— (color palette, images, icons, and fonts) in pre-production. The style represented on a mood board is then translated into digital form by visual designers.

40. SITEMAPS

Sitemaps are outlines or maps of the pages that make up a website. They show the relationship between the pages and links, apps, videos, or other components, and can take the form of a document or its own page on the website.

41. USER INTERFACE (UI)

User interface includes all the parts of a website, app, computer, smartphone, etc. that the user can manipulate and interact with. Display and touch screens, website menus, keyboards, your cursor—these are all part of a user interface.

42. USER EXPERIENCE (UX)

UX describes the emotions, attitudes, and ease-of-use a person has when using a product or service. UX Design is the practice of using design to improve communication between a product and its user in order to enhance the user's overall experience.

43. USER FLOW

User flow is the path typical users take when starting on a website and moving toward an action on the site. Creating a smooth path that is intuitive for users to follow is part of user experience (UX) design.

44. USER PERSONA

Personas are theoretical user profiles created by UX and web designers to define what kind of user a website is being built for and what their needs are. User personas are created using demographic information, user research, and analysis of customer experience metrics.

45. USER RESEARCH

User research is the process of investigating how users behave on a website and examining how that behavior can lead to a better website experience and design.

46. WIREFRAME

Wireframes are sketches of the key information that goes on each page of a website, essentially showing the site or page's "skeleton." Designers can then use this sketch as a starting point for laying out a website. These can be made in programs or written on a napkin—we've seen it all.

Tech Terms 101: Typography

47. FONT

Fonts or typefaces are what determine the look of your text—they are typically designed by people who specialize in type design. Fonts are organized by style

family's (Arial) and then grouped by weight with in that family (regular, italics, or bold).

48. KERNING

Kerning refers to the space between characters (letters, numbers, punctuation) and the process of adjusting that space to avoid unsightly gaps and improve the legibility of text.

49. LEADING (PRONOUNCED "LED-ING")

Leading is the vertical distance between lines of text on a website—in other words the space between lines, a subtle (but important!) part of design appearance.

50. TRACKING

Tracking is the amount of space between words (as opposed to individual characters), and like kerning and leading, is essential for legibility.

51. TYPE HIERARCHY

Hierarchy is a method of using different font sizes and styles on a website in order to organize the site and make it more visually appealing. A web page with titles in a larger font than its body text or bold text separating sections is using type hierarchy (like this very article!).

52. SERIF

Serifs are small lines attached to the end of letters or symbols in "serif" font family (Times New Roman, Lucida Bright), distinguishing those fonts from "sans serif" font families (fonts without serifs, like Arial, Helvetica).

Tech Terms 101: Front End development

53. FRONT END

Front end describes all the parts of a website that can be seen and interacted with by users. Front end web development usually involves coding with HTML, CSS, and JavaScript.

54. HTML (HYPERTEXT MARKUP LANGUAGE)

HTML is the standard language used to create web pages. It's the most basic building block you'll need for developing websites. You might remember basic HTML tags from early personal websites like Myspace, where you could customize your page with commands inside <>.

55. HTML5

HTML5 is (as of this writing) the latest version of HTML. HTML5 focuses on features that can be used on low-powered devices (making it ideal for creating mobile applications), the native ability to work with multimedia and graphic content, and new semantic web tag elements (features you use to structure your pages and documents).

56. ELEMENTS

Elements are individual HTML components of an document or webpage. For example, a paragraph in an HTML document is an element. Elements are made up of an opening tag ( ), a closing tag ( ), and information between them:

This is my paragraph!

57. META ELEMENTS

Meta elements are HTML elements that don't appear visibly for the user on a webpage, but give the web browsers additional information about the page—keywords, author of the document, last modified, etc.

58. SEMANTIC ELEMENTS

Semantic elements are HTML elements that provide information to the web browser processing the page as well as the developer building it. While non-semantic elements like and don't describe their content, semantic elements like <header>, <section>, and <article> define their content in their name.

59. STRUCTURAL ELEMENTS

Structural elements are the HTML elements used to organize the content of a web page. Structural elements like and are used to group block level and inline content together, respectively, while <header> contains the header content of a page, <footer> contains the footer content, etc.

60. ATTRIBUTES

Attributes are used to provide additional information about HTML elements. For example an HTML element like a paragraph can have an attribute of being aligned (left, center, or right). Attributes are included in the opening tag and made up of the attribute name, an equal sign, and a value in double quotes.

Example: <p = align="right">This is my paragraph!</p>

61. OPENING TAG AND CLOSING TAGS

Opening and closing tags are sets of angle brackets with an HTML element character(s) that contains a piece of content or part of the structure for a web page
Example of opening tag:
Example of closing tag:

62. SELF-CLOSING TAG

Self-closing tags are opening tags that don't have a corresponding closing tag but instead close themselves with a forward slash before the right angle bracket. If you look at this tag to insert an image, you'll see that it has the / which typically denotes closing a tag inside the opening tag.

Example:

Tech Terms 101: CSS

63. CSS (CASCADING STYLE SHEETS)

CSS is the language used to add style to documents created with HTML. Where HTML comes first and creates the foundation for a page, CSS comes along next and is used to create the page's layout, color, fonts, etc.

64. CSS3

CSS3 is the latest version of CSS, introducing features like rounded corners, shadows, gradients, transitions or animations, and new layouts like multi-columns and flexible box or grid layouts.

65. SELECTOR

Selector is a part of CSS code that defines which HTML element the CSS styling effects. For example if element "p" (say, a specific paragraph) is a selector, then:

```
<p style="font-size: 12px">
```

will change the font size of the selected element.

66. PROPERTY

A property in CSS is the part of a web page's foundational HTML code (font size, color, margin) whose appearance or style is changed by CSS.

67. VALUE

A value is the CSS code that defines the settings for for a CSS property. For example, if you're using CSS to change the font size and color of a certain block of text, the value describes what those changes will be. The value 1.5em, red, 20px will cause the property (the block of text) to have an indentation of 1.5 the font size, the text will appear red, and the font size will be 20pt.

68. DECLARATIONS

Declarations are parts of CSS code where a property and value for the selector are displayed inside a pair of curly brackets.

{font-size: 1.5em;}

Tech Terms 101: Back End development

69. BACK END

Back end refers to the "under the hood" part of a website or web service that makes it run (this includes

applications, web servers, and databases), and is typically not visible to the user interacting with the site or service.

70. APPLICATION

Applications are types of software (also called “apps” and often used to refer to mobile device software) designed to provide a function for a user or another app. Apps include everything from web browsers, to word processors, to photo and image editing tools, to chat programs like Skype and Google Hangouts.

71. API (APPLICATION PROGRAMMING INTERFACE)

An API is the interface used for building web applications. APIs provide building blocks for coders to use, and then the coders put those blocks together to create the program they are trying to build. Examples of APIs include Google Maps API (allows developers to embed Google maps on web pages), Twitter APIs (Twitter has two APIs—REST, which allows developers to access core Twitter data, and Search, which allows developers to interact with Twitter Search and trends data), and Amazon Product Advertising API (allows developers to access Amazon’s product database and advertise Amazon products on a website).

72. BUGS

Bugs are coding mistakes or unwanted pieces of code that keep a website or program from working properly.

73. DEVOPS (DEVELOPMENT OPERATIONS)

Devops is a software development process that focuses on helping development, operations, and quality teams understand each other and collaborate better. By unifying these three separate departments, Devops teams aim for shorter development cycles and more dependable software releases.

74. FRAMEWORKS

Frameworks are collection of programs and components commonly used in software development. Rather than spending time seeking out each of these components separately, developers can take advantage of frameworks where they are collected in one place. Examples of frameworks include Ruby on Rails, Bootstrap, AngularJS, and Joomla.

75. OBJECT-ORIENTED PROGRAMMING (OOP)

Object-oriented programming is a type of computer programming/software design that focuses on the creation of objects that have specific and unique attributes and abilities. In OOP, computer programs are made up of these created objects that then interact with one another. This is in contrast to earlier programming languages that focus more on the process of turning input data into output data rather than the data (objects) itself. Examples of OOP languages include Ruby, PHP, and Python.

76. SOFTWARE

Software is a program or set of instructions that tells a computer, phone, or tablet what to do. Software includes individual applications (web browsers, word processors) as well as system software like operating systems (Microsoft Windows), drivers (software that allows operating systems to communicate with hardware like printers), and utilities (tools like anti-virus programs or hard drive defragmenters).

77. TEXT EDITOR

A text editor is a type of software used to write plain text (without formatting), which is often used for coding and programming. After code has been written, it's formatted another program called a compiler (specific to the programming language you're using) to make it computer-readable, but code normally starts it's life on a plain text editor. Examples of text editors include SublimeText, TextEdit, TextWrangler, and Notepad++.

78. VERSION CONTROL

Version control is a tool used to keep track of changes to code and files on a website or app and allows the user to go back and restore earlier versions in case of bugs. Version control tools like Git are often built in to source code editors (programs used to write and edit code) like Visual Code Studio, or offered as part of a web hosting service.

79. WEB SERVERS

Web servers are computers used to store websites, online apps, documents, pictures, or other data, and can be accessed through the internet by way of applications like web browsers or file transfer protocol (FTP) clients. When you visit a website with the browser on your computer or smartphone, you are requesting it from a web server.

Tech Terms 101: Systems

80. CACHING

Caching is what happens when a web browser stores recurring website assets—like images and font styles—so that the website will load faster on repeat visits from the same user. By caching, your browser won't have to reinvent the wheel every time the site is requested.

81. CLOUD COMPUTING

Cloud computing is a practice where data is not stored locally on your own computer, but instead is spread out among a number of remote servers accessible through the internet. Service like Google Docs, Facebook, and Gmail are examples of cloud computing—you are interacting with data on your home computer that is stored externally in "the cloud."

82. FIREWALLS

Firewalls are systems designed to protect and secure a computer network—everything from a commercial web service to your home WiFi network—from external

security risks. Firewalls monitor inbound and outbound network traffic and determine whether or not to allow the traffic through based on a user-defined set of security standards.

83. ROUTERS

Routers are the devices used to connect personal computers to the internet via a service provider like cable or DSL. The device that your computer is either directly connected to via an ethernet cable or that you use to access your home WiFi network is a router.

84. UPTIME & DOWNTIME

Uptime and downtime describes how long a website, computer, or other system has been working (uptime) or not working (downtime).

85. VIRTUAL MACHINE (VM)

Virtual machine is software that makes it possible to use one computer operating system (like Windows 10) on a computer running another system (like a MacBook Pro). Examples include Virtual Box, Parallels, and VM Ware.

86. VPN (VIRTUAL PRIVATE NETWORK)

VPNs are networks that allow public internet connections to be used as private networks as a means of improving security. When you use a public WiFi network, for instance, your device and data are theoretically accessible by everyone else on the network. When using a VPN, you're still able to access the internet through the public network, but are shielded by the VPN. VPN's can be set

up through some web browsers (like Opera), or can be accessed through paid monthly services.

Tech Terms 101: Data

87. BIG DATA

Big data is a term for collections of data that are so large they can't be processed through traditional data processing systems. These collections come from sources like mobile devices, emails, search keywords, user database information, applications, and servers. By finding ways to comb through this data, companies can identify consumer patterns and use them to predict and optimize their business.

88. DATA ARCHITECTURE

How data is collected, stored, accessed, and used in companies and organizations

89. DATABASE

Data architecture describes the way data is collected, stored, accessed, and used in companies and organizations. It can be seen as the roadmap for how data flows across an organization's IT systems and applications.

90. DATA MODELING

Determining what kind of data is needed and how it will be structured and organized

91. DATA VISUALIZATION

Data visualization is the use of graphs, charts, tables, infographics, etc. in order to define and communicate data

being analyzed and the the findings that have come from it.

92. RELATIONAL DATABASE MANAGEMENT SYSTEM (RDMS)

Relational database management systems are used to organize data into tables—the data can then be accessed or reassembled without having to reorganize the database tables. Examples of RDMS include SAP and MySQL.

Tech Terms 101: Mobile

93. HYBRID APP

Hybrid apps are applications that will work on different platforms (computers, mobile devices, tablets) and are a combination of a native app (one that is stored locally on your device) and a web app (one that is accessed through a web browser on the Internet). Amazon and Apple App Stores, Twitter, Yelp, and Gmail are all examples of hybrid apps.

94. IDE (INTEGRATED DEVELOPMENT ENVIRONMENT)

IDE is a software application that includes tools like source code editors (text editors specifically designed for writing code), debuggers (tools for testing code), and build automation tools (tools for compiling code into machine-readable format and running automated tests). Examples of IDE's include Eclipse, Visual Studio, and VIM.

95. NATIVE APPS

Native apps are made specifically for certain platforms. They only run on the platforms they were built for, and are stored locally on those devices. Mail for iOS and Ingress for Android are both examples of native apps.

96. NFC (NEAR FIELD COMMUNICATION)

NFC is technology that lets mobile devices communicate using radio waves when they're very close to each other (about four inches or less) and is used for services like sharing files, pairing accessories, or wireless payments. Companies can use NFC to make products interactive with consumers' mobile devices, such as including NFC-enabled tags on products that allow users to register products, get discounts, or place new orders with their mobile device.

97. RESPONSIVE WEB DESIGN

Responsive web design is the practice of designing websites so that they adapt gracefully to different-sized devices like phones, tablets, wearable devices, etc. If you're able to visit a website on your phone and it looks just as proportional and seamless as it does on your computer, it's an example of responsive web design.

98. SDK (SOFTWARE DEVELOPMENT KIT)

SDK is a set of tools for creating specific types of software. SDK's are released by companies that control the platform the software is being developed for. For instance, apps developed for iOS require the iOS SDK,

Windows apps require the .NET Framework SDK, and Java apps require the Java Development Kit.

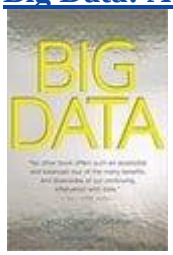
99. WEB APPS

Web apps are websites that look and feel like an app (as opposed to a series of linked pages). Facebook, Pandora, and Google Docs are all examples of web apps.

Bibliography.

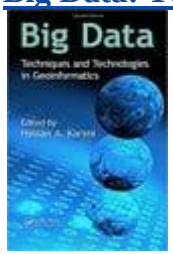
RESOURCES & REFERENCES

Big Data: A Revolution That Will Transform How We Live, Work, and Think



"Whether it is used by the NSA to fight terrorism or by online retailers to predict customers' buying patterns, big data is a revolution occurring around us, in the process of forever changing economics, science, culture, and the very way we think. But it also poses new threats, from the end of privacy as we know it to the prospect of being penalized for things we haven't even done yet, based on big data's ability to predict our future behavior. What we have already seen is just the tip of the iceberg. Big Data is the first major book about this earthshaking subject, with two leading experts explaining what big data is, how it will change our lives, and what we can do to protect ourselves from its hazards."

Big Data: Techniques and Technologies in Geoinformatics

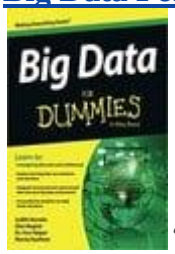


"Providing a perspective based on analysis of time, applications, and resources, this book familiarizes readers with geospatial applications that fall under the category of big data. It explores new trends in geospatial data collection, such as geo-crowdsourcing and advanced data collection technologies such as LiDAR point clouds. The book features a range of topics on big data techniques and technologies in geoinformatics including distributed computing, geospatial data analytics, social media, and volunteered geographic information. Big Data: Techniques and Technologies in Geoinformatics tackles these challenges head on, integrating coverage of techniques and technologies for storing, managing, and computing geospatial big data."

Designing Data-Intensive Applications; The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

"In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively."

Big Data For Dummies

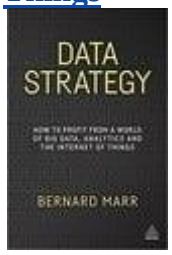


"Big data management is one of the major challenges facing business, industry, and not-for-profit organizations. Data sets such as customer transactions for a mega-retailer, weather patterns monitored by meteorologists, or social network activity can quickly outpace the capacity of traditional data management tools. If you need to develop or manage big data solutions, you'll appreciate how these four experts define, explain, and guide you through this new and often confusing concept. You'll learn what it is, why it matters, and how to choose and implement solutions that work. Big Data For Dummies cuts through the confusion and helps you take charge of big data solutions for your organization."

Big Data: Principles and best practices of scalable realtime data systems

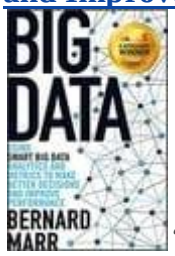
"Big Data teaches you to build big data systems using an architecture designed specifically to capture and analyze web-scale data. This book presents the Lambda Architecture, a scalable, easy-to-understand approach that can be built and run by a small team. You'll explore the theory of big data systems and how to implement them in practice. In addition to discovering a general framework for processing big data, you'll learn specific technologies like Hadoop, Storm, and NoSQL databases. This book requires no previous exposure to large-scale data analysis or NoSQL tools. Familiarity with traditional databases is helpful."

Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things



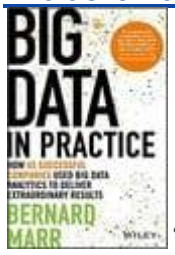
"Bernard Marr's Data Strategy is a must-have guide to creating a robust data strategy. Explaining how to identify your strategic data needs, what methods to use to collect the data and, most importantly, how to translate your data into organizational insights for improved business decision-making and performance, this is essential reading for anyone aiming to leverage the value of their business data and gain competitive advantage. Packed with case studies and real-world examples, advice on how to build data competencies in an organization and crucial coverage of how to ensure your data doesn't become a liability, Data Strategy will equip any organization with the tools and strategies it needs to profit from big data, analytics and the Internet of Things."

Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance



"There is so much buzz around big data. We all need to know what it is and how it works – that much is obvious. But is a basic understanding of the theory enough to hold your own in strategy meetings? Probably. But what will set you apart from the rest is actually knowing how to USE big data to get solid, real-world business results – and putting that in place to improve performance. Big Data will give you a clear understanding, blueprint, and step-by-step approach to building your own big data strategy. This is a well-needed practical introduction to actually putting the topic into practice. Illustrated with numerous real-world examples from a cross section of companies and organisations, Big Data will take you through the five steps of the SMART model: Start with Strategy, Measure Metrics and Data, Apply Analytics, Report Results, Transform."

Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results



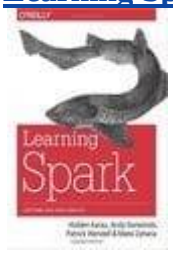
"From technology, media and retail, to sport teams, government agencies and financial institutions, learn the actual strategies and processes being used to learn about customers, improve manufacturing, spur innovation, improve safety and so much more.

Organised for easy dip-in navigation, each chapter follows the same structure to give you the information you need quickly. For each company profiled, learn what data was used, what problem it solved and the processes put it place to make it practical, as well as the technical details, challenges and lessons learned from each unique scenario."

The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios

"Comprising dozens of examples that address different industries and departments (healthcare, transportation, finance, human resources, marketing, customer service, sports, etc.) and different platforms (print, desktop, tablet, smartphone, and conference room display) The Big Book of Dashboards is the only book that matches great dashboards with real-world business scenarios. By organizing the book based on these scenarios and offering practical and effective visualization examples, The Big Book of Dashboards will be the trusted resource that you open when you need to build an effective business dashboard."

Learning Spark: Lightning-Fast Big Data Analysis



"Data in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning."

Big Data at Work: Dispelling the Myths, Uncovering the Opportunities

"When the term "big data" first came on the scene, bestselling author Tom Davenport (Competing on Analytics, Analytics at Work) thought it was just another example of technology hype. But his research in the years that followed changed his mind. With

dozens of company examples, including UPS, GE, Amazon, United Healthcare, Citigroup, and many others, this book will help you seize all opportunities—from improving decisions, products, and services to strengthening customer relationships. It will show you how to put big data to work in your own organization so that you too can harness the power of this ever-evolving new resource."

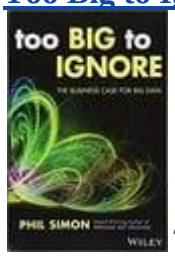
Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data

"Data Science and Big Data Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!"

Big Data For Beginners: Understanding SMART Big Data, Data Mining & Data Analytics For improved Business Performance, Life Decisions & More (Data … Computer Programming, Growth Hacking, ITIL)

"Big Data For Beginners! The Ultimate Beginners Crash Course To Understanding And Interpreting Big Data! Are You Ready To Learn How To Understand SMART Big Data, Data Mining & Data Analytics For improved Business Performance, Life Decisions & More? If So You've Come To The Right Place – Regardless Of How Little Experience You May Have! Here's A Preview Of What Big Data For Beginners! Contains… A Conundrum Called 'Big Data' How To Understand Big Data Better What Can Big Data Do For You? Understanding The Analytics (And The Importance) The Obstacles And Importance Of The Big Data Situation We're In A Closer Look At Key Big Data Challenges Generating Business Value through Data Mining And Much, Much More!"

Too Big to Ignore: The Business Case for Big Data (Wiley and SAS Business Series)



"Big Data is no temporary blip of fad. In fact, it is only going to intensify in the coming years, and its ramifications for the future of business are impossible to

overstate. Too Big to Ignore explains why Big Data is a big deal. Simon provides commonsense, jargon-free advice for people and organizations looking to understand and leverage Big Data. Rife with case studies, examples, analysis, and quotes from real-world Big Data practitioners, the book is required reading for chief executives, company owners, industry leaders, and business professionals."

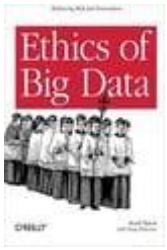
People Analytics in the Era of Big Data: Changing the Way You Attract, Acquire, Develop, and Retain Talent

"People Analytics in the Era of Big Data provides a blueprint for leveraging your talent pool through the use of data analytics. Written by the Global Vice President of Business Intelligence and Predictive Analytics at Monster Worldwide, this book is packed full of actionable insights to help you source, recruit, acquire, engage, retain, promote, and manage the exceptional talent your organization needs. With a unique approach that applies analytics to every stage of the hiring process and the entire workforce planning and management cycle, this informative guide provides the key perspective that brings analytics into HR in a truly useful way."

Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses

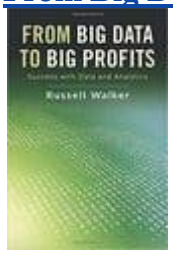
"The availability of Big Data, low-cost commodity hardware and new information management and analytics software has produced a unique moment in the history of business. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue and profitability. The Age of Big Data is here, and these are truly revolutionary times. This timely book looks at cutting-edge companies supporting an exciting new generation of business analytics."

Ethics of Big Data: Balancing Risk and Innovation



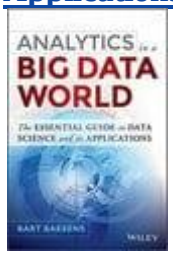
"What are your organization's policies for generating and using huge datasets full of personal information? This book examines ethical questions raised by the big data phenomenon, and explains why enterprises need to reconsider business decisions concerning privacy and identity. Authors Kord Davis and Doug Patterson provide methods and techniques to help your business engage in a transparent and productive ethical inquiry into your current data practices. With this book, you'll learn how to align your actions with explicit company values and preserve the trust of customers, partners, and stakeholders."

From Big Data to Big Profits: Success with Data and Analytics



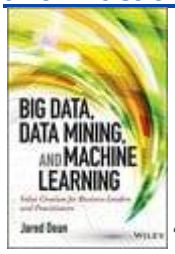
"In From Big Data to Big Profits, Russell Walker investigates the use of Big Data to stimulate innovations in operational effectiveness and business growth. Walker examines the nature of Big Data and how businesses can use it to create new monetization opportunities. Using case studies of Apple, Netflix, Google, LinkedIn, Zillow, Amazon, and other leaders in the use of Big Data, Walker explores how digital platforms such as mobile apps and social networks are changing the nature of customer interactions and the way Big Data is created and used by companies. Such changes, as Walker points out, will require careful consideration of legal and unspoken business practices as they affect consumer privacy."

Analytics in a Big Data World: The Essential Guide to Data Science and its Applications (Wiley and SAS Business Series)



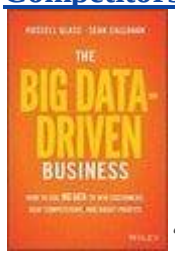
"The book draws on author Bart Baesens' expertise on the topics of big data, analytics and its applications in e.g. credit risk, marketing, and fraud to provide a clear roadmap for organizations that want to use data analytics to their advantage, but need a good starting point. Baesens has conducted extensive research on big data, analytics, customer relationship management, web analytics, fraud detection, and credit risk management, and uses this experience to bring clarity to a complex topic. For organizations looking to enhance their capabilities via data analytics, this resource is the go-to reference for leveraging data to enhance business capabilities."

Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners (Wiley and SAS Business Series)



"Big data is big business. But having the data and the computational power to process it isn't nearly enough to produce meaningful results. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners is a complete resource for technology and marketing executives looking to cut through the hype and produce real results that hit the bottom line. Providing an engaging, thorough overview of the current state of big data analytics and the growing trend toward high performance computing architectures, the book is a detail-driven look into how big data analytics can be leveraged to foster positive change and drive efficiency."

The Big Data-Driven Business: How to Use Big Data to Win Customers, Beat Competitors, and Boost Profits



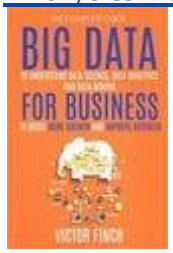
"The Big Data-Driven Business: How to Use Big Data to Win Customers, Beat Competitors, and Boost Profits makes the case that big data is for real, and more than just big hype. Better information allows for better decisions, better targeting, and better reach. Big data has become an indispensable tool for the most effective marketers in the business, and it's becoming less of a competitive advantage and more like an industry standard. Remaining relevant as the marketplace evolves requires a full understanding and application of big data, and The Big Data-Driven Business provides the practical guidance businesses need."

Big Data Marketing: Engage Your Customers More Effectively and Drive Value

"Many of today's businesses find themselves caught in a snarl of internal data, paralyzed by internal silos, and executing antiquated marketing approaches. As a result, consumers are losing patience, shareholders are clamoring for growth and differentiation, and marketers are left struggling to untangle the massive mess. Big Data Marketing provides a strategic road map for executives who want to clear the chaos and start driving competitive advantage and top line growth. Using real-world examples, non-technical

language, additional downloadable resources, and a healthy dose of humor, Big Data Marketing will help you discover the remedy offered by data-driven marketing."

Big Data For Business: Your Comprehensive Guide To Understand Data Science, Data Analytics and Data Mining To Boost More Growth and Improve Business (Data Analytics Book Series)

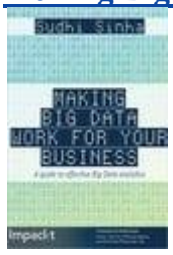


"This Big Data For Business: Your Comprehensive Guide To Understand Data Science, Data Analytics and Data Mining To Boost More Growth and Improve Business. is your must have guide to explore and learn about the impact of Big Data For Business, and understand how you can starts forming ideas on how you can use it for your next business growth. The Bottom Line: What are you waiting for? Start today by making the smartest investment you could possibly make. An investment in yourself, your knowledge and your business growth."

Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault

"Today, the world is trying to create and educate data scientists because of the phenomenon of Big Data. And everyone is looking deeply into this technology. But no one is looking at the larger architectural picture of how Big Data needs to fit within the existing systems (data warehousing systems). Taking a look at the larger picture into which Big Data fits gives the data scientist the necessary context for how pieces of the puzzle should fit together. Most references on Big Data look at only one tiny part of a much larger whole. Until data gathered can be put into an existing framework or architecture it can't be used to its full potential. Data Architecture a Primer for the Data Scientist addresses the larger architectural picture of how Big Data fits with the existing information infrastructure, an essential topic for the data scientist."

Making Big Data Work for Your Business

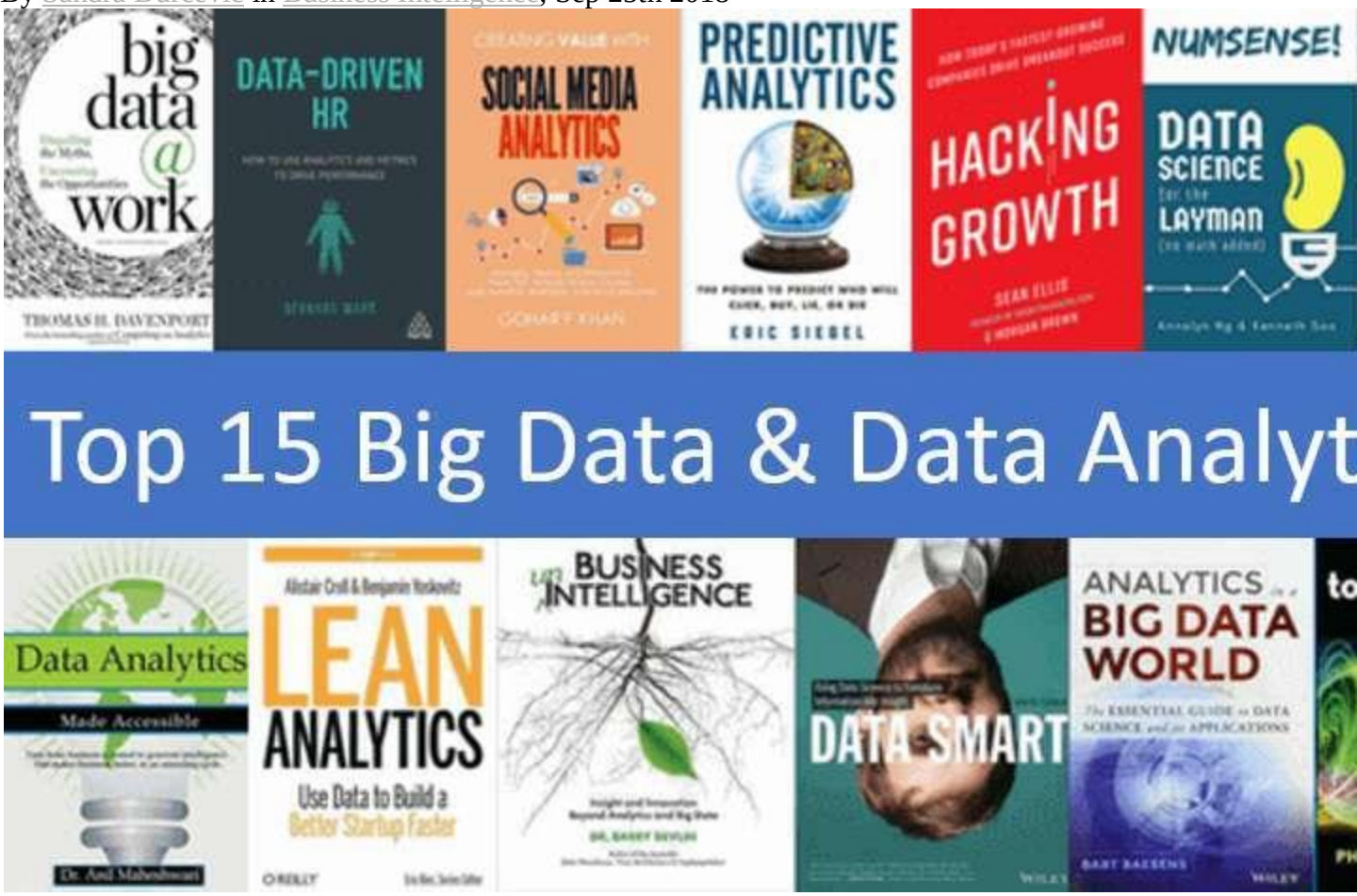


"Big Data has brought about a revolution in the way we do business. Essential business decisions can today be informed by the wealth of data now at our disposal. However, while Big Data may appear to be the answer to every business problem, for many,

gaining real value from data – gaining business insights is a difficult task. Big Data, for many, is a big problem itself, with many struggling to reap the rewards that it promises. In this accessible and stimulating guide, Sudhi Sinha, management, technology and sustainability expert, provides a unique perspective on Big Data and how to derive maximum value from it – with sharp and careful analytics."

Unlock The Power of Your Data With These 15 Big Data & Data Analytics Books

By Sandra Durcevic in Business Intelligence, Sep 25th 2018



The saying “knowledge is power” has never been more relevant, thanks to the widespread commercial use of big data and data analytics.

This trend has been brought about by the new demands of the modern marketplace, and it's here to stay. The rate at which data is generated has increased exponentially in recent years. To put this into perspective, 40,000 search queries are performed per second via Google alone – this equates to 3.46 million searches per day and 1.2 trillion each year.

Companies, both big and small, are seeking the best ways to leverage their data into a competitive advantage. With that in mind, we have prepared a list of top 15 definitive data analytics and big data books, along with magazines and authentic readers' reviews upvoted by the Goodreads community. Whether you are a complete novice or a seasoned business intelligence professional, you will find here some books on data analytics that will help you cultivate your understanding of this essential field. And with that understanding, you'll be able to tap into the potential of data analytics to create strategic advantages, exploit your metrics to shape them into stunning business dashboards, and identify new business opportunities or at least participate in the process.

Exclusive Bonus Content: Not Sure Which Big Data Book To Read?

Download our summarized, free guide on top 15 best books on big data!

Before we delve any deeper, here are three big data analytics insights to put its relevance and importance into perspective.

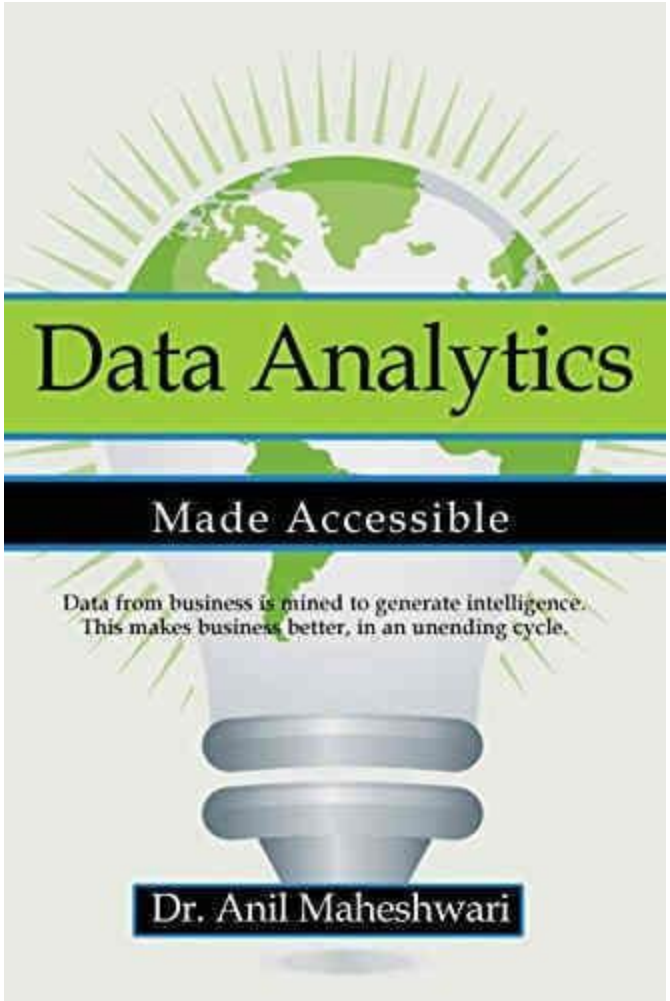
Essential Big Data And Data Analytics Insights

- In a mere five years from now, the number of smart connected devices on the planet will be more than 50 billion – all of which will generate data that can be shared, collected, and analyzed.

