INTRODUCTION TO BIG DATA ANALYTICS

Lecture 1

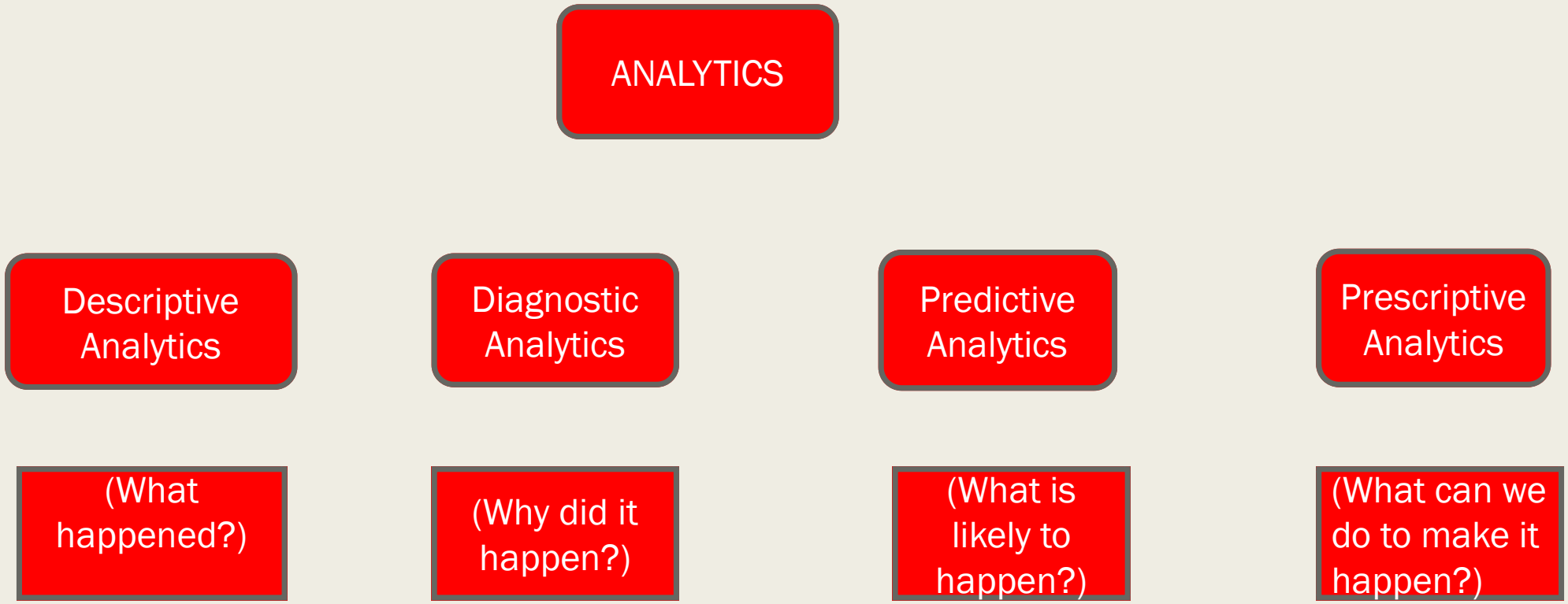
Learning Objectives

- Learn about Analytics
- Analyze the growing level of data.
- Demonstrate the characteristics and types of big data.
- Learn why traditional data storage cannot be used to store big data.
- Learn the availability of modern technologies developed to handle big data.
- Learn Big Data solution application for different case studies.

What is Analytics?

- Analytics is this process of extracting and creating information from raw data by filtering, processing, categorizing, condensing and contextualizing the data.
- This information obtained is then organized and structured to infer knowledge about the system and/or its users, its environment, and its operations and progress towards its objectives, thus making the systems smarter and more efficient.

