Big Data Technology

Jnaneshwar Bohara

Chapter 1: Introduction to Big Data

- Big Data Overview
- Background of Data Analytics
- Role of Distributed System in Big Data
- Role of data Scientist
- Current Trend in Big Data Analytics

1.1 Big Data Overview

- Collection of data sets **so large and complex** that it becomes **difficult to process** using **on-hand database management tools** or **traditional data processing applications**.

Big Data

- Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.

From Bits to GeopBytes

1024 Bytes	1 Kilobyte
1024 Kilobytes	1 Megabyte
1024 Megabytes	1 Gigabyte
1024 Gigabytes	1 Terabyte
1024 Terabytes	1 Petabyte
1024 Petabytes	1 Exabyte
1024 Exabytes	1 Zettabyte
1024 Zettabytes	1 Yottabyte
1024 Yottabytes	1 Brontobyte
1024 Brontobytes	1 Geopbyte

One geopbyte is 1024^10 or1267650600228229401496703205376 bytes.
Or simply a 1 followed by 30 digits. (Not zeroes)

Big Data Statistics

- **In 2020**, there will be around 40 trillion gigabytes of data (**40 zettabytes**)
- **90%** of all data has been created in the **last two years**
- Today it would take a person approximately **181 million years to download** all the data from the internet
- Internet users generate about **2.5 quintillion bytes** of data each day.
- **In 2019**, internet users spent **1.2 billion years online.**

Big Data Statistics

- **Social media** accounts for **33% of the total time spent** online.
- **In 2019**, there are **2.3 billion active Facebook users**, and they generate a lot of data.
- **Twitter** users send over **half a million tweets every minute**.
- **97.2%** of organizations are investing in **big data and AI**.
- Using big data, **Netflix saves $1 billion per year** on customer retention.

Big Data Statistics

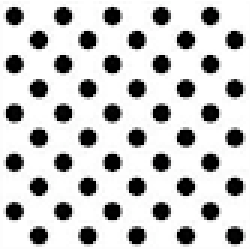
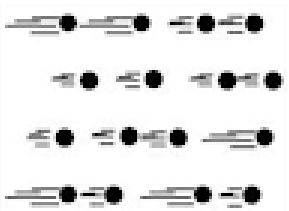
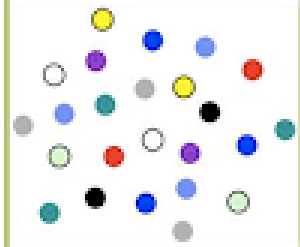
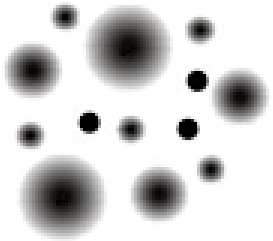
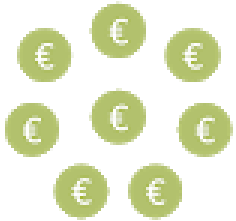
- **In 2020**, the big data market is expected to **grow by 14%**.
- **Job listings for data science** and analytics will reach around **2.7 million by 2020**.
- **By 2020**, every person will generate **1.7 megabytes** in just a **second**.
- **Automated analytics** will be vital to big data by **2020**.

https://techjury.net/stats-about/big-data-statistics/#gref

Characteristics of Big Data

- **Volume** - Data at rest (too big)
- **Variety -** Data in many forms (too complex)
- **Velocity -**Data in motion(too fast)
- **Veracity** - Data in doubt(uncertainty)
- **Value** - Data into money

Characteristics of Big Data

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,...	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

Volume: Scale of Data

- Refers to the vast amounts of data generated every second.
- We are not talking Terabytes but Brontobytes or Geopbytes.
- If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute.

Variety: Different Forms of Data

- This refers to the different types of data we can now use.
- In the past we focused on structured data that fits neatly into tables or relational databases, such as financial data.
- In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.)
- Big data technology means we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

Velocity: Analysis of Streaming Data

- Refers to the speed at which new data is generated and the speed at which data moves around.
- Just think of social media messages going viral in seconds.
- Technology allows us now to analyse the data while it is being generated (in-memory analytics), without ever putting it into databases.

Veracity: Uncertainty of Data

- Refers to the messiness or trustworthiness of the data.
- With many forms of big data, quality and accuracy are less controllable
- Big data and analytics technology now allows us to work with these type of data.