- The White House has invested an incredible $200 million in big data projects – a true testament to the growing importance and relevance of big data analytics across sectors.
- As of this moment, just 5% of all accessible data is analyzed and used – just think of the potential.

The Best Data Analytics And Big Data Books Of All Time

1) Data Analytics Made Accessible, by A. Maheshwari



Best for: the new intern who has no idea what data science even means

An excerpt from a rave review:

"I would definitely recommend this book to everyone interested in learning about Data Analytics from scratch and would say it is the best resource available among all other Data Analytics books."

If we had to pick one book for an absolute newbie to the field of Data Science to read, it would be this one. Updated for 2018, "Bussiness Intelligence and Data Mining Made Accessible" is inarguably the best book there is on data analytics, and does exactly what its name implies: it explains data analytics in an easy way, and makes it understandable and digestible for the uninitiated.

The book promotes easy understanding through:

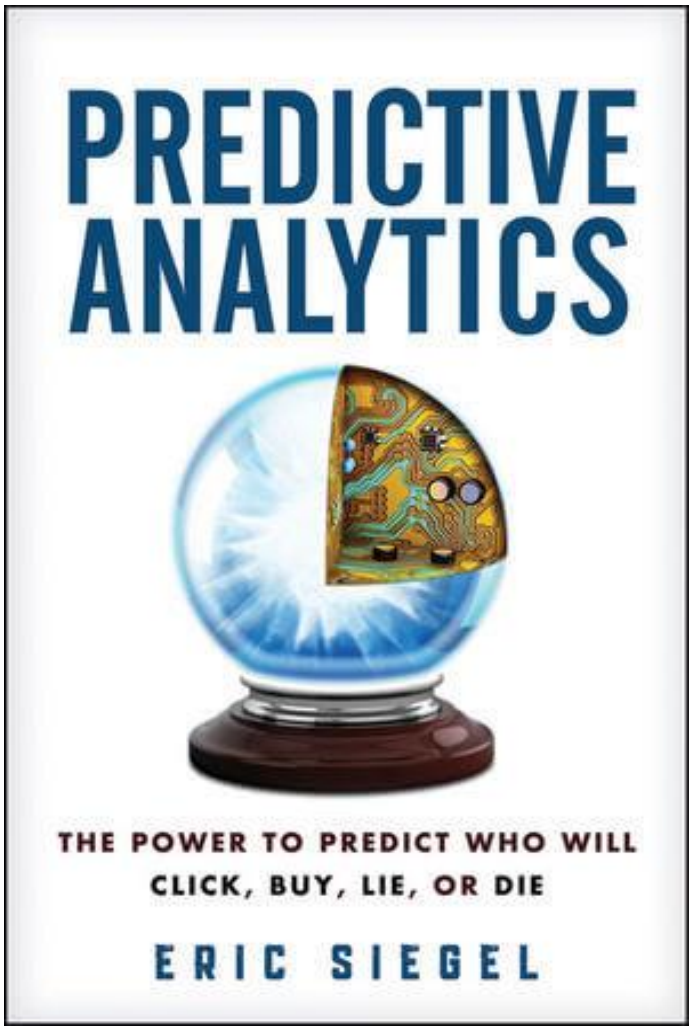
- Concrete, real-world examples at the beginning of each chapter
- An intuitively organized layout structured like a semester-long college course
- Case studies in each chapter to tie the material together

Due to its scope of content and clear explanation, "Data Analytics Made Accessible" has been made a college textbook for many universities in the US and worldwide. The author, Anil Maheshwari, Ph.D., has both practical and intellectual knowledge of data analytics; he worked in data science at IBM for 9 years before becoming a professor.

The book also has some "crowdsourced" material, as the 2017 edition had 4 chapters added based on feedback from reviewers and readers. At 156 pages on Kindle, this is a book you could finish in one (long) sitting if you were so inclined, and that you can also

use as an inspiration when you work on your business intelligence strategy.

2) Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by E. Siegel



Best for: someone who has heard a lot of buzz about predictive analytics, but doesn't have a firm grasp on the subject

An excerpt from a rave review:

"The Freakonomics of big data."

—Stein Kretsinger, founding executive, Advertising. com

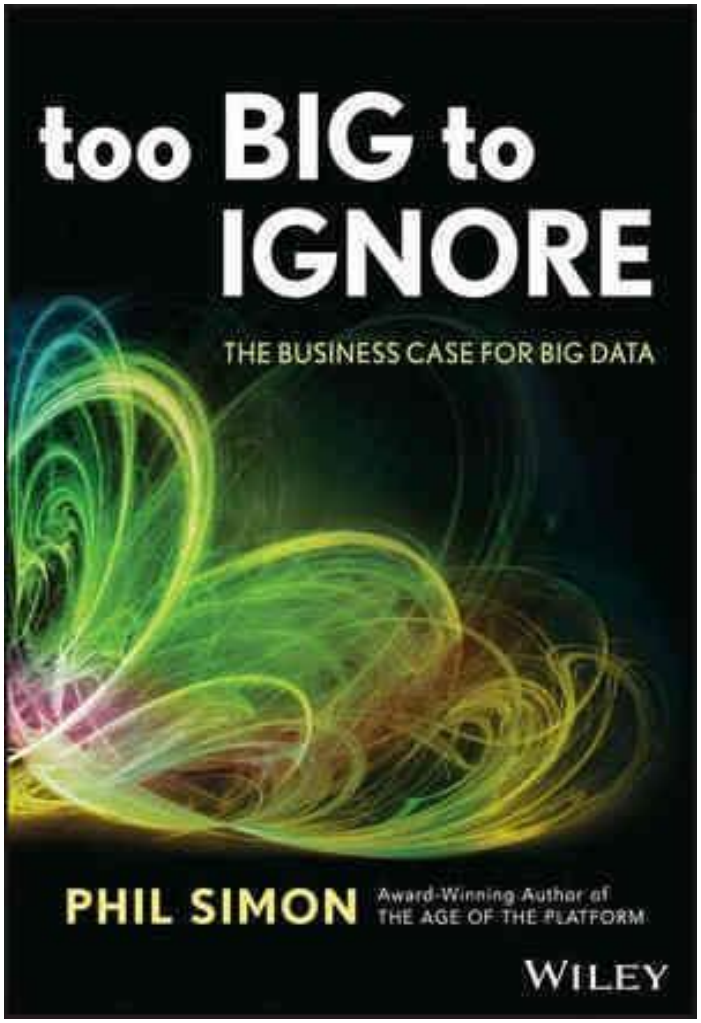
We have included predictive analytics in our list of the most prominent business intelligence trends, as it has been widely recognized as the strategy that makes it possible to unleash the power of big data. From a business perspective, predictive analytics is used to analyze current data and historical facts in order to better understand customers, products, and competitors and to identify potential risks and opportunities for a company.

However, due to its vast application, predictive analytics should not concern only business professionals. Most people are aware that companies collect our GPS locale, text messages, credit card purchases, social media posts, Google search history, etc., and this book will give you an insight into their data collecting procedures and the reasons behind them.

Eric Siegel's data analytics book is an eye-opening read for anyone who wants to learn what predictive analytics is, and how predictive analytics can be deployed across a wide range of disciplines. It is not a manual, so a data scientist looking for instructions would be disappointed. Although there is some discussion of algorithms including linear regression or decision trees, it's easy to understand even for a layman.

Siegel's book makes it clear that predictive analytics is not a sneaky procedure used by companies to sell more, but a significant leap in technology which, by predicting human behavior, can help combat financial risk, improve health care, reduce spam, toughen crime-fighting, and yes, boost sales. It was lately revised and updated in January 2016.

3) Too Big to Ignore: The Business Case for Big Data, by award-winning author P. Simon



Best for: the member of your management team who rolls their eyes whenever big data or predictive analytics are brought up

An excerpt from a rave review:

"Simon provides a very thorough exploration for non-technologists into the new world of "Big Data" with many illustrations of how companies are beginning to exploit this resource to their advantage."

There are two types of people who should read this book: people who don't believe in the merits of big data and predictive analytics, and people who are so interested in these topics that they love learning about the current use cases of these technologies and this is what makes it one of the best big data books.

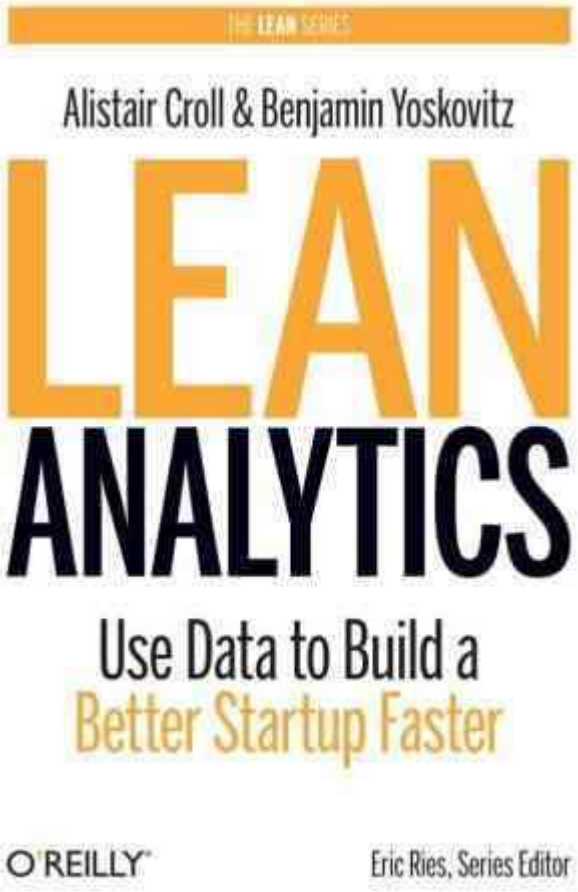
"Too Big To Ignore" examines many examples of how companies (and local governments!) are using big data to their advantage, including:

- Progressive Insurance's use of GPS trackers/accelerometers which determine customer safety ratings
- Google's ability to predict local flu outbreaks by measuring spikes in flu-related local searches
- The government of Boston fixing potholes using data that residents enter into their smartphones

The author, Phil Simon, being a speaker who has made keynotes at EA, Cisco, Zappos, and Netflix, is an expert at making technical information simple. Simon makes the case that big data is not only an area of potential innovation- it's a crucial factor that your company must address now to survive in the modern marketplace. His argument contains urgency and clarity, centering around this point: big data is no fad. It's a huge change in how business is conducted, and it's already happening.

Remarkably free of jargon and filled with case studies and examples, "Too Big To ignore" is an excellent introduction to big data, as seen through the lens of: what can big data do for me and my business?

4) Lean Analytics: Use Data to Build a Better Startup Faster, by A. Croll and B. Yoskovitz



Best for: anyone in your company who wants to deeply understand your customers through the use of data analytics

Excerpts from rave reviews:

"As useful for today's multi-billion dollar companies as it is for entrepreneurs."

— John Stormer, Salesforce.com

“Your competition will use this book to outgrow you.”

— Mike Volpe, Hubspot

Eric Reis started a global movement by releasing the book “The Lean Startup”. The philosophy of the book revolved around getting feedback from customers as quickly as possible and iterating rapidly based on that feedback. It was only a matter of time before the “lean philosophy” was applied to data analytics.

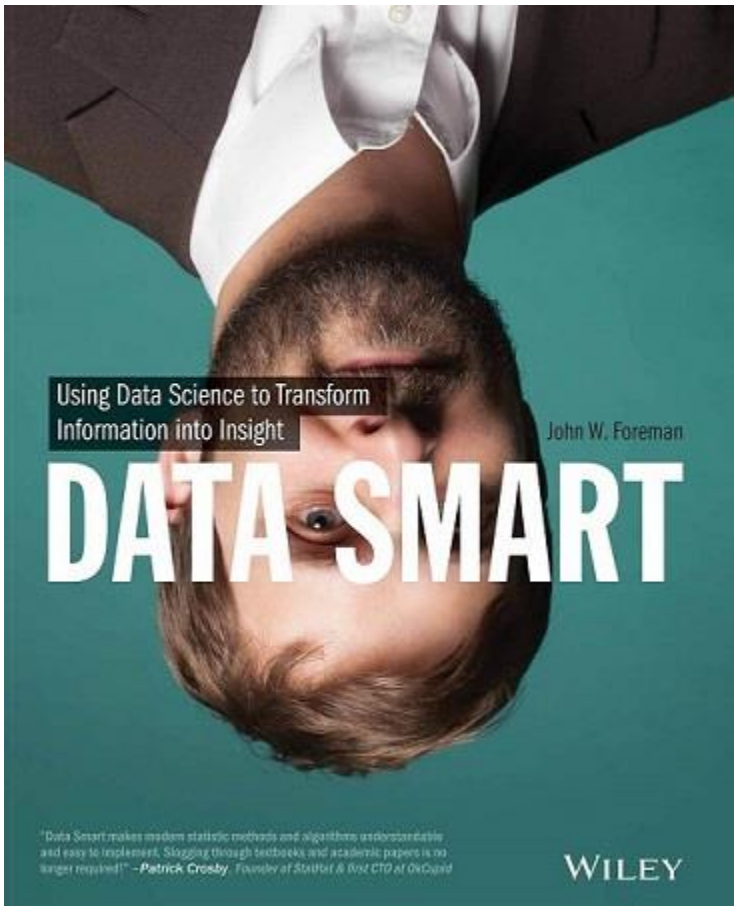
However, don’t be deceived – just as you don’t need to be a literal startup to gain a lot of value from Eric Ries’ book, companies of all sizes and shapes can learn a lot of valuable information from “Lean Analytics”. The book has three main ideas:

1. The biggest risk your company faces is investing a lot of time and resources into building something that the market doesn’t want.
2. Product/market fit is THE most important factor to get right.
3. By using the right analytics metrics, you can determine which products or services to focus on or build – and how to market them.

In today’s world, every company faces the potential to be disrupted. It’s up to you: do you want to disrupt your own company from the inside by being an intrapreneur, or are you going to let someone else disrupt you in the market?

Reading this book will give you the toolkit you need to make sure the former happens and not the latter.

5) Data Smart: Using Data Science to Transform Information into Insight, by J. W. Foreman



Best for: a somewhat technical reader who is good with Excel, but doesn't know much about data science

An excerpt from a rave review:

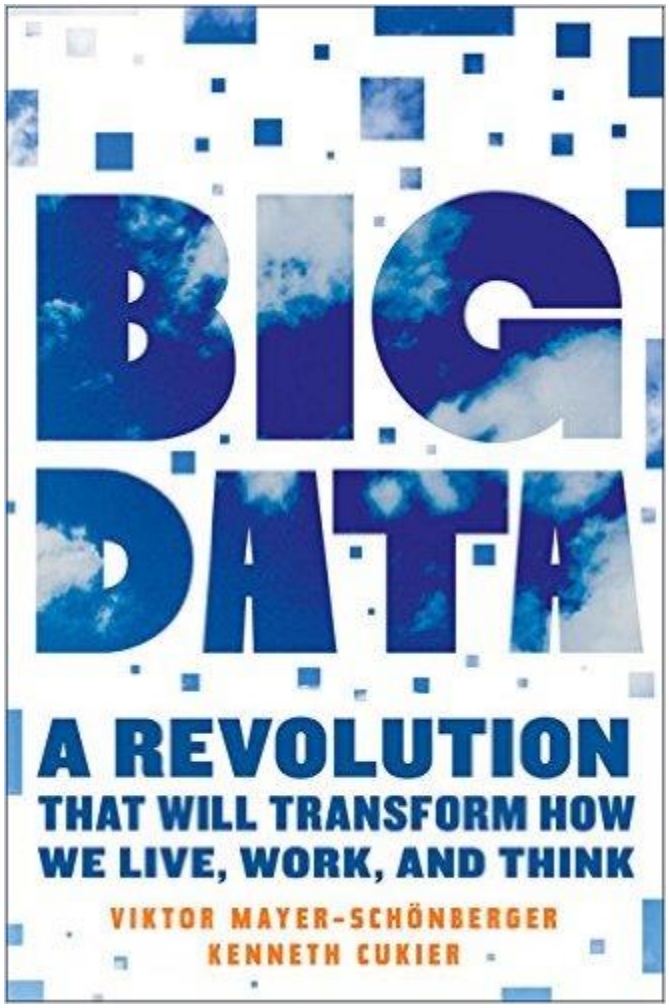
"What I like most about the book is that it doesn't try to wave a magic data wand to cure all of your company's ills. Instead it focuses on a few areas where data and analytic techniques can deliver a concrete benefit, and gives you just enough to get started."

‘Data Smart’ contains concrete hints on which analytic techniques to apply to effectively crunch data. It’s a useful read for anyone with a little background in applied mathematics and a spreadsheet program on their PC. It is a well thought out and designed tutorial with many easy-to-understand real-world examples for a business professional who must work with data sets.

Each chapter covers a different technique in a spreadsheet, including nonlinear programming and genetic algorithms, clustering, graph modularity, data mining in graphs, supervised AI through logistic regression, ensemble models, forecasting, seasonal adjustments, and prediction intervals through Monte Carlo simulation as well as moving from spreadsheets into the R programming language.