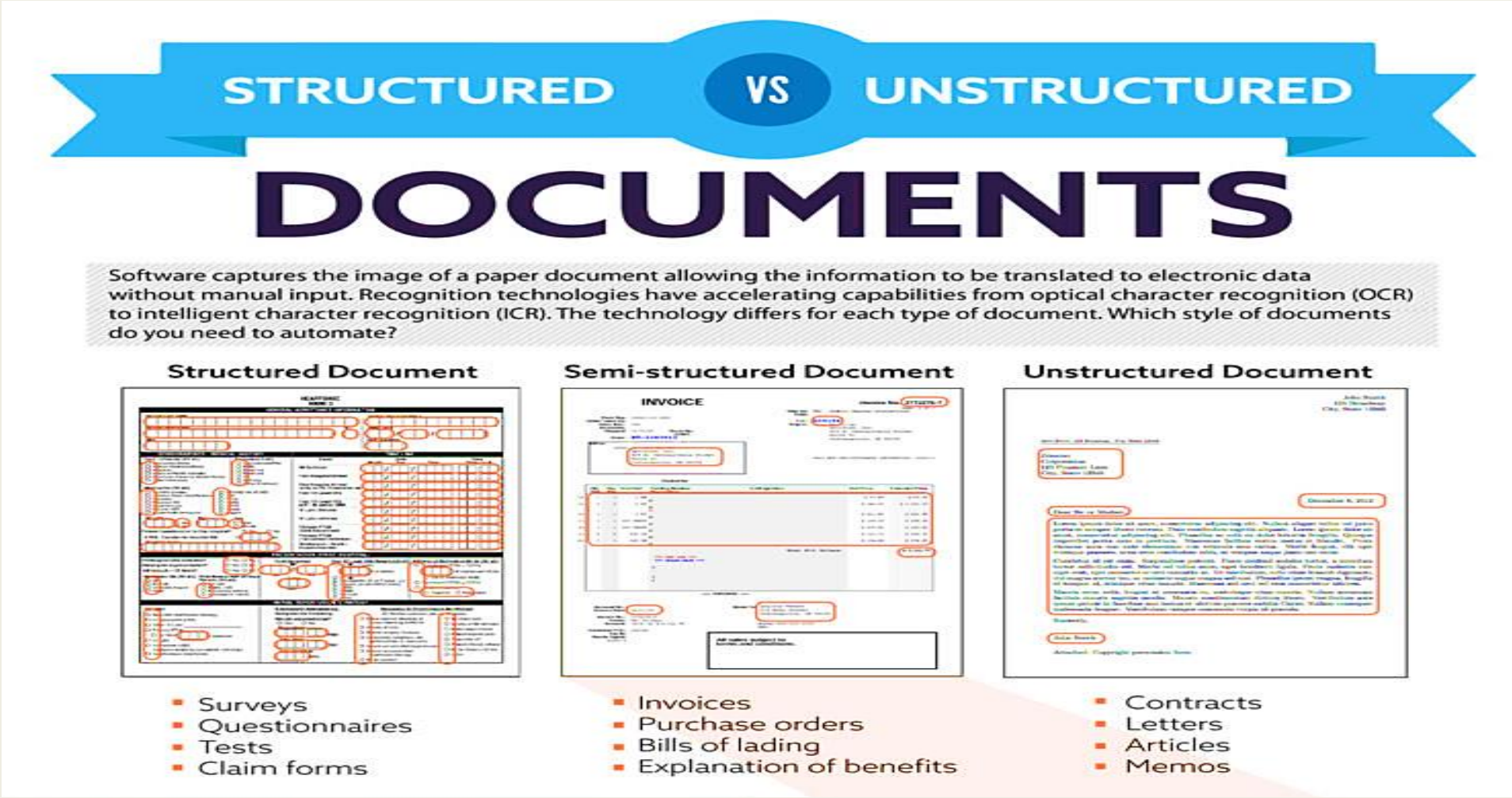
Types of Analytics



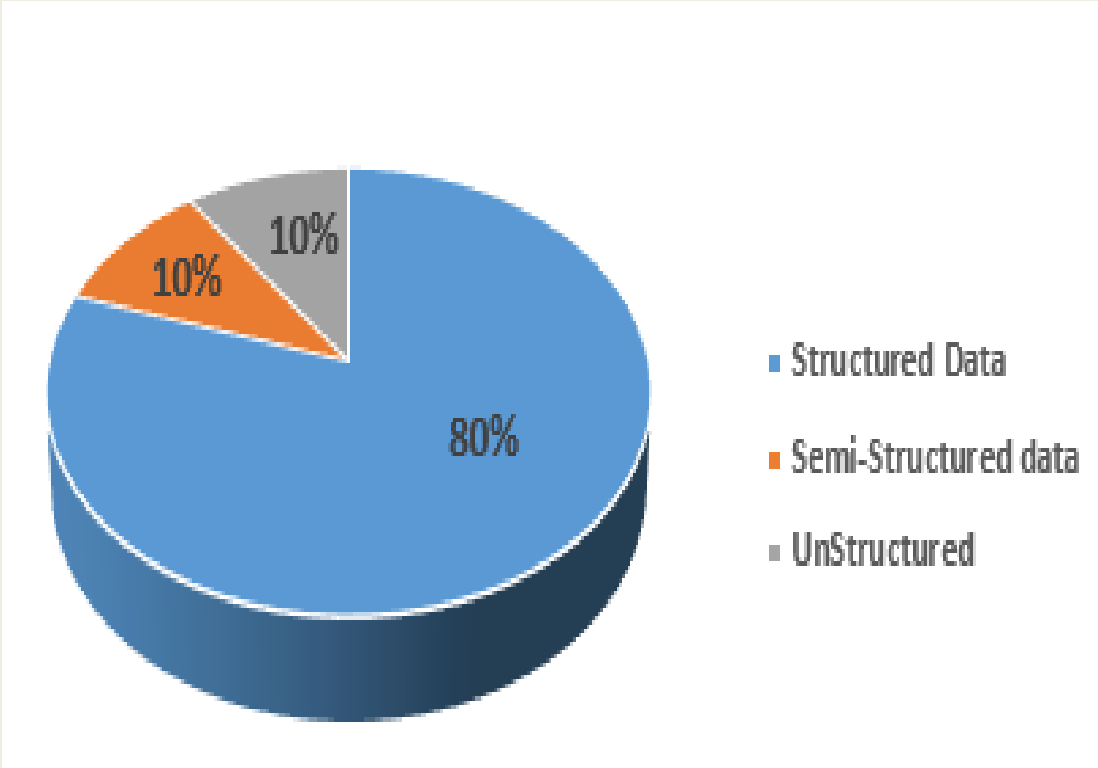
Formats of Big Data or Digital Data

- ***Structured:*** Organized data format with a fixed schema. Ex: RDBMS
- ***Semi-Structured:*** Partially organized data which does not have a fixed format. Ex: XML, JSON
- ***Unstructured:*** Unorganized data with an unknown schema. Ex: Audio, video files etc.