Value: Turning Big Data into Value

- Having access to big data is no good unless we can turn it into value.
- Companies are starting to generate amazing value from their big data.

1.2 Background of Data Analytics

- Big data analytics is the process of examining large amounts of data of a variety of types.
- The primary goal of big data analytics is to help companies make better business decisions.
- Analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs.

Data Analytics

- Big data Consist of

  - Uncovered hidden patterns.
  - Unknown correlations and other useful information.
  - Such information can provide business benefits.
  - More effective marketing and increased revenue.

Data Analytics

- Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as **predictive analysis** and **data mining.**
- But the unstructured data sources used for big data analytics may not fit in traditional data warehouses.
- Traditional data warehouses may not be able to handle the processing demands posed by big data.

Data Analytics

- The technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce.
- Knowledge about these technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems.
- Big Data analytics initiatives include
- internal data analytics skills
- high cost of hiring experienced analytics professionals,
- challenges in integrating Hadoop systems and data warehouses

Data Analytics

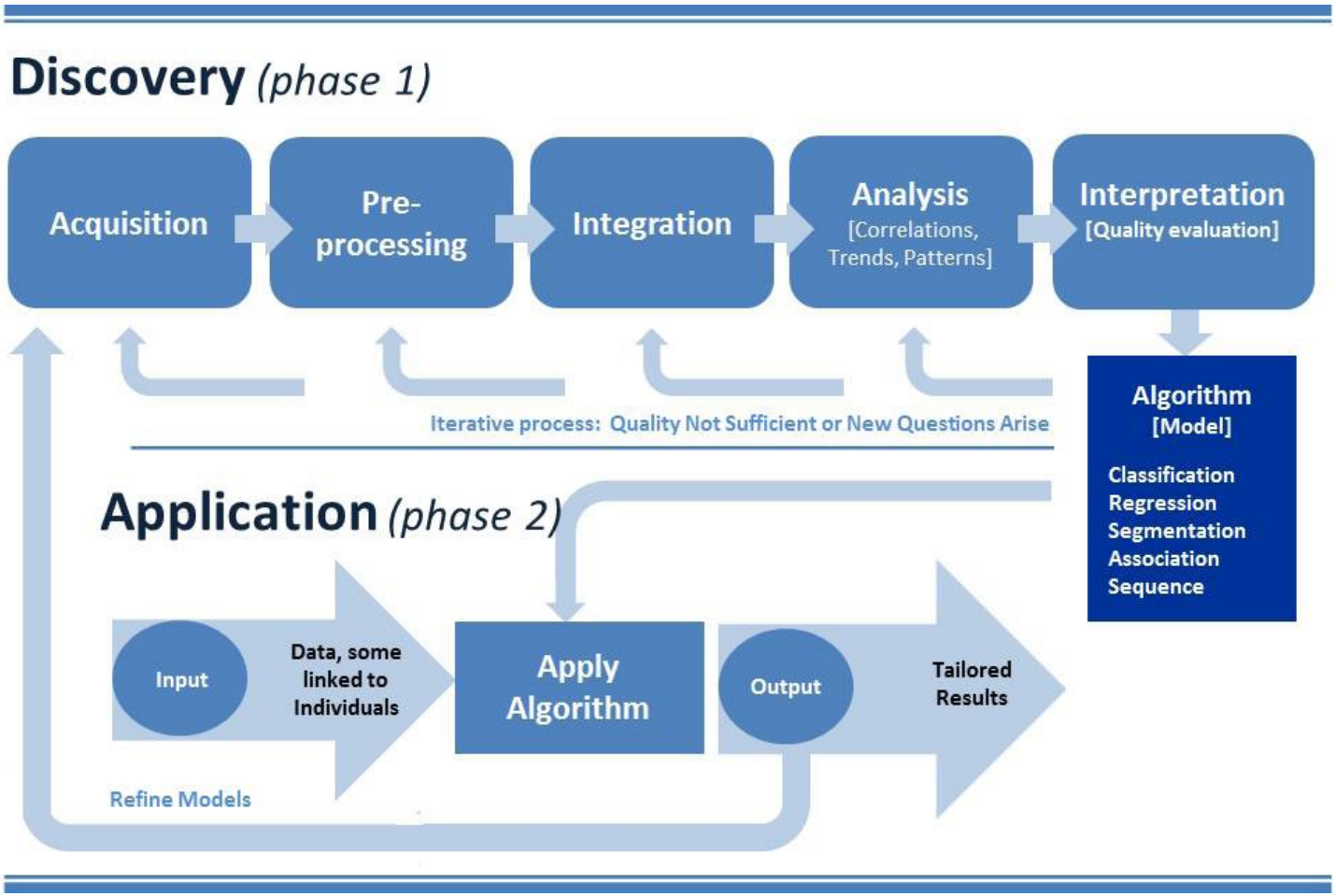
- Big Analytics delivers competitive advantage compared to the traditional analytical model.
- Big Analytics describes the efficient use of a simple model applied to volumes of data that would be too large for the traditional analytical environment.
- Research suggests that a simple algorithm with a large volume of data is more accurate than a sophisticated algorithm with little data.