‘Data Smart’ contains enough practical information to actually start performing analyses by using good old Microsoft Excel. Its goal isn’t to revolutionize your business with additional software, but rather to make incremental improvements to processes with accessible analytic techniques. However, once you start working with larger enterprise-level data sets with millions of rows and hundreds of columns of information, Excel may not be capable of handling such volumes. At this point, turning to self-service business intelligence would be the most affordable and effective solution.

6) Big Data: A Revolution That Will Transform How We Live, Work, and Think by V. Mayer-Schönberger and K. Cukier



Best for: the reader interested in how big data can improve the quality of our lives (and not just in a business sense)

An excerpt from a rave review:

"An optimistic and practical look at the Big Data revolution — just the thing to get your head around the big changes already

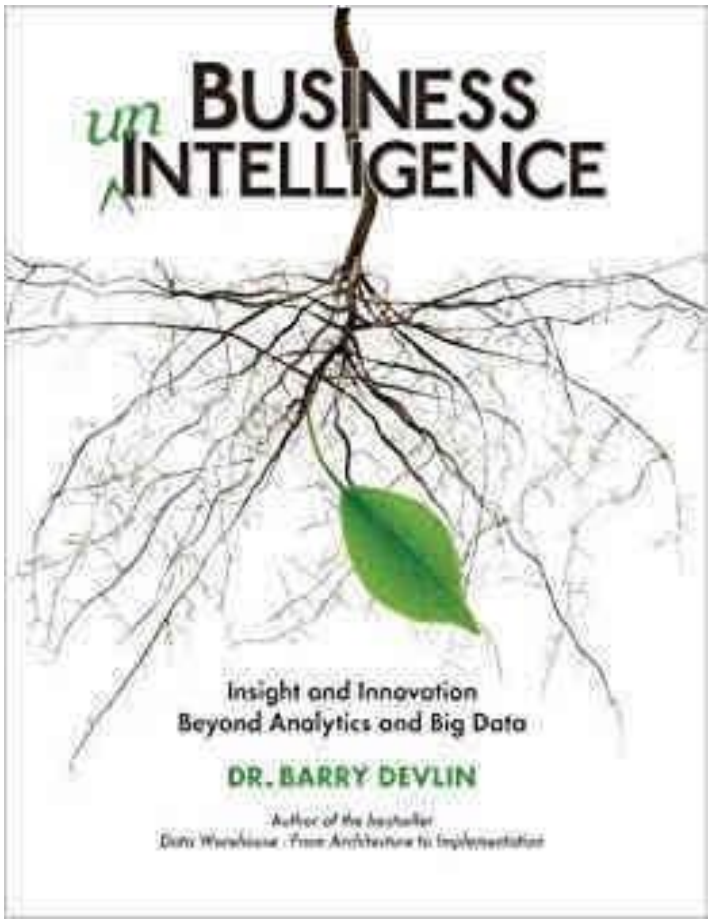
underway and the bigger changes to come ." —Cory Doctorow, boingboing.com

This is another big data book that provides readers with a more general view on key issues around Big Data, with the authors offering their opinions and insights on how the technology will proceed. This would be a perfect read for people new to the subject who want to understand in what way big data can be leveraged to improve people's life quality – from identifying consumers' shopping patterns to predicting flu outbreaks.

The book also sheds light on how big data's key characteristics (volume, variety, velocity, and veracity) will change the way we process and manage data. It mentions the completeness of data (as opposed to sampling), the power to quantify and digitize new formats of information that were previously inaccessible, as well as the ability to use new databases (like Hadoop and NoSQL) and statistical tools (machine learning and data mining) to describe huge quantities of data.

Exclusive Bonus Content: Not Sure Which Data Analytics Book To Read?
Download our free guide on top 15 best books on data analytics!

7) Business UnIntelligence: Insight and Innovation Beyond Analytics and Big Data, by B. Devlin



Best for: the seasoned business intelligence professional who is ready to think deep and hard about important issues in data analytics and big data

An excerpt from a rave review:

"…a tour de force of the data warehouse and business intelligence landscape. It drills into every nook and cranny of the industry, the great successes as well as the depths of insanity (and there's plenty

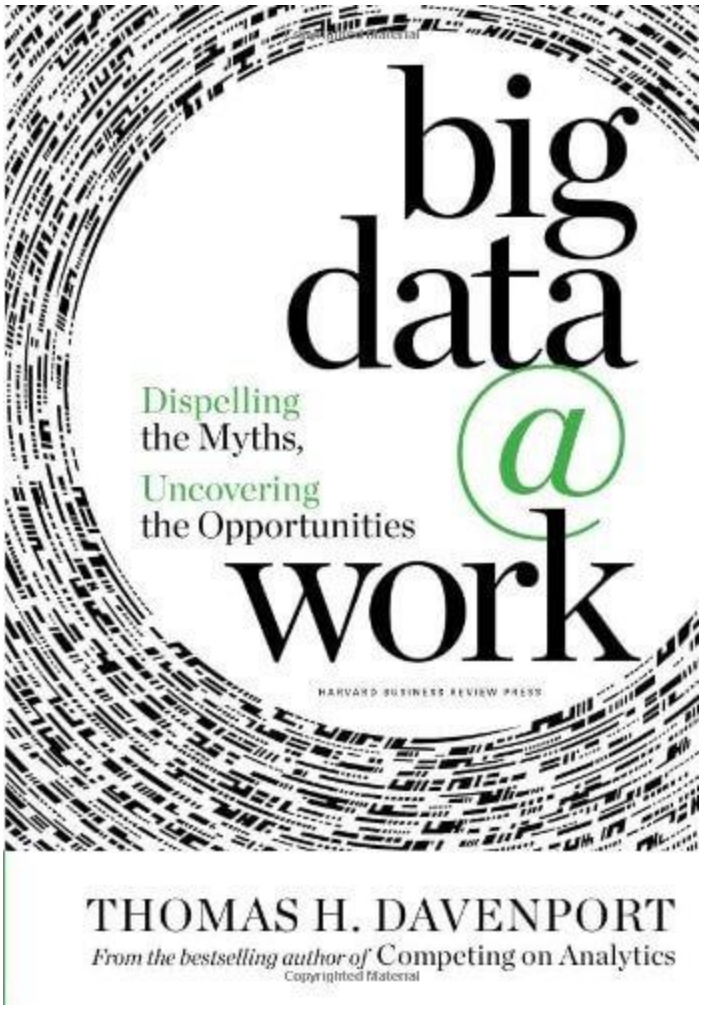
of both revealed). This book details what the true 'Father of Data Warehousing' thinks of his children and it's not always pretty…"

This book is most useful for someone who lives and breathes BI – and who is ready to critically look at their ideas surrounding the field. In this at-times contrarian and unflinching book, Dr. Barry Devlin shows how modern BI often fails to deal with data from mobile, social media, and the Internet of Things in a meaningful way. Devlin also makes the argument that modern business decisions must be made from a combination of data-driven (rational) and emotional (intuitive) sources, as opposed to only using data – and that business intelligence must reflect those needs.

The book additionally serves as a history of the field of business intelligence, big data, and data analytics, as Devlin details the past, present, and future of the field. He does so in order to challenge many of the assumptions in modern data analytics and data gathering, by showing how quickly the old best practices have become outdated due to the sheer volume and velocity of modern data sources.

If you're ready to be challenged to think differently, "Business unIntelligence" is amongst the best data analytics books to do so.

8) Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, by T. H. Davenport



Best for: managers who want to start and manage the big data journey in both small and large organizations

An excerpt from a rave review:

"It's a required reading for managers that need a straightforward, hype-free introduction to big data, a clear and clarifying "signal" in

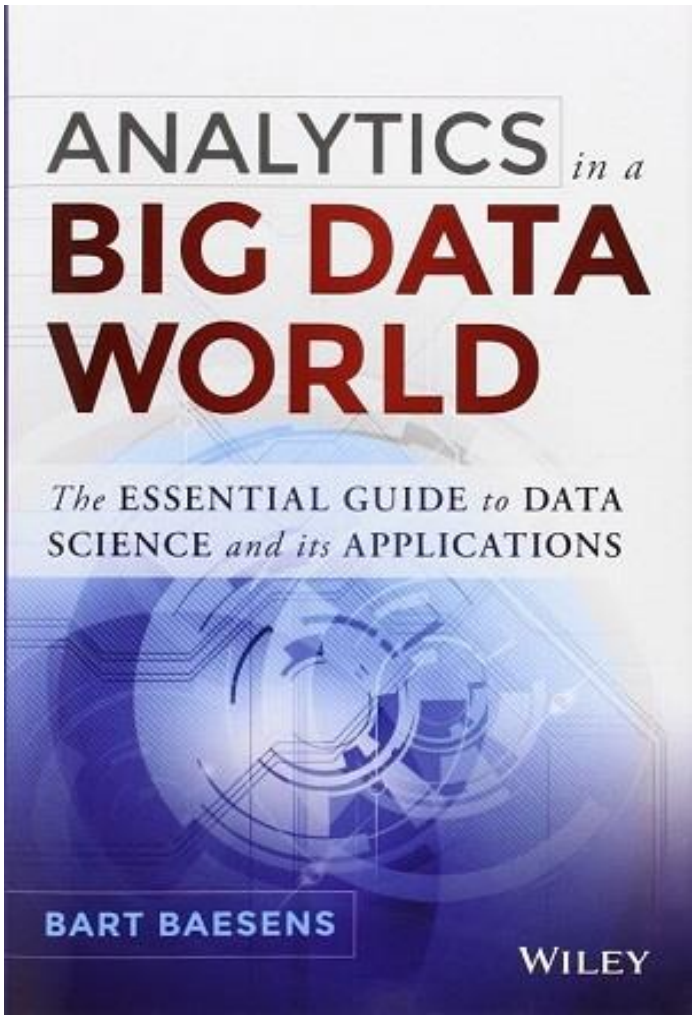
the incredible noise around the confusing and mislabeled term." — Forbes

With tips on how to develop a strategy and a plan of action regarding big data, what technology you need to embrace it and how to hire the right kinds of people to crunch big data, this book is clearly manager-oriented.

It also offers an overview of big data technologies, explains what is needed to succeed with big data, and gives examples of both successful and failed data practices undertaken by startups, online firms, and large companies. The author also introduces the concept of "analytics 3.0" to describe how companies can combine traditional analytics with a big data approach. He recognizes big online companies like Google or Facebook as the originators of best big data tools and technologies, as well as data-driven management reporting and best practices.

'Big Data at Work' is a pleasant read, however, this approachability may be a merit for some readers and a flaw for others. Critics point out that the book offers rather a breezy approach to the subject as it refrains from using technical language, thus it avoids answering some of the rudimentary questions.

9) Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, by B. Baesens



Best for: business data analysts, consultants, and graduate students in business analytics

An excerpt from a rave review:

"In a domain overwhelmed with hype and hyperboles, 'Analytics in a Big Data World' provides a no-nonsense, focused coverage on specifics and implementation best practices."

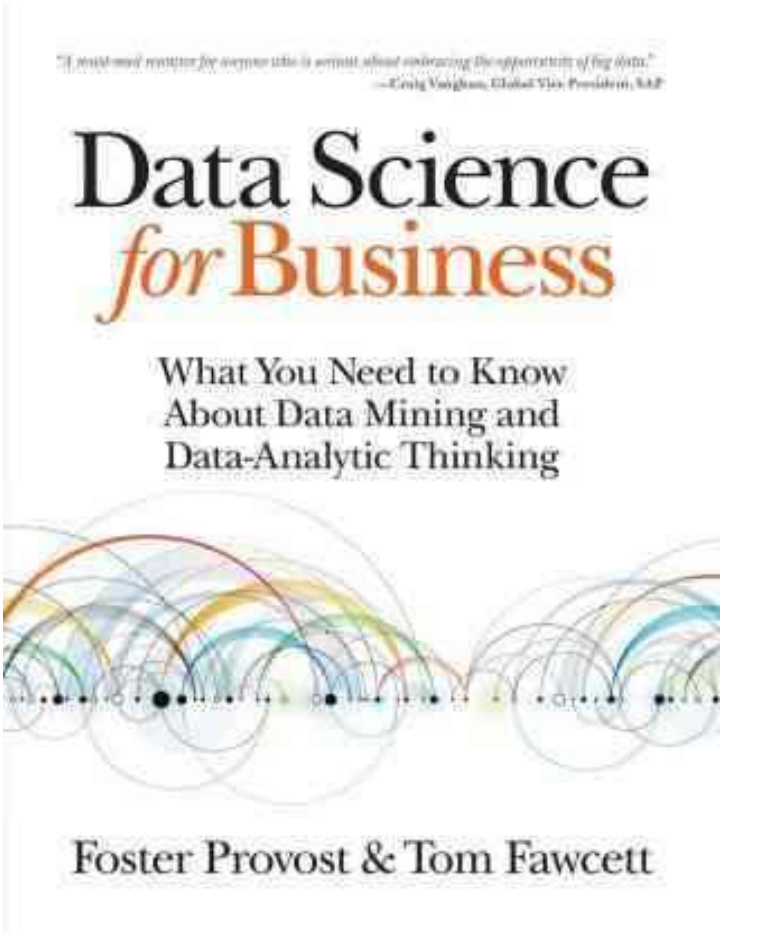
This is a real data analytics manual that would suit readers who already have the basic knowledge of data mining and business intelligence and are looking for structural and technical instructions on how to conduct big data analytics in real-world business management.

With a very strong practical focus “Analytics in a Big Data World” starts with providing the readers with the basic nomenclature, the analytics process model, and its relation to other relevant disciplines, such as statistics, machine learning, and artificial intelligence. The author then proceeds with highlighting the most important steps of the process model, such as sampling, treatment of missing values, and variable selection. The subsequent chapters focus on predictive and descriptive analytics.

Additionally, numerous case studies on risk management, fraud detection, customer relationship management, and web analytics are included and described in detail. In the seventh chapter, the author provides us with concrete instructions on which business analytics tools, and practices, to use to put analytics to work. Topics covered here range from backtesting and benchmarking approaches to data quality issues, software tools, and model documentation practices.

Designed to be an accessible resource, this essential big data book does not include an exhaustive coverage of all analytical techniques. Instead, it highlights data analytics techniques that really provide added value in business environments.

10) Data Science For Business: What You Need to Know About Data Mining & Data-Analytic Thinking, by F. Provost & T. Fawcett



Best for: someone who has read a few intro books on data science and is ready to challenge themselves and dive deeper

An excerpt from a rave review:

"The book strikes a satisfyingly good balance between technical fundamentals and business applications: just enough numbers and technical details for a solid foundation, complemented with

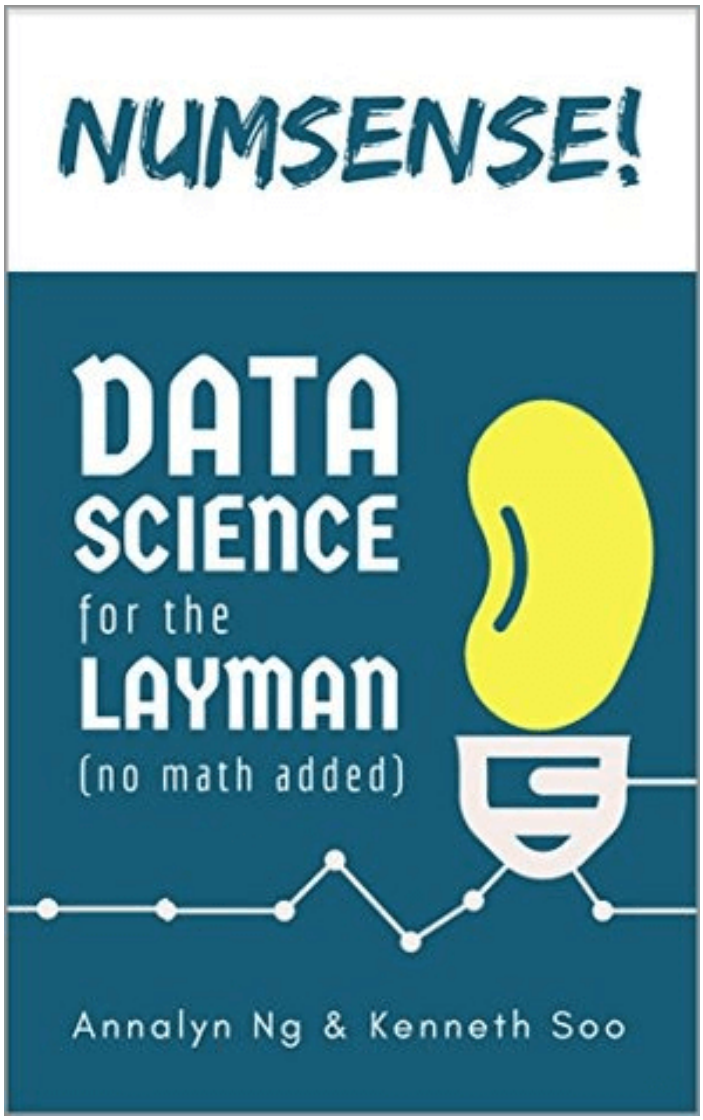
numerous business cases and examples to see how the tech stuffs fall into place."

Many books about data analytics and big data focus on the "how" of data science – the techniques and mechanisms. "Data Science for Business" does that as well, but also goes into the "why" of data science and provides insight into some useful ways to think about data science in a business setting.

The book reviews some underlying principles of data analytics, and is a great read for an aspiring data-driven decision maker who wants to intelligently participate in using big data and analytics to improve their company's strategic and tactical choices.

Finally, "Data Science for Business" goes into just enough detail explaining the data mining techniques used today, using plenty of scientific thinking without overwhelming the reader with numbers and equations. This is facilitated by the use of technical sections which the reader can choose to skip or devour according to their interest.