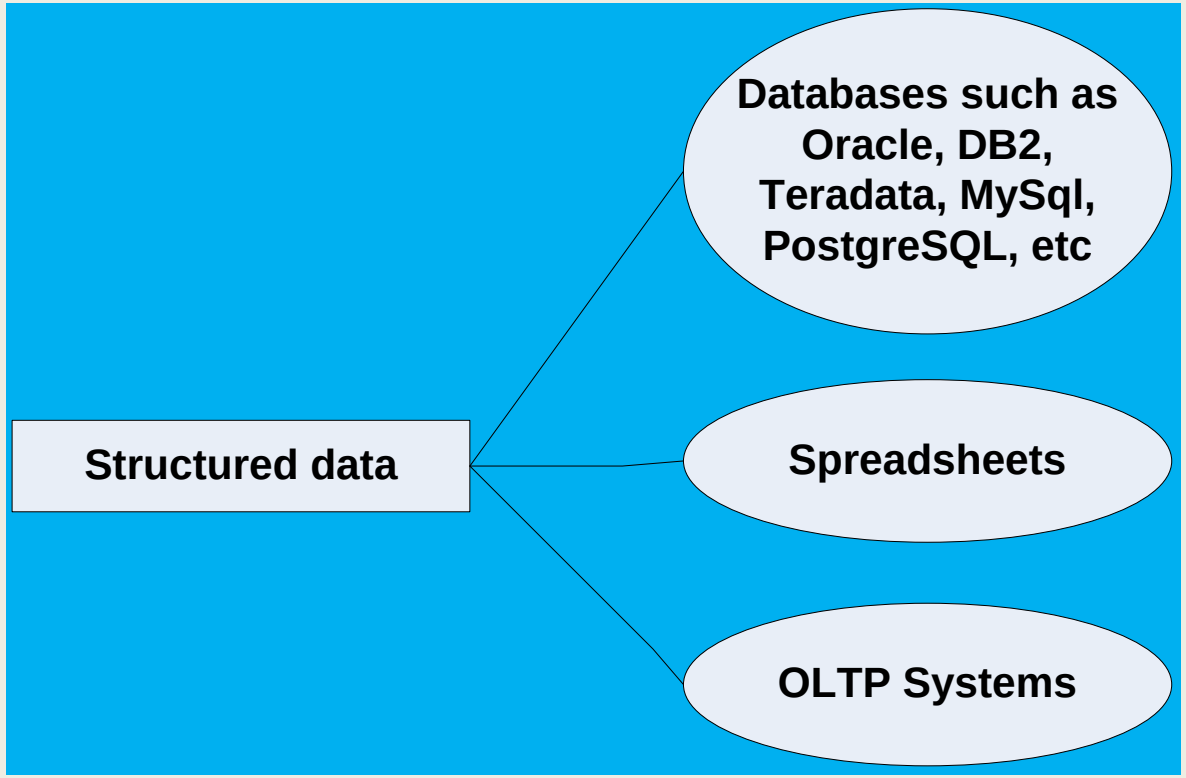
Formats of Big Data



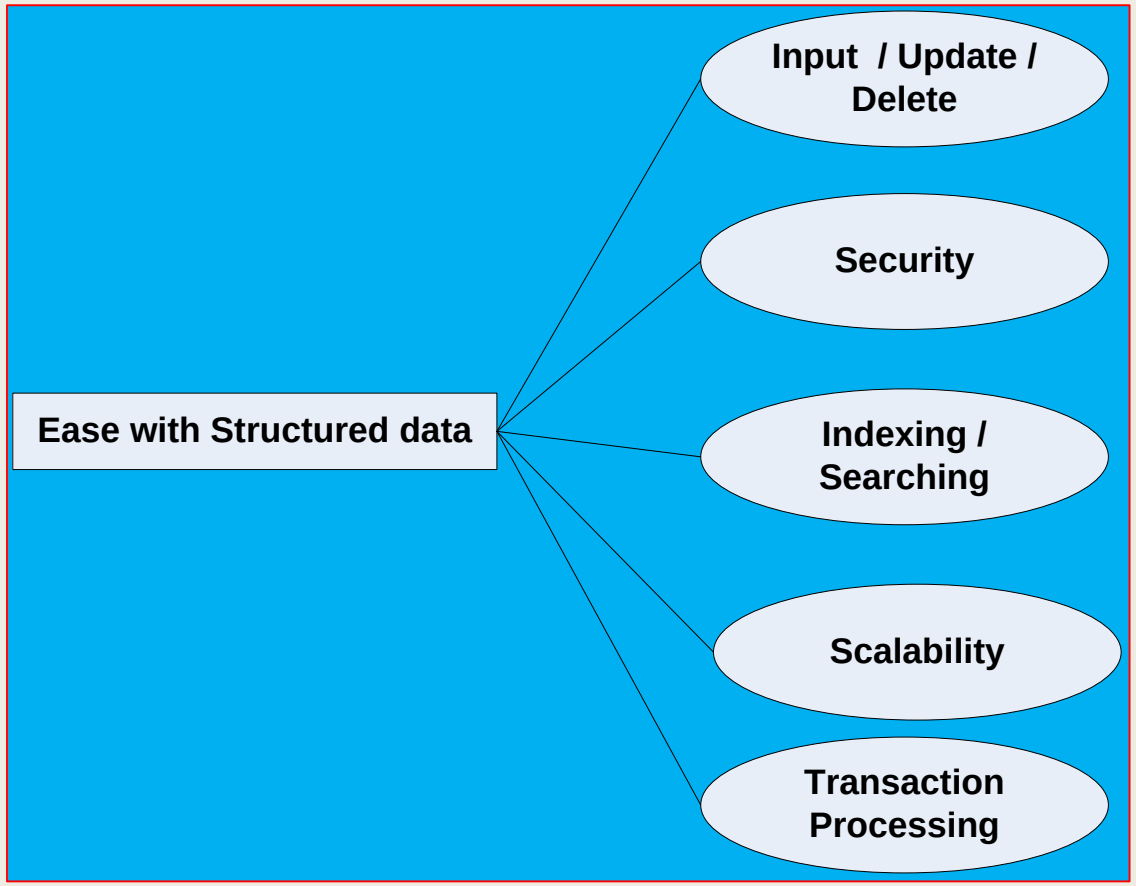
Approximate Percentage Distribution of Digital Data