Data Analytics

- Big Analytics supporting the following objectives for working with Big Data Analytics:

1. Avoid sampling / aggregation;
2. Reduce data movement and replication;
3. Bring the analytics as close as possible to the data.
4. Optimize computation speed.

The Process of Data Analytics



Data Analytics Process: Discovery

- The knowledge discovery phase involves
  - gathering data to be analyzed.
  - pre-processing it into a format that can be used.
  - consolidating it for analysis,
  - analyzing it to discover what it may reveal.
  - and interpreting it to understand the processes by which the data was analyzed and how conclusions were reached.

Data Analytics Process: Discovery

- ***Acquisition***
  - Data acquisition involves collecting or acquiring data for analysis.
  - Acquisition requires access to information and a mechanism for gathering it.
- ***Pre-processing***
  - Pre-processing is necessary if analytics is to yield trustworthy , useful results.
  - places it in a standard format for analysis.

Data Analytics Process: Discovery

- ***Integration***
  - Integration involves consolidating data for analysis.
    - Retrieving relevant data from various sources for analysis
    - Eliminating redundant data or clustering data to obtain a smaller representative sample
- ***Analysis***
    - Searching for relationships between data items in a database, or exploring data in search of classifications or associations.
    - Analysis can yield descriptions or predictions.
    - Analysis based on interpretation, organizations can determine whether and how to act on them.

Data Analytics Process: Discovery

- ***Interpretation***
  - Analytic processes are reviewed by data scientists to understand results and how they were determined.
  - Interpretation involves retracing methods, understanding choices made throughout the process and critically examining the quality of the analysis.
  - It provides the foundation for decisions about whether analytic outcomes are trustworthy.

Data Analytics Process: Application

- ***Application***
  - Associations discovered amongst data in the knowledge phase of the analytic process are incorporated into an algorithm and applied.
  - In the application phase organizations gather the benefits of knowledge discovery.
  - Through application of derived algorithms, organizations make determinations upon which they can act.

1.3 Role of Distributed System in Big Data

- What is a Distributed System?
  - A distributed system consists of a collection of autonomous computers, connected through a network and distribution middleware, which enables computers to coordinate their activities and to share the resources of the system, so that users perceive the system as a single, integrated computing facility.

Big data is distributed data

- Big data is **distributed data** : Data is so massive it cannot be stored or processed by a single node.
- The way to scale fast and affordably is to use commodity hardware to distribute the storage and processing of our massive data streams across several nodes, adding and removing nodes as needed.

Distributed data generation is fueling big data growth

- The reason we have data problems so big that we need large-scale distributed computing architecture to solve is that the creation of the data is also large-scale and distributed.
- Most of us walk around carrying devices that are constantly pulsing all sorts of data into the cloud and beyond – our locations, our photos, our tweets, our status updates, our connections, even our *heartbeats*.