11) Numsense! Data Science for the Layman: No Math Added by Annalyn Ng & Kenneth Soo



Best for: Any lay person with no prior background in math or analytics, who wants to work in this field or to manage other data scientists

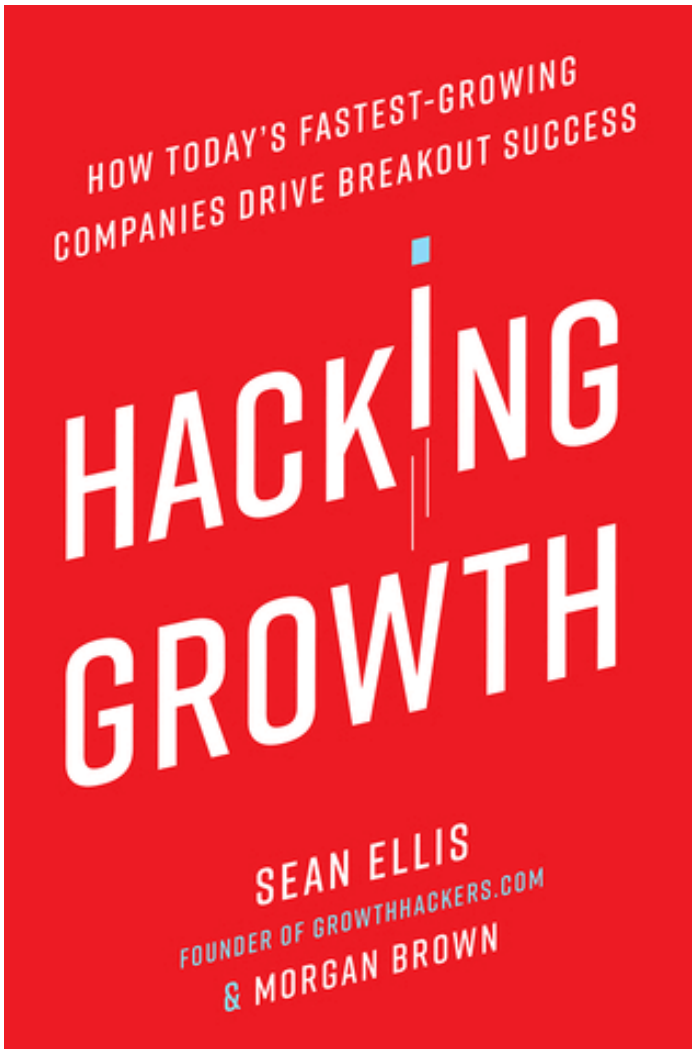
An excerpt from a rave review:

“Numsense! Data Science for the Layman is a great little book. Not only could it be a fine introduction for someone with little if any knowledge of data science, but it also provides nice summaries of several different areas for those with familiarity. Five stars for doing what the title says.”

For big data books geared toward the practical application of digital insights, Numsense! is one of the best on the market. Not only does this digestible guide speak to the reader in a clear, decipherable language, but it is also rich in actionable tips in areas including A/B testing, social network analysis, regression analytics, clustering, and more.

Boasting inspiring real-world examples and a comprehensive glossary of terms, this data analysis book is a must-read for anyone looking to embark on a lifelong journey toward analytical enlightenment.

12) Hacking Growth: How Today's Fastest-Growing Companies Drive Breakout Success by Sean Ellis & Morgan Brown



Best for: a budding startup entrepreneur looking to grow and evolve their empire by leveraging the power of big data

An excerpt from a rave review:

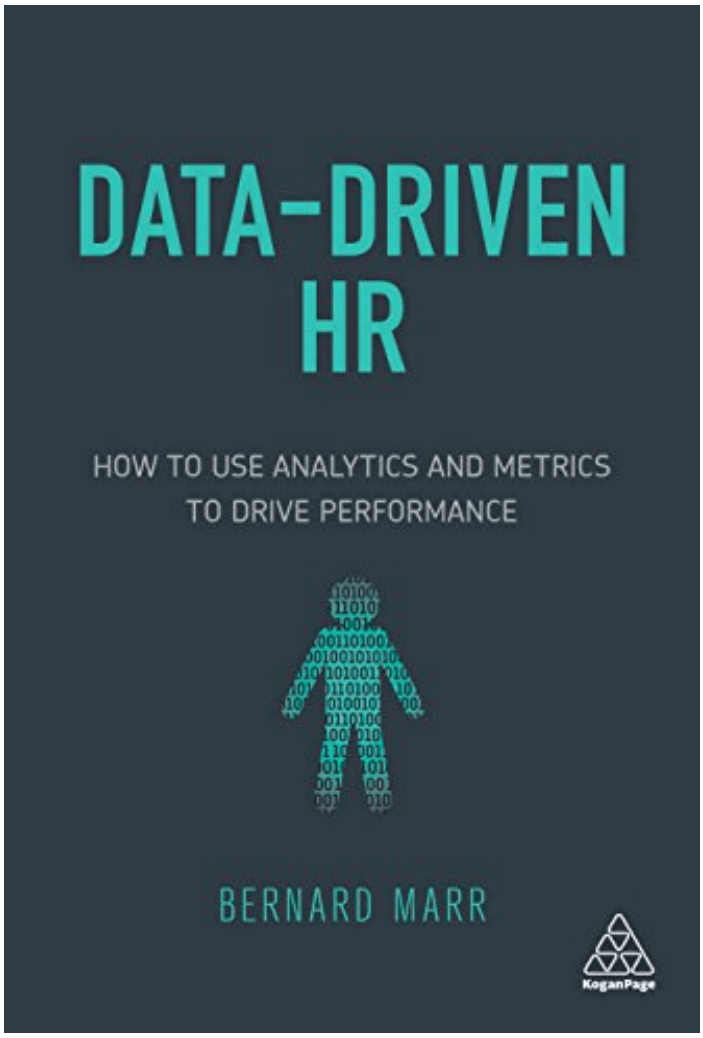
"Must read book for anyone interested in the subject matter. The author(s) lay out a very thorough yet concise picture of what growth

hacking involves and a step by step method on how to do it. They convincingly show that growth hacking methods or mindset can and should apply for you whether you work for a startup or a large company."

Growth hacking is a relatively new phenomenon, bestowing the term of using key insights, data, and digital strategies to connect with your target audience on a more meaningful, more personal level. And if executed the right way, it works.

Of all the growth hacking-themed books available today, this is the most inspiring, the most understandable and ultimately, the most rewarding. Not only will you gain tangible insight into how brands like Airbnb and Pinterest became global sensations, but you'll also gain access to a toolkit for growth hacking based on informed data-driven decisions.

13) Data-Driven HR: How to Use Analytics and Metrics to Drive Performance by Bernard Marr



Best for: business leaders, executives, and HR directors looking to improve their business practices through real-world big data analytics

An excerpt from a rave review:

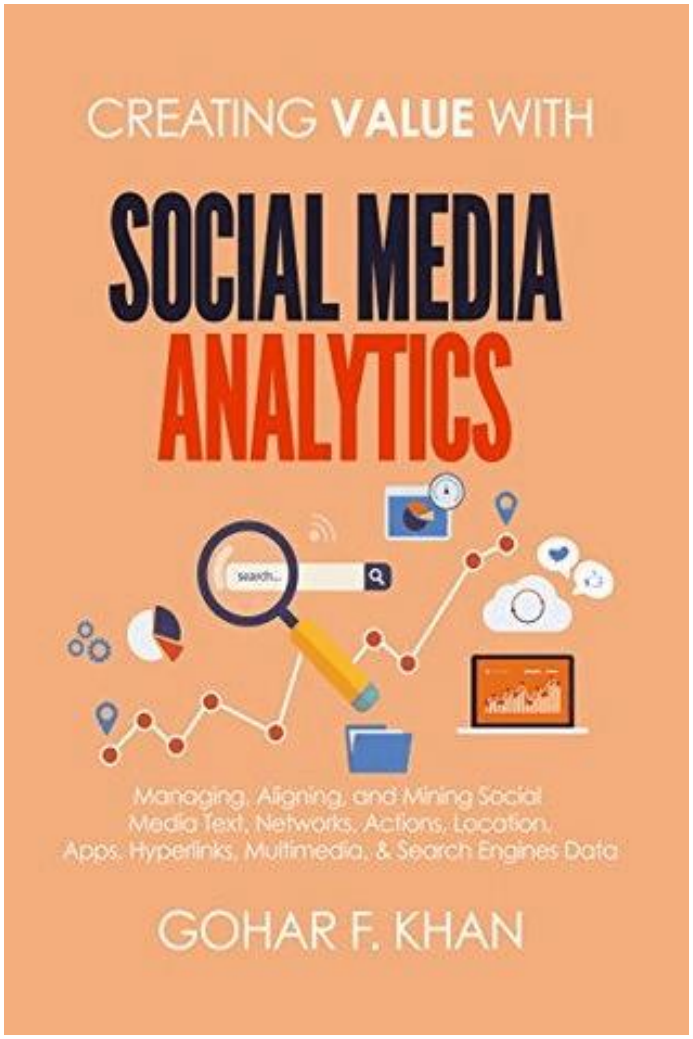
“A practical, inspirational guide for human resources (HR) professionals, this book sets out to show how data collected by a company can be utilised to improve the HR function and, in turn, the company-as-a-whole. It is a lot more than just storing HR information on a computer, yet many professionals have yet to see the utility and potential of data-driven HR. A book like this can change that!”

In today’s world, data serves to enhance the productivity, output, and efficiency of all sectors, disciplines and departments – and human resources is no exception.

For the HR professional looking to establish detailed HR KPIs, leverage the value of digital metrics and insights to improve areas such as training & development, data protection, staff management, and organizational efficiency, this is one of the best books on digital data you could ever read.

Crammed with practical insights and easy-to-follow case studies, this HR-based big data bible will serve as an invaluable reference in your quest for human resources perfection.

14) Creating Value With Social Media Analytics: Managing, Aligning, and Mining Social Media Text, Networks, Actions, Location, Apps, Hyperlinks, Multimedia, & Search Engines Data by Gohar F. Khan



Best for: any individual looking to get under the skin of data-based insights and metrics through renowned social media platforms

An excerpt from a rave review:

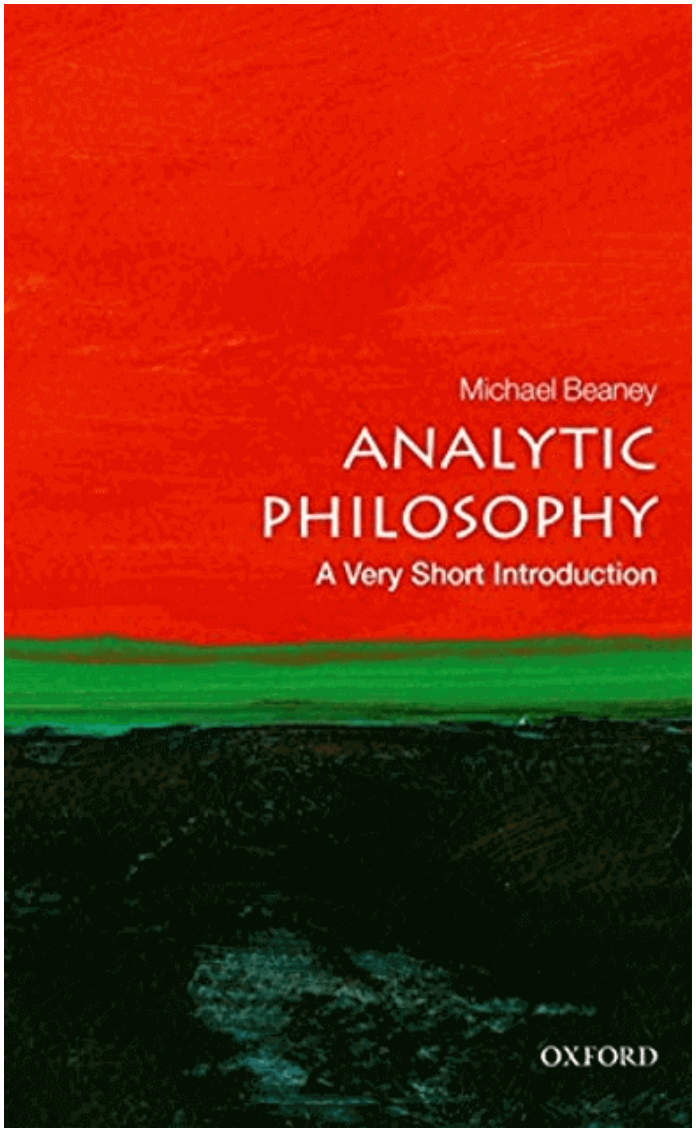
""Gohar Khan is a pioneer in the emerging domain of social media analytics. This latest text is a must-read for business leaders, managers, and academicians, as it provides a clear and concise understanding of business value creation through social media data from a social lens." -Laeeq Khan, Director, Social Media Analytics Research Team, Ohio University.

If you have a solid working grasp on the functionality of the world's most prominent social media platforms and digital marketing KPIs, but you'd like to squeeze more value from each channel, this big data book is a must-read.

Not only is the author's knowledge on the subject vast and deeply impressive, but it is also presented in such a way that budding data scientists, digital marketers, social media executives, and business leaders can extract priceless nuggets of information with ease.

By using big data analytics to refine and drive your social media strategy, you stand to set yourself apart from the competition – and this big data book will help you do just that.

15) Analytic Philosophy: A Very Short Introduction by Michael Beaney



Best for: individuals looking to understand the history, origins, and core philosophies of the analytical, data-driven mindset

An excerpt from a rave review:

"A concise, delightfully accessible, and intellectually stimulating introduction to philosophy in the analytic tradition, especially its

formative phase." – Erich Reck, Professor, University of California at Riverside

One of the most prolific data analysis books in existence, this insightful, informative, and refreshing work of prose serves as the ideal supplement to the more practical books and toolkits on our list.

Digging deep into the very ideation of the subject and the premise behind analytic thinking, this book defines precisely why big data analytics is so valuable while offering digestible concepts that will serve as the very foundations of everything you do with the digital insights available to you. A real must-read for anyone with a thirst for big data enlightenment.

*"The most valuable commodity I know of is information." – **Gordon Gekko**, **Wall Street***

Exclusive Bonus Content: Not Sure Which Data Analysis Book To Read?

Download our free guide on top 15 best books on data analysis!

If you found our list of the best data analytics and big data books useful, but your hunger for knowledge hasn't been satisfied yet, take a look at our best business intelligence books or our data visualization books post to keep growing in your understanding of data science. And if you'd like to put your newfound knowledge of big data analytics into practice, explore our online dashboard tool.

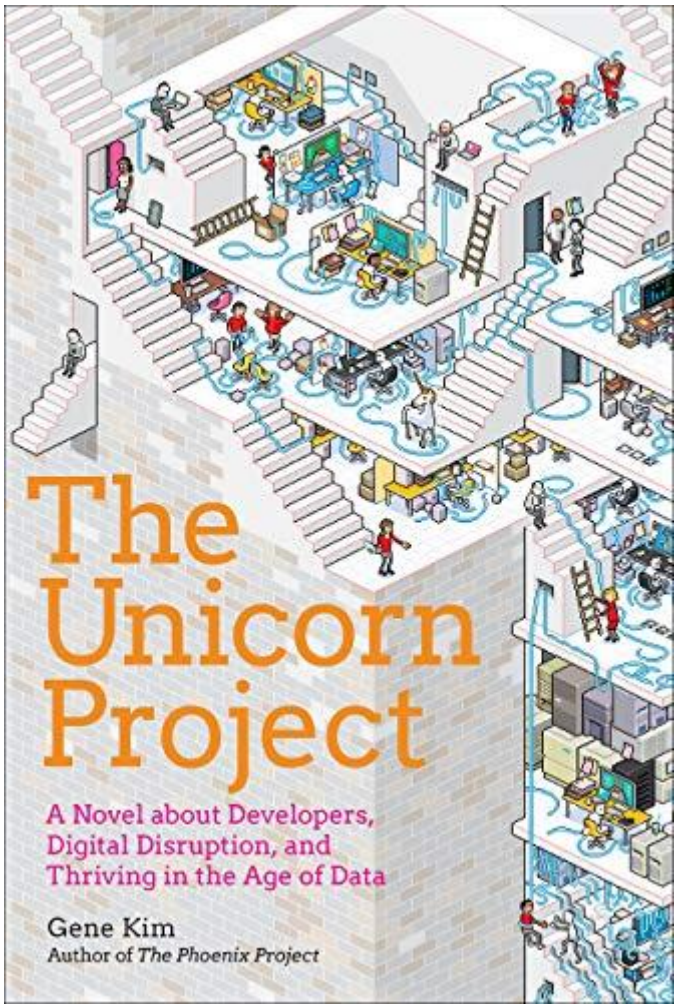
So, what are the best big data books? Here is a summary:

1. *Data Analytics Made Accessible*, by A. Maheshwari
2. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* by E. Siegel
3. *Too Big to Ignore: The Business Case for Big Data*, by award-winning author P. Simon
4. *Lean Analytics: Use Data to Build a Better Startup Faster*, by A. Croll and B. Yoskovitz
5. *Data Smart: Using Data Science to Transform Information into Insight*, by J. W. Foreman
6. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* by V. Mayer-Schönberger and K. Cukier
7. *Business UnIntelligence: Insight and Innovation Beyond Analytics and Big Data*, by B. Devlin
8. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*, by T. H. Davenport
9. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, by B. Baesens
10. *Data Science For Business: What You Need to Know About Data Mining & Data-Analytic Thinking*, by F. Provost & T. Fawcett
11. *Numsense! Data Science for the Layman: No Math Added* by Annalyn Ng & Kenneth Soo
12. *Hacking Growth: How Today's Fastest-Growing Companies Drive Breakout Success* by Sean Ellis & Morgan Brown
13. *Data-Driven HR: How to Use Analytics and Metrics to Drive Performance* by Bernard Marr
14. *Creating Value With Social Media Analytics: Managing, Aligning, and Mining Social Media Text, Networks, Actions, Location, Apps, Hyperlinks, Multimedia, & Search Engines Data* by Gohar F. Khan
15. *Analytic Philosophy: A Very Short Introduction* by Michael Beaney