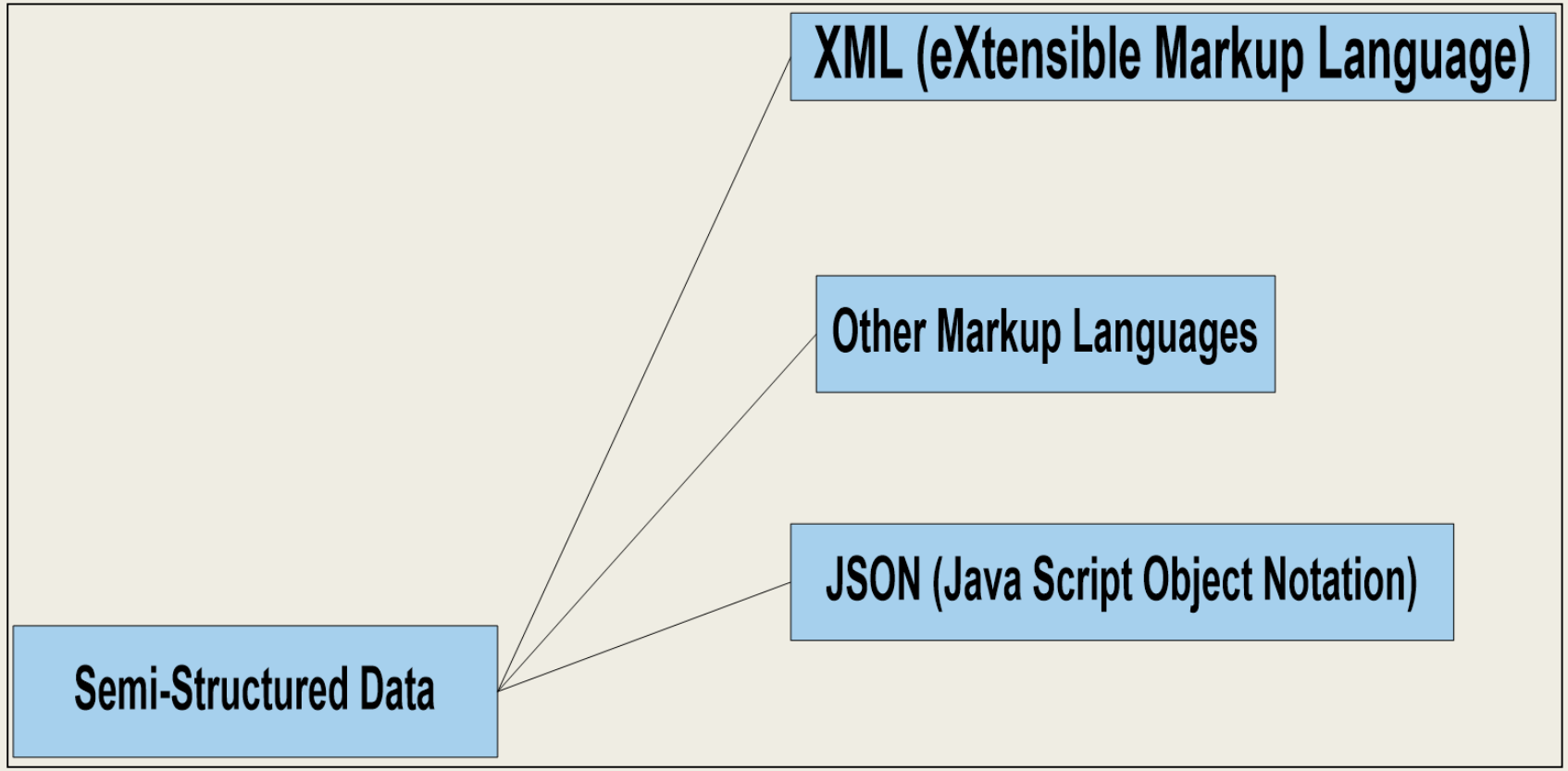
Sources of Structured Data



Ease with Structured Data



Sources of Semi-structured Data



Sources of Semi-structured Data

Sample JSON document

{

_id:9,