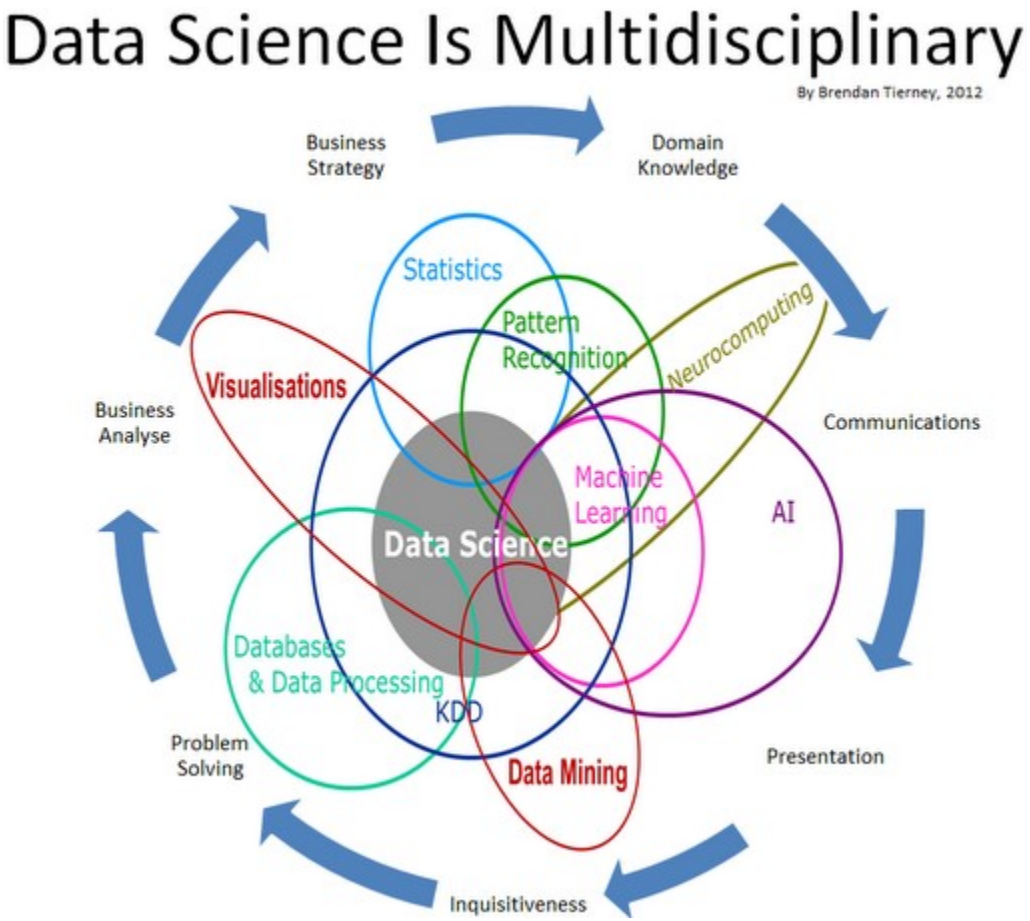
“Hadoop” and “MapReduce

- Hadoop: An open source platform for consolidating, combining and understanding large-scale data in order to make better business decisions.
- 2 key parts to Hadoop:
  - HDFS (Hadoop distributed file system) which lets you store data across multiple nodes.
  - MapReduce which lets you process data in parallel across multiple nodes.

1.4 Role of data Scientist

- **Data Science**
  - **Data science**, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
  - It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Data Science



Data Scientist

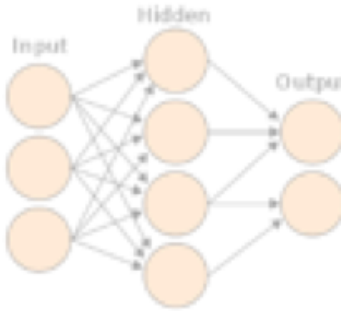
- High ranking professional with training and curiosities to make discovery in the world of big data.
- The people who understand how to fish out answers to important business questions from today's tsunami of unstructured information.
- Newly coined term , in 2008 by D.J Patil and Jeff Hammerbacher
- A hybrid of data hacker, analyst, communicator,and trusted adviser. The combination is extremely powerful—and rare

Data Scientist

- Sudden appearance of Data Scientist on the business scene reflects the fact that companies are now wrestling with information that comes in varieties and volumes never encountered before.
- If the organization stores multiple petabytes of data, if the information most critical to the business resides in forms other than rows and columns of numbers, or if answering the biggest question would involve a “mashup” of several analytical efforts, it has got a big data opportunity.