16 Best New Big Data Books To Read In 2020

As featured on CNN, Forbes and Inc – BookAuthority identifies and rates the best books in the world, based on public mentions, recommendations, ratings and sentiment

1



The Unicorn Project

A Novel about Developers, Digital Disruption, and Thriving in the Age of Data

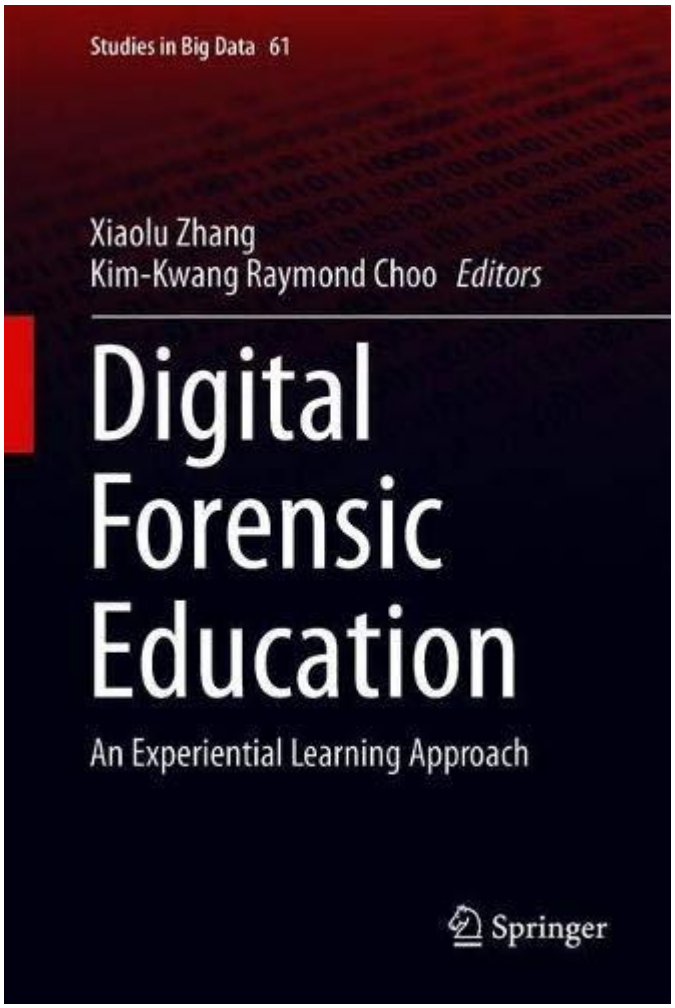
Gene Kim

|Nov 25, 2019|

4.25

With The Phoenix Project we learned about the Three Ways. Now with The Unicorn Project, Gene Kim introduces the Five Ideals. This highly anticipated follow-up to the bestselling title The Phoenix Project takes another look at Parts Unlimited, this time from the perspective of software development.
In The Unicorn Project, we follow Maxine, a senior lead developer and architect, as she is exiled to the Phoenix Project, to the horror of her friends and colleagues, as punishment for contributing to a payroll outage.

2



Digital Forensic Education

An Experiential Learning Approach

Xiaolu Zhang, Kim-Kwang Raymond Choo

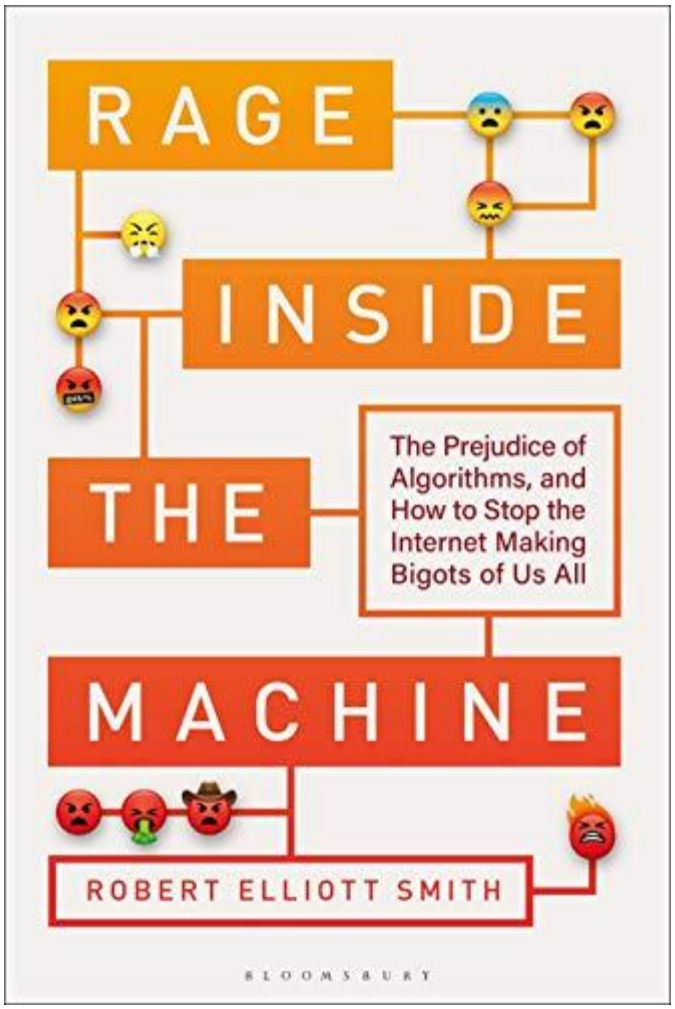
Oct 9, 2019

4.24

In this book, the editors explain how students enrolled in two digital forensic courses at their institution are exposed to experiential

learning opportunities, where the students acquire the knowledge and skills of the subject-matter while also learning how to adapt to the ever-changing digital forensic landscape. Their findings (e.g., forensic examination of different IoT devices) are also presented in the book. Digital forensics is a topic of increasing importance as our society becomes “smarter” with more of the “things” around us been internet- and inter-connected (e.g.

3



Rage Inside the Machine

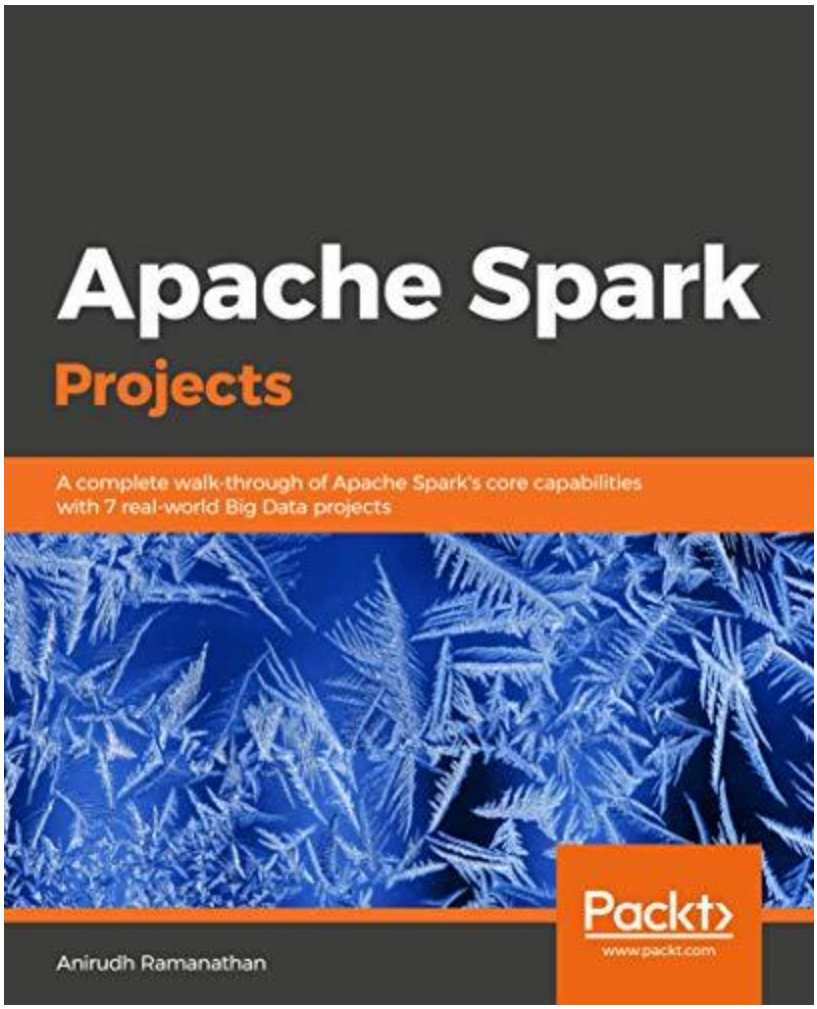
The Prejudice of Algorithms, and How to Stop the Internet Making Bigots of Us All

Robert Elliott Smith
|Aug 26, 2019|

4.21

An investigative journey into the sources of morality in artificial intelligence and how this impacts our society.
Offering a rigorous, fresh perspective on how technology has brought us to this place, Mechanisms of Morality challenges the long-held assumption that technology is an apolitical and amoral force. Having worked in the field of artificial intelligence for over 30 years, Smith reveals the mounting evidence that the mechanical actors in our lives have, or at least express, morals: they're just not the morals of the progressive modern society that we imagined we were moving towards.

4



Apache Spark Projects

A complete walk-through of Apache Spark's core capabilities with 7 real-world Big Data projects

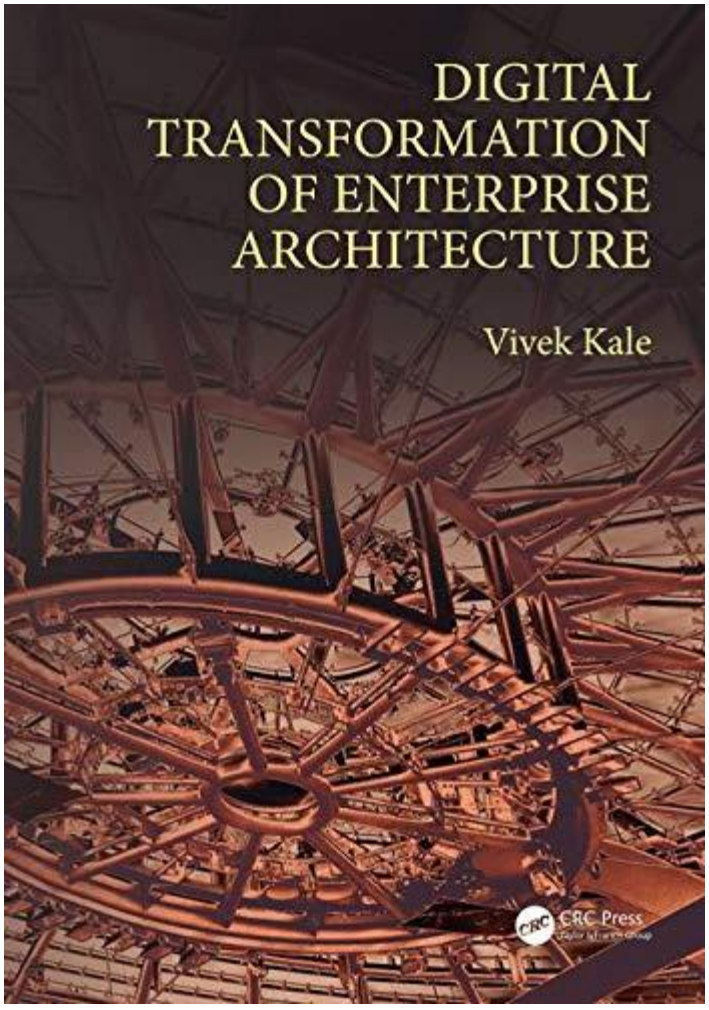
Anirudh Ramanathan

Aug 8, 2019

4.24

Explore the potential of Apache Spark and its ecosystem through real-world applications.
Key Features
A unique, practical guide with 7 end to end projects demonstrating the power of Apache Spark
Shows the readers how to perform real-time Big Data processing using different components of Apache Spark.

5



Digital Transformation of Enterprise Architecture

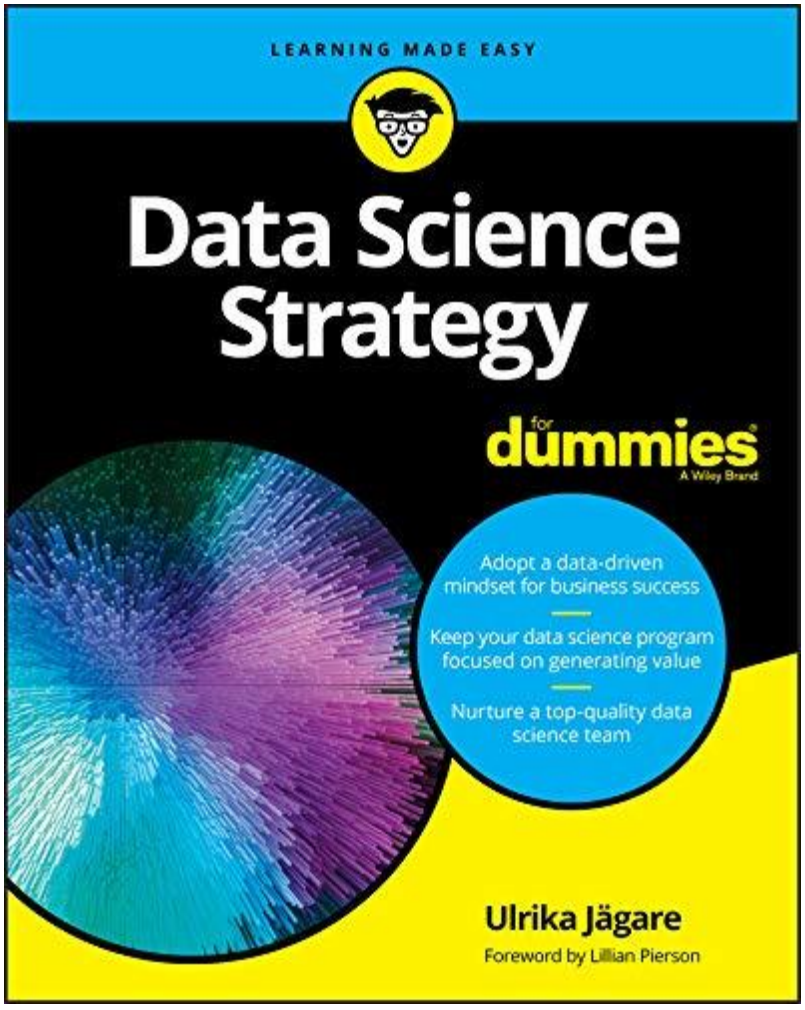
Vivek Kale

Jul 24, 2019

4.24

Digital Transformation of Enterprise Architecture is the first book to propose Enterprise Architecture (EA) as the most important element (after Business Models) for digital transformation of enterprises. This book makes digital transformation more tangible by showing the rationale and typical technologies associated with it, and, these technologies in turn reveal the essence of digital transformation This book proposes that it is the perennial quest for interoperability & portability, scalability, availability etc. that has directed and driven the evolution of the IT/IS industry in the past 50 years.

6



Data Science Strategy For Dummies)

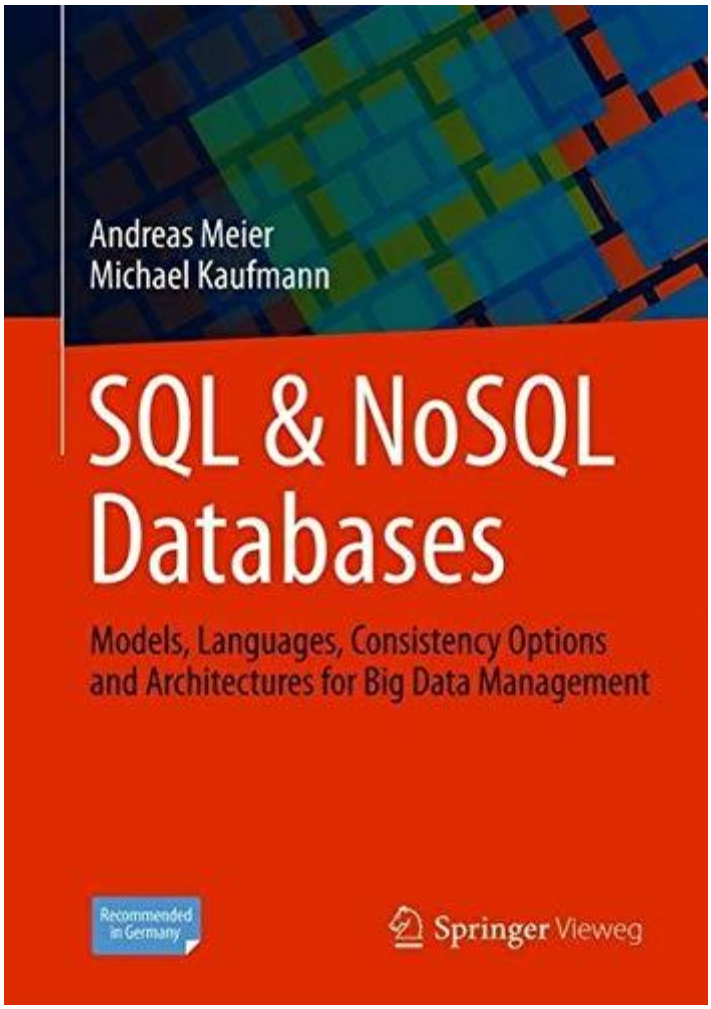
Ulrika Jägare

|Jul 10, 2019|

4.23

All the answers to your data science questions
Over half of all businesses are using data science to generate insights and value from big data. How are they doing it? Data Science Strategy For Dummies answers all your questions about how to build a data science capability from scratch, starting with the “what” and the “why” of data science and covering what it takes to lead and nurture a top-notch team of data scientists
With this book, you’ll learn how to incorporate data science as a strategic function into any business, large or small.