BookTitle: “Fundamentals of Business Analytics”,

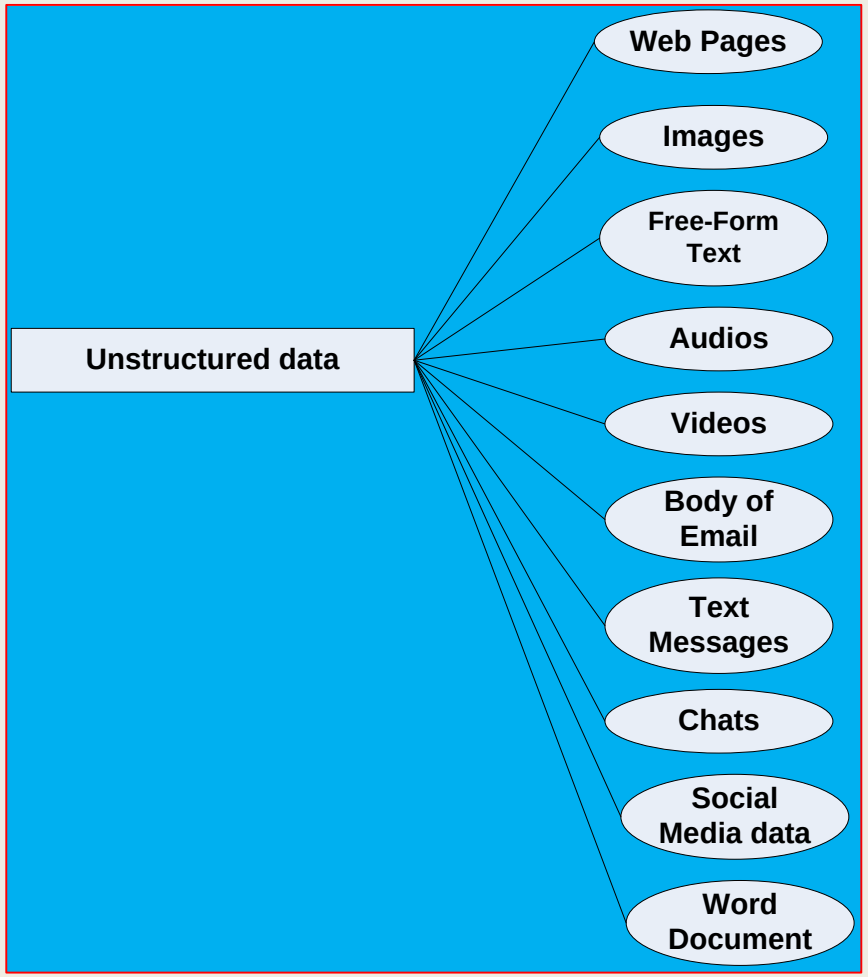
AuthorName: “Seema Acharya”,

Publisher: “Wiley India”,

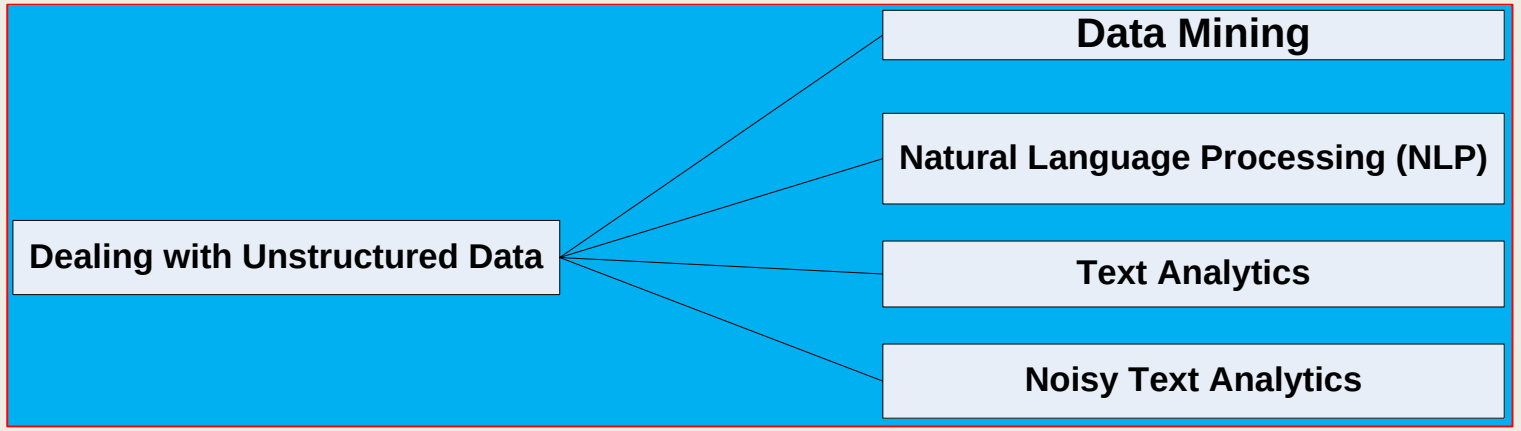
YearofPublication: “2011”