Data Scientist

The Data Scientist



- A New Role Exists – the **Data Scientist**
  - One Part Scientist/Statistician
  - Two Parts Sleuth/Artist
  - One Part Programmer
  - Focused on *data* not models
- Working with **analysts** to create business value

Confidential Think Big Analytics

Data scientist: a brand new profession

- Data Scientist: The Sexiest Job of the 21st Century [Harward Business Review 2013]
- Data scientist? A guide to 2015's hottest profession [Mashable 2015]
- “It’s official – data scientist is the best job in America” [Forbes, 2016]
- "This hot new field promises to revolutionize industries from business to government, health care to academia."

  — *The New York Times*

Successful Data Scientist Characteristics

- **Intellectual curiosity, Intuition**
  - Find needle in a haystack(something that is difficult to locate in a much larger space)
  - Ask the right questions – value to the business
- **Communication and engagements**
- **Presentation skills**
  - Let the data speak but tell a story
  - Story teller – drive business value not just data insights
- **Creativity**
  - Guide further investigation
- **Business Savvy**
  - Discovering patterns that identify risks and opportunities
  - Measure

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Role/Skill of Data Scientist

Data Scientist should have skill set to

- use technologies that make taming big data possible, including Hadoop (the most widely used framework for distributed file system processing) and related open-source tools, cloud computing, and data visualization.
- make discoveries while swimming in pool of data
- bring structure to large quantities of formless data and make analysis possible
- identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set

Role/Skill of Data Scientist

Data Scientist should have skill set to

- communicate what they've learned and suggest its implications for new business directions
- be creative in displaying information visually and making the patterns they find clear and compelling
- fashion their own tools and even conduct academic-style research
- write code
- desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested

Data Scientist Job Description

- Amazon's Shopper Marketing & Insights team focuses on serving the advertisers and our overall ad business to provide strategic media planning, customer insights, targeting recommendations, and measurement and optimization of advertising.
- We are hiring outstanding Data Scientists who will use innovative statistical and machine learning approaches to drive advertising optimization and contribute to the creation of scalable insights. The ideal candidate should have one hand on the white-board writing equations and one hand on the keyboard writing code.

Data Scientist

- A recent study by the McKinsey Global Institute concludes, "a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.).“
- The report estimates that there will be four to five million jobs in the U.S. requiring data analysis skills by 2018, and that large numbers of positions will only be filled through training or retraining. The authors also project a need for 1.5 million more managers and analysts with deep analytical and technical skills "who can ask the right questions and consume the results of analysis of big data effectively.“

https://datascience.berkeley.edu/about/what-is-data-science/

1.5 Current Trend in Big Data Analytics (Big Data Trends 2020)

- **Augmented analytics**
  - combines data analysis with ML and NLP
- **Edge Computing**
  - brings computation and data storage closer to the location where it is needed
- **In-Memory Computing**
  - used to perform real-time data analysis

Big Data Trends 2020

- **DataOps**
  - used to improve the quality and reduce the cycle time of data analytics
- **Chief Data Officers (CDOs)**
  - responsible for governance and utilization of information within a company
- **Continuous Intelligence (CI)**
  - integrates real-time analytics with business operations

https://addepto.com/big-data-trends-2020/

Big Data Applications

- Healthcare
- Manufacturing
- Media & Entertainment
- Internet of Things (IoT)
- Government
- Ecommerce
- Disaster Management
- Digital Marketing
- Telecommunication
- Retail Industry
- Finance
- Education