7



SQL & NoSQL Databases

Models, Languages, Consistency Options and Architectures for Big Data Management

Andreas Meier, **Michael Kaufmann**
|Jun 23, 2019|

4.24

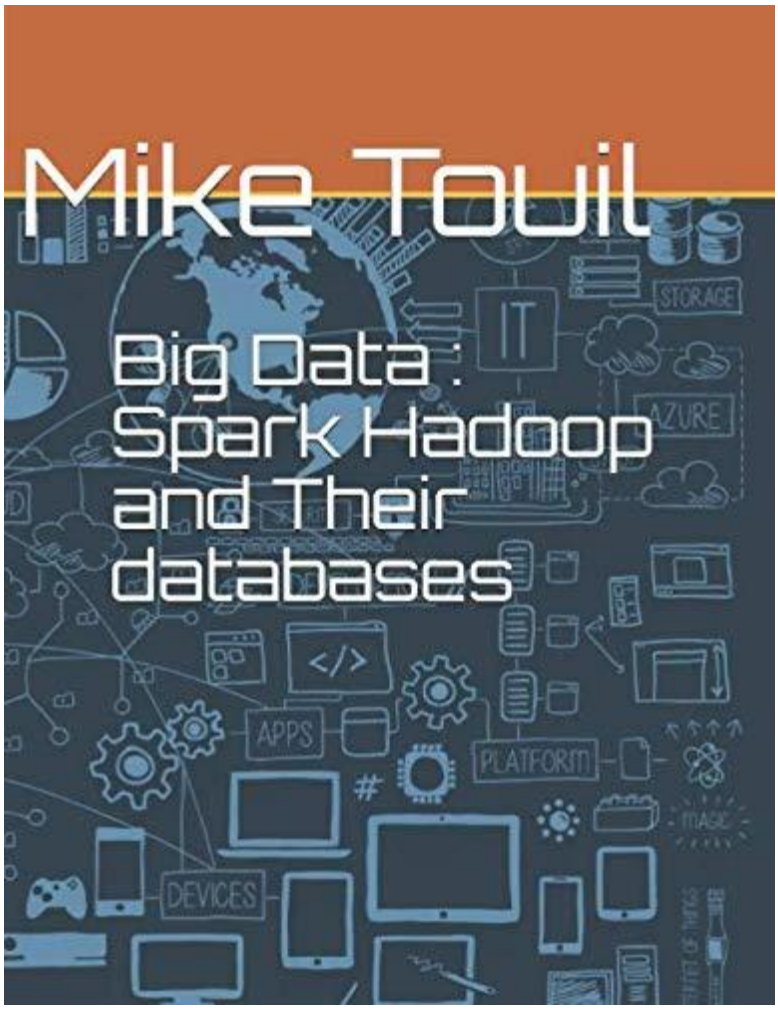
This book offers a comprehensive introduction to relational (SQL) and non-relational (NoSQL) databases. The authors thoroughly review the current state of database tools and techniques, and examine coming innovations

The book opens with a broad look at data management, including an overview of information systems and databases, and an explanation of contemporary database types:

SQL and NoSQL databases, and their respective management systems

The nature and uses of Big Data.

8



Big Data

Spark Hadoop and Their databases

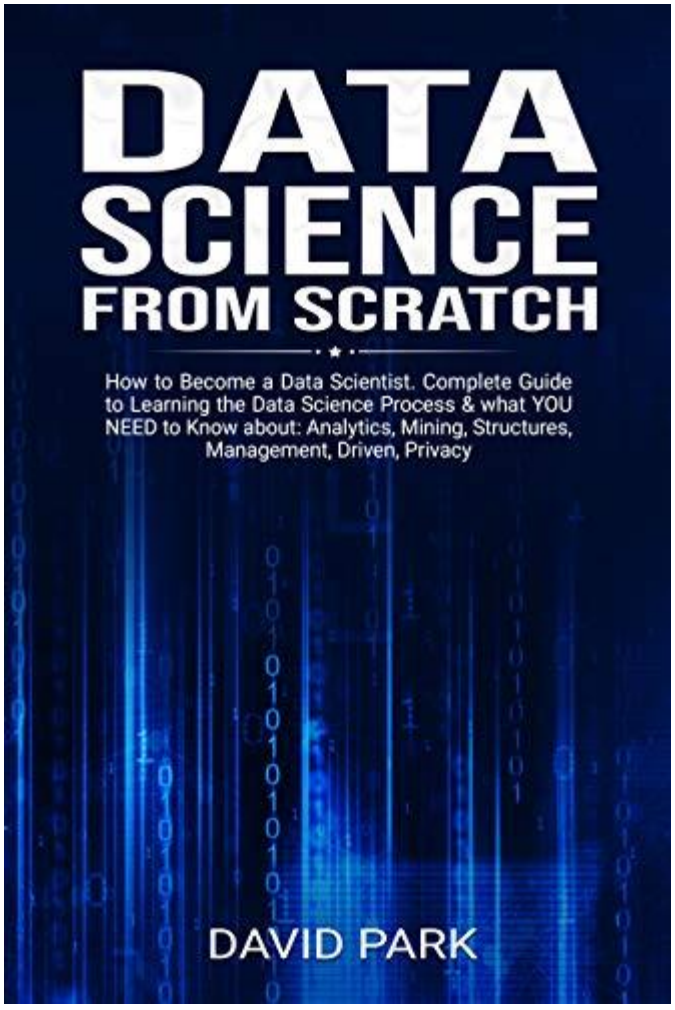
Mike Touil

Jun 3, 2019

4.24

Big data is a field in which we are very envolved nowadays, this is the ideal book to learn more about Hadoop Spark their components and also the databases they come up with.

9



Data Science From Scratch

How to Become a Data Scientist. Complete Guide to Learning the Data Science Process & what YOU NEED to Know about: Analytics, Mining, Structures, Management, Driven, Privacy

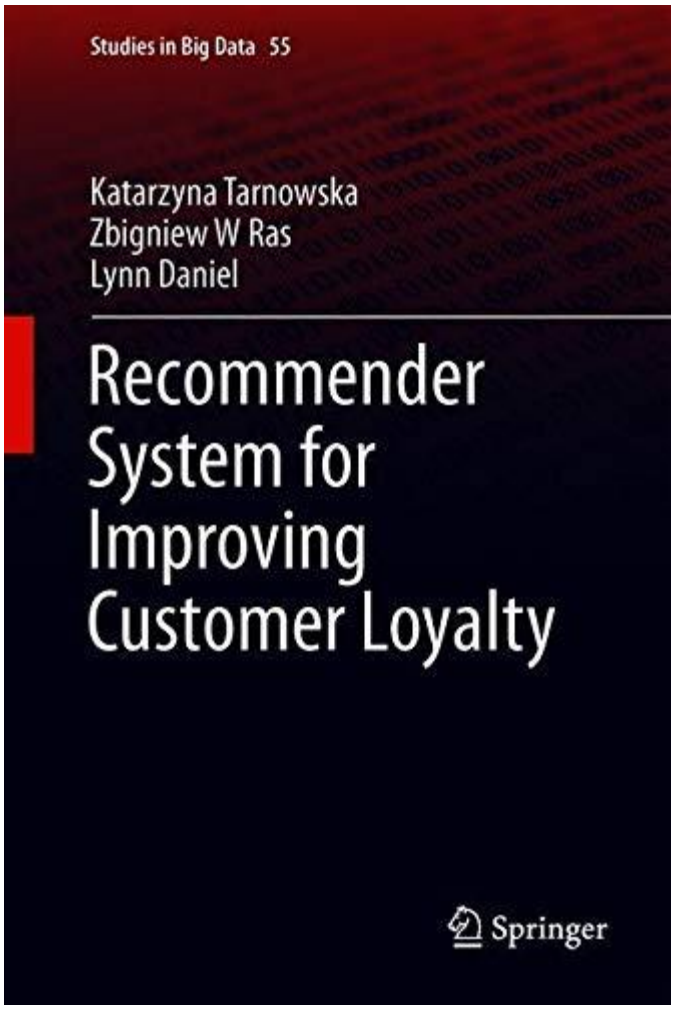
David Park

May 20, 2019

4.59

Data science is the application of a combination of mathematical, statistical, analytical and programming skills for the collection, organization, and interpretation of data to allow effective and proper management of the business whose data it is
The job of such a scientist is trending all over the world. The demand for such scientists is huge, more than the number of available candidates. A recent report explained that the need for these scientists has increased by more than 50% since last year. These scientists often referred to as big data wranglers, are a perfect blend of mathematician and computer scientist.

10


Studies in Big Data 55
Katarzyna Tarnowska
Zbigniew W Ras
Lynn Daniel
Recommender System for Improving Customer Loyalty
Springer

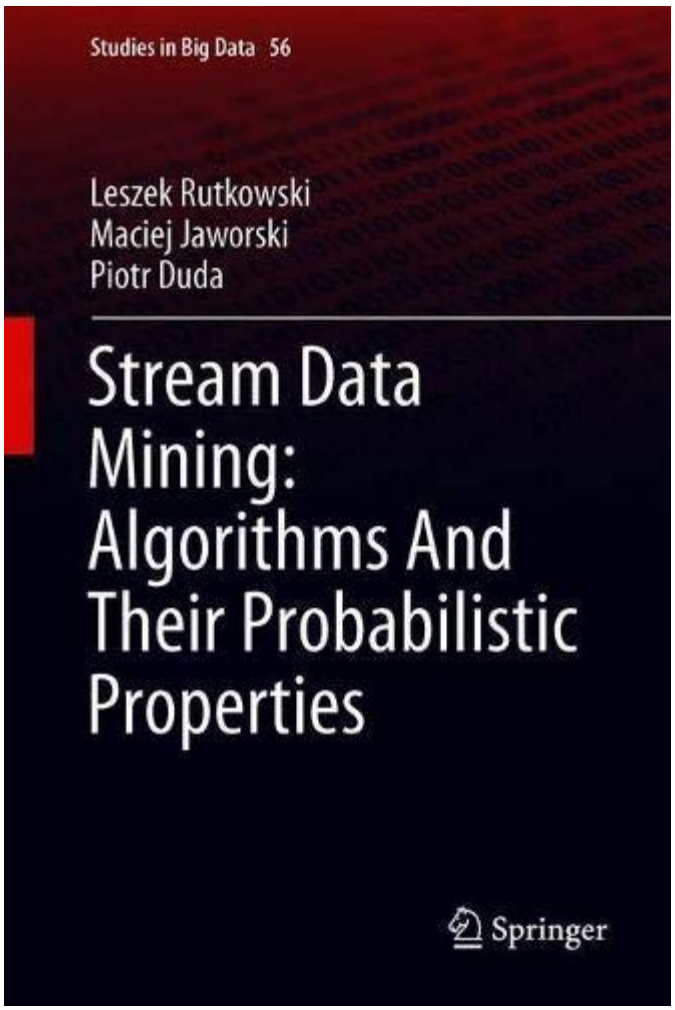
Recommender System for Improving Customer Loyalty

Katarzyna Tarnowska, Zbigniew W Ras, Lynn Daniel
May 20, 2019

4.24

This book presents the Recommender System for Improving Customer Loyalty. New and innovative products have begun appearing from a wide variety of countries, which has increased the need to improve the customer experience. When a customer spends hundreds of thousands of dollars on a piece of equipment, keeping it running efficiently is critical to achieving the desired return on investment. Moreover, managers have discovered that delivering a better customer experience pays off in a number of ways.

11


Studies in Big Data 56
Leszek Rutkowski
Maciej Jaworski
Piotr Duda
Stream Data Mining: Algorithms And Their Probabilistic Properties
Springer

Stream Data Mining

Algorithms And Their Probabilistic Properties

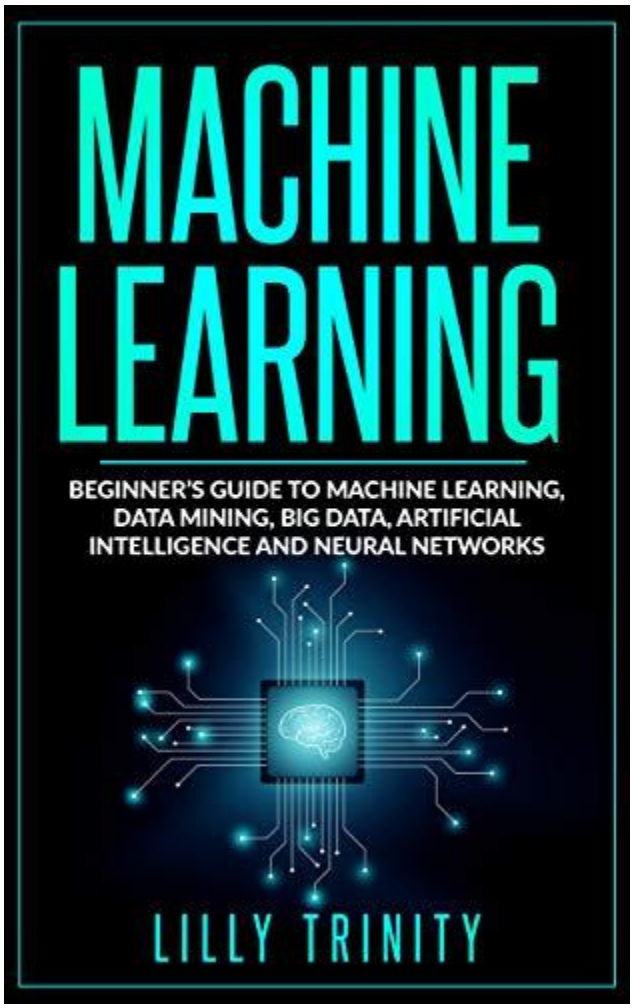
Leszek Rutkowski, Maciej Jaworski, Piotr Duda (You?)

May 16, 2019

4.24

This book presents a unique approach to stream data mining. Unlike the vast majority of previous approaches, which are largely based on heuristics, it highlights methods and algorithms that are mathematically justified. First, it describes how to adapt static decision trees to accommodate data streams; in this regard, new splitting criteria are developed to guarantee that they are asymptotically equivalent to the classical batch tree. Moreover, new decision trees are designed, leading to the original concept of hybrid trees.

12



Machine Learning

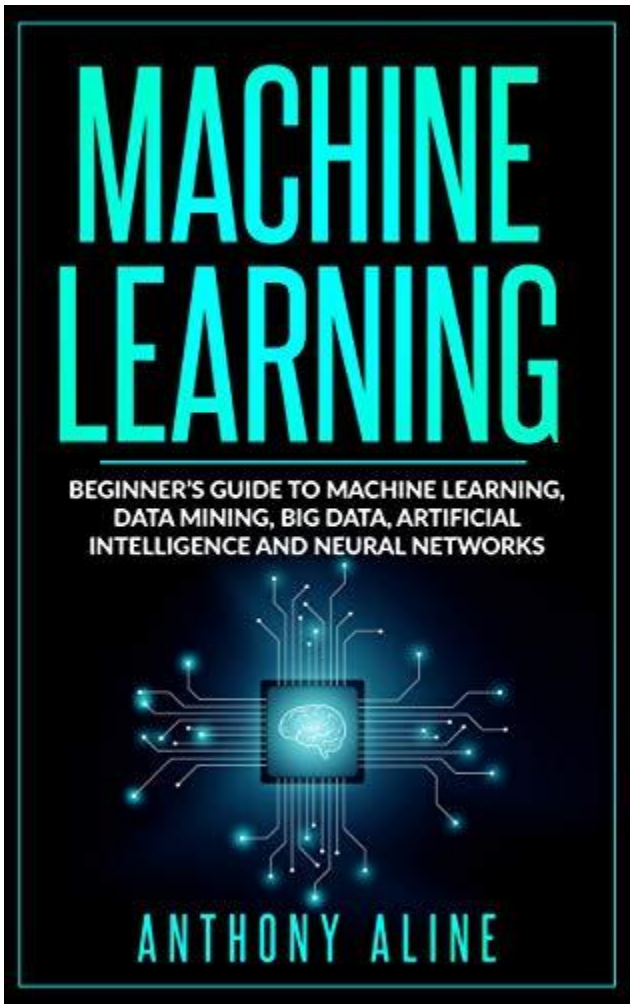
Beginner’s Guide to Machine Learning, Data Mining, Big Data, Artificial Intelligence and Neural Networks

Lilly Trinity

May 11, 2019

4.7

If you’re looking for a way to become an expert Robot and impress your friends with the programs you can make from scratch, then keep reading..
Artificial Intelligence, and in particular, Machine Learning is here today and it is shaping our world. It is shaping and simplifying the way we live, work, travel and communicate. Hence it is important for you to understand what it is and how it works.
This guide has been designed to help you gain an understanding of machine learning, artificial intelligence and big data in a simple way.