}

Sources of Unstructured Data



Dealing with Unstructured Data



Answer a few quick questions

- Which category (structured, semi-structured, or unstructured) will you place a Web Page in?
- Which category (structured, semi-structured, or unstructured) will you place Word Document in?
- Which category (structured, semi-structured, or unstructured) will you place streaming data in?
- Which category (structured, semi-structured, or unstructured) will you place tab-delimited text files in?
- Which category (structured, semi-structured, or unstructured) will you place BibTex files in?
- State a few examples of human generated and machine-generated data.

Lecture :2

CHARACTERISTICS OF DATA

- Composition
- Condition
- Context

Motivation behind Big Data

- According to an estimate by IBM, 2.5 quintillion bytes of data is created every day. A recent report by DOMO estimates the amount of data generated every minute on popular online platforms. Below are some key pieces of data from the report:
  - *Facebook users share nearly 4.16 million pieces of content*
  - *Twitter users send nearly 300,000 tweets*
  - *Instagram users like nearly 1.73 million photos*
  - *YouTube users upload 300 hours of new video content*

Motivation behind Big Data

- Apple users download nearly 51,000 apps.
- Skype users make nearly 110,000 new calls
- Amazon receives 4300 new visitors
- Uber passengers take 694 rides
- Netflix subscribers stream nearly 77,000 hours of video

Evolution of Big Data

	Data Generation and Storage	Data Utilization	Data Driven
Complex and Unstructured			Structured data, Unstructured data, Multimedia data
Complex and Relational		Relational databases: Data-intensive applications	
Primitive and Structured	Mainframes: Basic data storage		
Existence	1970s and before	Relational(1980s and 1990s)	2000s and beyond

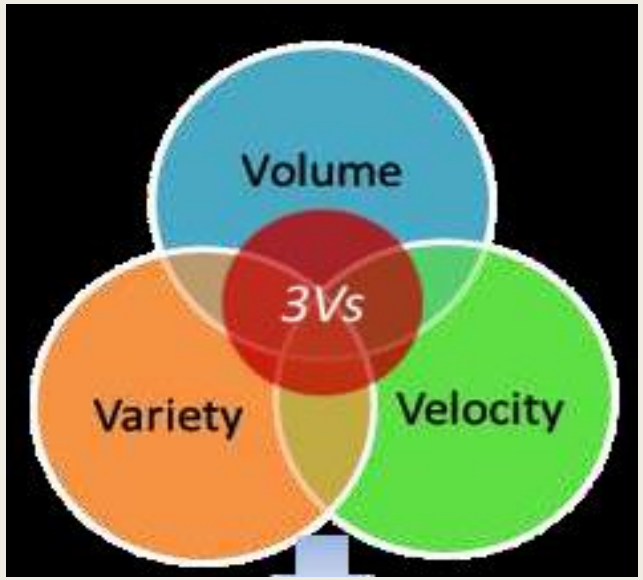
What is Big Data?

- Big data is defined as collections of datasets whose volume, velocity or variety is so large that it is difficult to store, manage, process and analyze the data using traditional databases and data processing tools.
- Big data refers to the massive datasets that are collected from a variety of data sources for business needs to reveal new insights for optimized decision making.