Some Big Data Use Cases

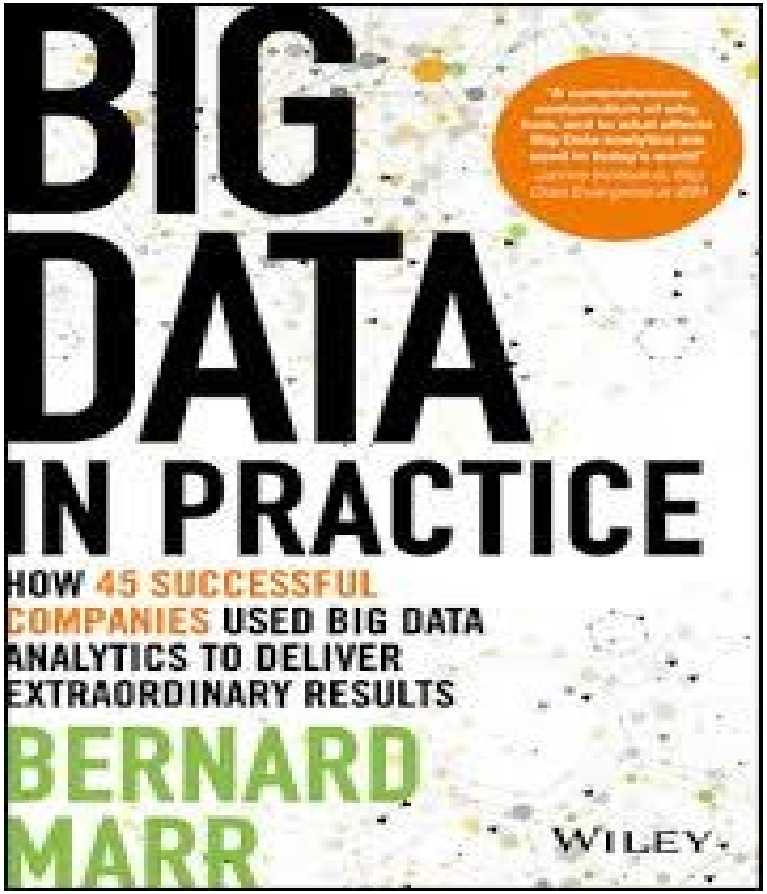
- Credit Card Fraud Detection
- Sentiment Analysis
- Delivering personalized customer experience
- Preventing customer churn
- Genomic Research

How is Big Data Used

- Understanding and Targeting Customers
- Understanding and Optimizing Business Processes
- Improving Sports Performance
- Improving and Optimizing Cities and Countries
- Improving Healthcare and Public Health

https://www.bernardmarr.com/default.asp?contentID=1076

Big Data in Practice



Big Data in COVID-19 Pandemic

- Identification of infected cases
  - It is capable of storing the complete medical history of all patients, due to its capability of storing a massive amount of data
  - By providing the captured data, this technology helps in identification of the infected cases and undertake further analysis of the level of risks

Big Data in COVID-19 Pandemic

- Travel history
  - Used to store the travel history of the people to analyze the risk
  - Helps to identify people who may be in contact with the infected patient of this virus

Big Data in COVID-19 Pandemic

- Fever symptoms
  - Big data can keep the record of fever and other symptoms of a patient and suggest if medical attention is required
  - Helps to identify the suspicious cases and other misinformation with the appropriate data

Big Data in COVID-19 Pandemic

- Identification of the virus at an early stage
  - Quickly helps to identify the infected patient at an early stage
  - Helps to analyze and identify persons who can be infected by this virus in future

Big Data in COVID-19 Pandemic

- Identification and analysis of fast-moving disease
  - Helps to effectively analyze the fast-moving disease as efficiently as possible
  - Potential to handle appropriate information regarding the disease

Big Data in COVID-19 Pandemic

- Information during lockdown
  - This technology collects information regarding this virus during the lockdown
  - Track and monitor the movement of people and entire health management

Big Data in COVID-19 Pandemic

- People entered or leaving the affected area
  - It helps to analyze the number of people entered or leaving from the affected city
  - With these vast amount of data, health specialist can quickly identify the chances of the virus in those peoples

Big Data in COVID-19 Pandemic

- Faster development of medical treatments
  - Assist in fast-tracking the development of new medicines and equipment needed for current and future medicinal needs
  - Provides previous data of virus inhabited or spread and, thus, helps in gaining a giving advantage over newer pandemic/epidemic with previously analyzed results

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193/

Role Of Big Data In The Fight Against COVID-19

- China's Surveillance Infrastructure Used to Track Exposed People
- Mobile App for Contact Tracing
- Official Dashboards Track the Virus and Outbreak Analytics
- Big Data Analytics and Successes in Taiwan

https://www.linkedin.com/pulse/vital-role-big-data-fight-against-covid-19-coronavirus-bernard-marr

Scope of Big Data

- Increasing demand for Data Analytics
- Increasing enterprise adoption of Big Data
- Big Data finds application across various parallels of the industry
- Huge Job Opportunities & Meeting the Skill Gap
- Promises exponential salary growth
- Key Decision-Making Power

Challenges of Big Data

- Dealing with data growth
- Generating insights in a timely manner
- Recruiting and retaining big data talent
- Integrating disparate data sources
- Securing big data
- Organizational resistance

Thank You !