Machine Learning

Beginner’s Guide to Machine Learning, Deep Learning, Data Mining, Big Data, Artificial Intelligence and Neural Networks

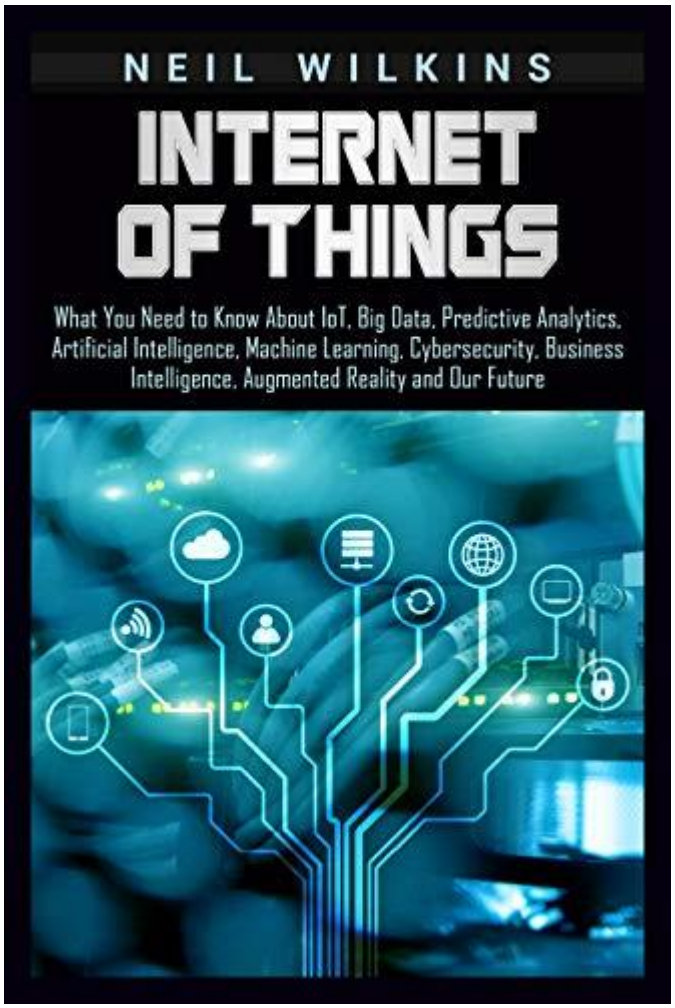
Anthony Aline)
|Mar 28, 2019|

4.67

2522

Artificial Intelligence, and in particular, Machine Learning is here today and it is shaping our world. It is shaping and simplifying the way we live, work, travel and communicate. Hence it is important for you to understand what it is and how it works

This guide has been designed to help you gain an understanding of machine learning, artificial intelligence and big data in a simple way We will walk you step-by-step into the world of Machine Learning.



Internet of Things

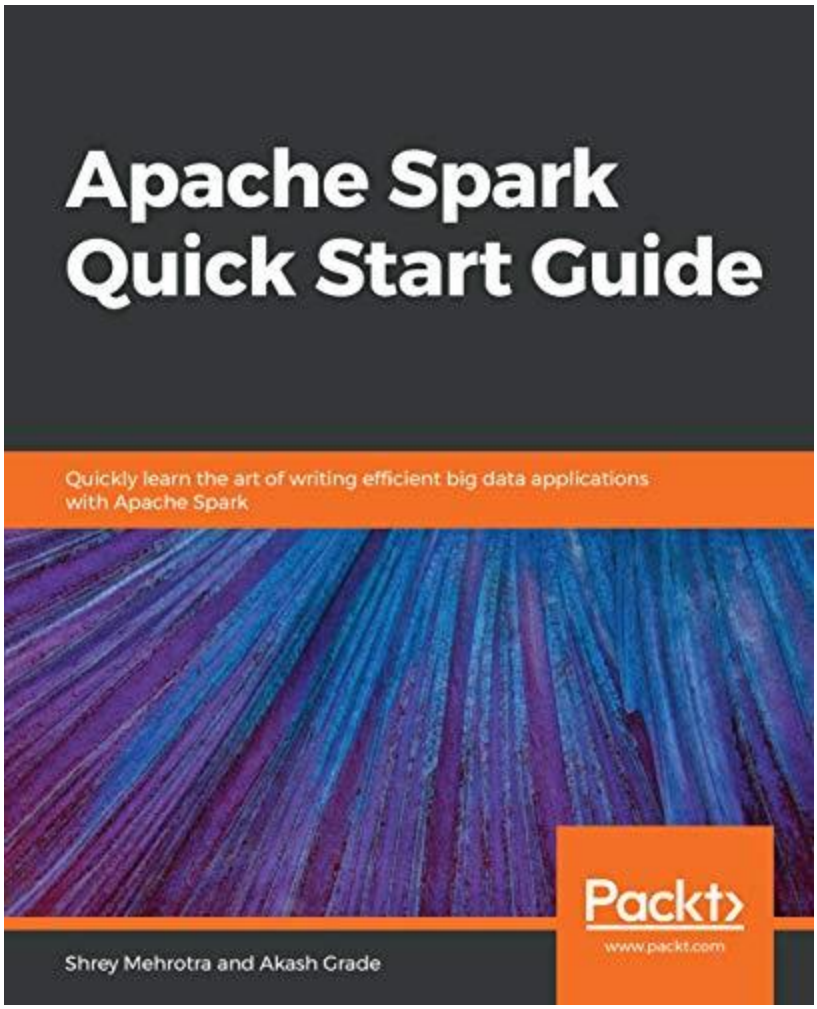
What You Need to Know About IoT, Big Data, Predictive Analytics, Artificial Intelligence, Machine Learning, Cybersecurity, Business Intelligence, Augmented Reality and Our Future

Neil Wilkins (You?)
|Mar 7, 2019|

4.57

If you want to learn about the Internet of Things, then keep reading…
You were just woken up in the middle of the night by smart lightbulbs in your house blasting at full power for no reason. Your bleary-eyed investigation shows they tried to download a firmware update and failed. At that moment, Alexa starts quietly whispering sweet nonsense to herself in the corner and Roomba starts slamming into the nearest wall.
What do you do?
Is your house haunted or have the machines finally started an uprising? Neither – it’s just another day in the IoT wonderland.

15



Apache Spark Quick Start Guide

Quickly learn the art of writing efficient big data applications with Apache Spark

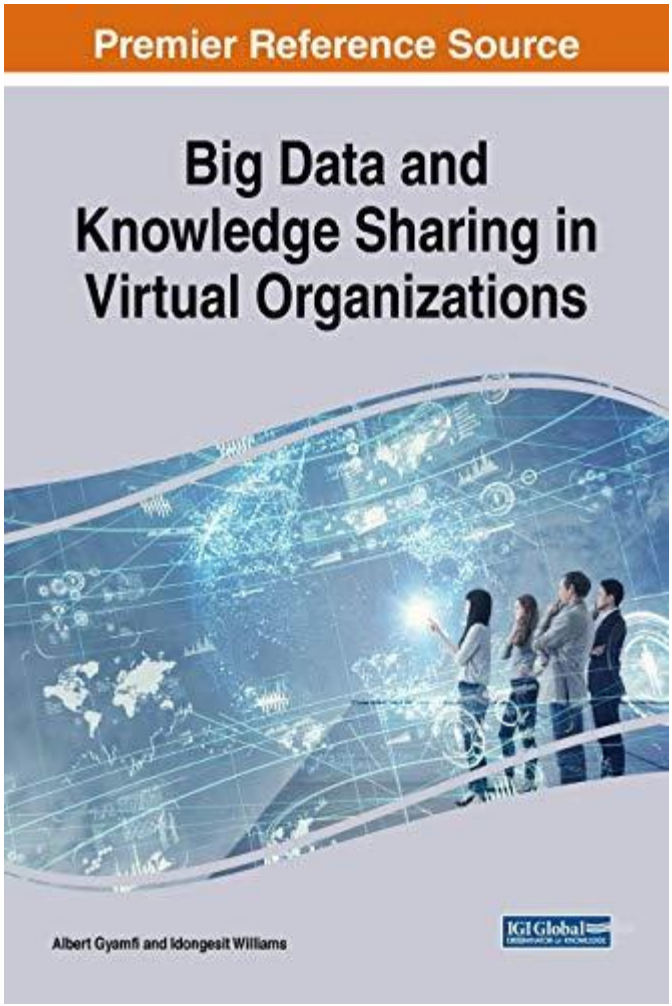
Shrey Mehrotra, **Akash Grade**

|Jan 30, 2019|

4.24

A practical guide for solving complex data processing challenges by applying the best optimizations techniques in Apache Spark.
Key Features
Learn about the core concepts and the latest developments in Apache Spark
Master writing efficient big data applications with Spark's built-in modules for SQL, Streaming, Machine Learning and Graph analysis.

16



Big Data and Knowledge Sharing in Virtual Organizations

Albert Gyamfi

|Dec 31, 2018|

4.24

Knowledge in its pure state is tacit in naturedifficult to formalize and communicatebut can be converted into codified form and shared through both social interactions and the use of IT-based applications and systems. Even though there seems to be considerable synergies between the resulting huge data and the convertible knowledge, there is still a debate on how the increasing amount of data captured by corporations could improve decision making and foster innovation through effective knowledge-sharing practices.

Popular Big Data Books

Big Data: A Revolution That Will Transform How We Live, Work, and Think (Hardcover)

by

Viktor Mayer-Schönberger

(shelved 45 times as *big-data*)

avg rating 3.71 — 7,090 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are (Hardcover)

by

Seth Stephens-Davidowitz (Goodreads Author)

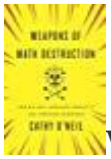
(shelved 34 times as *big-data*)

avg rating 3.92 — 22,887 ratings — published 2017

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Hardcover)

by

Cathy O'Neil

(shelved 31 times as *big-data*)
avg rating 3.87 — 12,298 ratings — published 2016

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Learning Spark: Lightning-Fast Big Data Analysis (Kindle Edition)

by

Holden Karau (Goodreads Author)

(shelved 23 times as *big-data*)
avg rating 3.92 — 466 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Dataclysm: Who We Are (When We Think No One's Looking)

by

Christian Rudder

(shelved 23 times as *big-data*)
avg rating 3.74 — 10,243 ratings — published 2014

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data: Principles and best practices of scalable realtime data systems (Paperback)

by

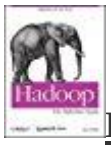
Nathan Marz

(shelved 23 times as *big-data*)
avg rating 3.81 — 398 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Hadoop: The Definitive Guide (Paperback)

by

Tom White (Goodreads Author)

(shelved 21 times as *big-data*)
avg rating 3.95 — 727 ratings — published 2009

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Designing Data-Intensive Applications (ebook)

by

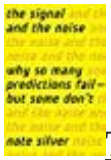
Martin Kleppmann (Goodreads Author)

(shelved 17 times as *big-data*)
avg rating 4.72 — 2,108 ratings — published 2015

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



The Signal and the Noise: Why So Many Predictions Fail - But Some Don't (Hardcover)

by

Nate Silver

(shelved 16 times as *big-data*)

avg rating 3.98 — 38,916 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems (Paperback)

by

Donald Miner

(shelved 16 times as *big-data*)

avg rating 3.87 — 89 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data at Work: Dispelling the Myths, Uncovering the Opportunities (Hardcover)

by

Thomas H. Davenport

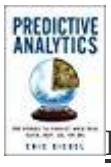
(shelved 12 times as *big-data*)

avg rating 3.63 — 363 ratings — published 2014

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (Hardcover)

by

Eric Siegel

(shelved 12 times as *big-data*)
avg rating 3.67 — 1,717 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

The Human Face of Big Data (Hardcover)

by

Rick Smolan (Goodreads Author) (Creator)

(shelved 11 times as *big-data*)
avg rating 4.00 — 203 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Data Science for Business: What you need to know about data mining and data-analytic thinking (Paperback)

by

Foster Provost

(shelved 11 times as *big-data*)
avg rating 4.16 — 1,474 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Advanced Analytics with Spark (Paperback)

by

Sandy Ryza

(shelved 10 times as *big-data*)
avg rating 3.97 — 108 ratings — published 2015

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Hadoop: The Definitive Guide (ebook)

by

Tom White

(shelved 10 times as *big-data*)
avg rating 3.91 — 197 ratings — published 2009

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work (Paperback)

by

Q. Ethan McCallum

(shelved 10 times as *big-data*)
avg rating 3.52 — 94 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Data Smart: Using Data Science to Transform Information into Insight (Paperback)

by

John W. Foreman

(shelved 9 times as *big-data*)
avg rating 4.15 — 814 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists (Paperback)

by

Philipp K. Janert

(shelved 9 times as *big-data*)
avg rating 4.06 — 290 ratings — published 2010

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale (Paperback)

by

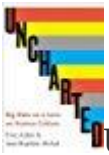
Neha Narkhede

(shelved 8 times as *big-data*)
avg rating 4.15 — 279 ratings — published

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Uncharted: Big Data and an Emerging Science of Human History (Hardcover)

by

Erez Aiden (Goodreads Author)

(shelved 8 times as *big-data*)
avg rating 3.73 — 609 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Doing Data Science (Paperback)

by

Rachel Schutt

(shelved 8 times as *big-data*)
avg rating 3.72 — 464 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Mining of Massive Datasets (Hardcover)

by

Anand Rajaraman

(shelved 8 times as *big-data*)
avg rating 4.35 — 201 ratings — published 2011

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Programming Hive (Paperback)

by

Edward Capriolo

(shelved 7 times as *big-data*)
avg rating 3.70 — 82 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark (Paperback)

by

Holden Karau (Goodreads Author)

(shelved 7 times as *big-data*)
avg rating 3.93 — 71 ratings — published

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

I Heart Logs: Event Data, Stream Processing, and Data Integration (Paperback)

by

Jay Kreps

(shelved 7 times as *big-data*)

avg rating 3.84 — 180 ratings — published 2014

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine (Paperback)

by

Clinton Gormley

(shelved 7 times as *big-data*)

avg rating 4.25 — 207 ratings — published 2014

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance (Kindle Edition)

by

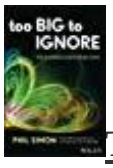
Bernard Marr

(shelved 7 times as *big-data*)
avg rating 3.61 — 277 ratings — published 2015

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Too Big to Ignore: The Business Case for Big Data (Hardcover)

by

Phil Simon (Goodreads Author)

(shelved 7 times as *big-data*)
avg rating 3.73 — 89 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Real-Time Big Data Analytics: Emerging Architecture (Kindle Edition)

by

Mike Barlow

(shelved 7 times as *big-data*)
avg rating 3.53 — 163 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data Now: Current Perspectives from O'Reilly Radar (ebook)

by

O'Reilly Radar Team

(shelved 7 times as *big-data*)
avg rating 3.35 — 254 ratings — published 2011

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Hadoop in Action (Paperback)

by

Chuck Lam

(shelved 7 times as *big-data*)
avg rating 3.67 — 113 ratings — published 2010

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Planning for Big Data: A CIO's Handbook to the Changing Data Landscape (ebook)

by

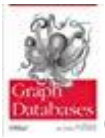
Edd Wilder-James (Goodreads Author)

(shelved 6 times as *big-data*)
avg rating 3.42 — 155 ratings — published 2004

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Graph Databases (Paperback)

by

Ian Robinson

(shelved 6 times as *big-data*)
avg rating 3.65 — 344 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Cassandra: The Definitive Guide (Paperback)

by

Eben Hewitt

(shelved 6 times as *big-data*)
avg rating 3.64 — 196 ratings — published 2010

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results (Kindle Edition)

by

Bernard Marr

(shelved 6 times as *big-data*)
avg rating 3.38 — 143 ratings — published 2016

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World (Paperback)

by

Bruce Schneier

(shelved 6 times as *big-data*)
avg rating 4.01 — 3,014 ratings — published 2015

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Storm Applied (ebook)

by

Sean T. Allen

(shelved 6 times as *big-data*)
avg rating 3.92 — 26 ratings — published 2015

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data for Dummies (Paperback)

by

Judith Hurwitz

(shelved 6 times as *big-data*)
avg rating 3.32 — 151 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Python for Data Analysis (Paperback)

by

Wes McKinney

(shelved 6 times as *big-data*)
avg rating 4.10 — 1,347 ratings — published 2011

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data Now: 2012 Edition (ebook)

by

O'Reilly Media

(shelved 6 times as *big-data*)
avg rating 3.39 — 89 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Hadoop Operations (ebook)

by

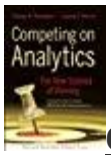
Eric Sammer

(shelved 6 times as *big-data*)
avg rating 3.90 — 83 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars



Competing on Analytics: The New Science of Winning (Hardcover)

by

Thomas H. Davenport

(shelved 5 times as *big-data*)
avg rating 3.64 — 1,032 ratings — published 2007

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Agile Data Science: Building Data Analytics Applications with Hadoop (Paperback)

by

Russell Jurney

(shelved 5 times as *big-data*)
avg rating 3.29 — 65 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Building Data Science Teams (Kindle Edition)

by

D.J. Patil

(shelved 5 times as *big-data*)
avg rating 3.67 — 278 ratings — published 2011

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Big Data Glossary (Paperback)

by

Pete Warden

(shelved 5 times as *big-data*)
avg rating 3.12 — 59 ratings — published 2011

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Hadoop Real-World Solutions Cookbook (ebook)

by

Jonathan R. Owens

(shelved 5 times as *big-data*)
avg rating 3.44 — 9 ratings — published 2013

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Hadoop in Practice (Paperback)

by

Alex Holmes

(shelved 5 times as *big-data*)
avg rating 3.87 — 38 ratings — published 2012

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

Event Streams in Action (Paperback)

by

Alexander Dean

(shelved 4 times as *big-data*)
avg rating 3.24 — 21 ratings — published

Want to Read

Rate this book

1 of 5 stars2 of 5 stars3 of 5 stars4 of 5 stars5 of 5 stars

The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World (Hardcover)

by

Pedro Domingos

(shelved 4 times as *big-data*)
avg rating 3.79 — 4,127 ratings — published 2015

Want to Read

Rate this book