Gartner- 3Vs Definition

- Douglas Laney [28] in his white paper published by Meta group, which Gartner subsequently acquired in 2004. Douglas noticed that due to surging of e-commerce activities, data has grown along three dimensions, namely:
  - *Volume, which means Incoming data stream and Cumulative volume of data.*
  - *Velocity, which represents the pace data used to support interaction and generated by interactions*
  - *Variety, which signifies the variety of incompatible and inconsistent data formats and data structures.*

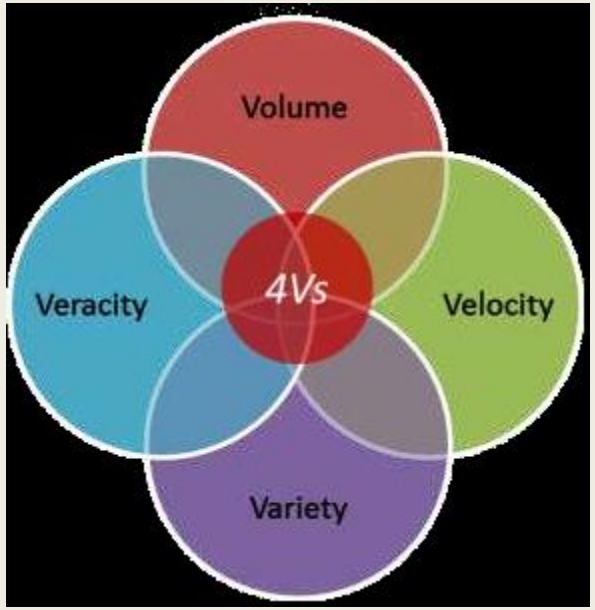
Gartner- 3Vs Definition



IBM- 4Vs Definition

- IBM added another attribute or “V” for “Veracity” on the top of Douglas Laney’s 3Vs notation, which is so called as Four Vs of Big Data. It defines each “V” as following :
  - *Volume stands for scale of data*
  - *Velocity denotes to analyzing streaming data*
  - *Variety indicates different forms of data*
  - *Veracity implies uncertainty of data*

IBM- 4Vs Definition



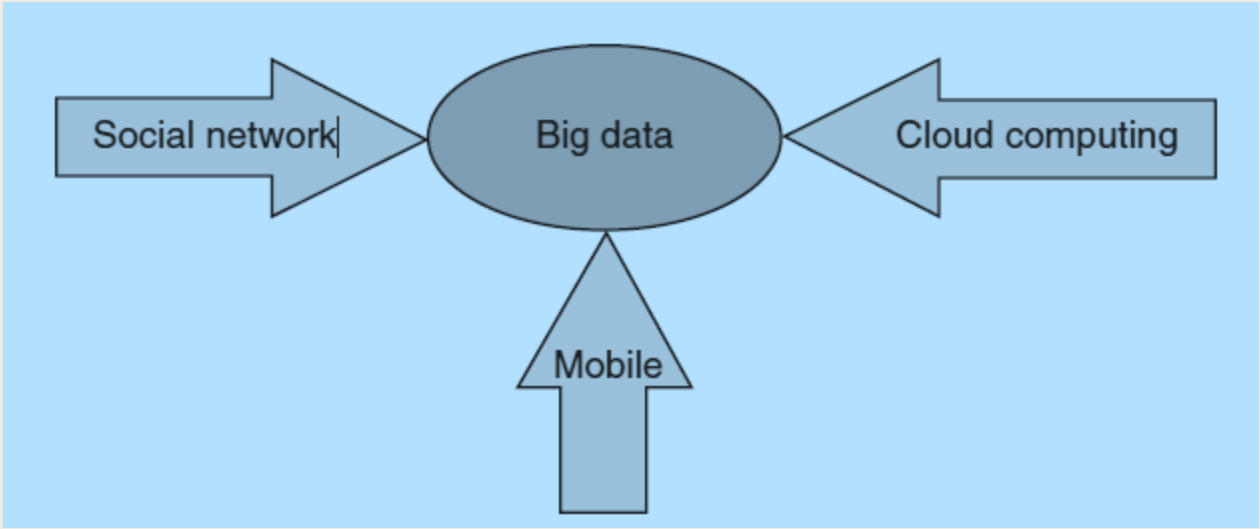
Microsoft - 6Vs Definition

- For the sake of maximizing the business value, Microsoft extended Douglas Laney's 3Vs attributes to 6 Vs which it added Variability, Veracity and Visibility:
  - *Volume stands for scale of data*
  - *Velocity denotes to analyzing streaming data*
  - *Variety indicates different forms of data*
  - *Veracity focuses on trustworthiness of data sources.*

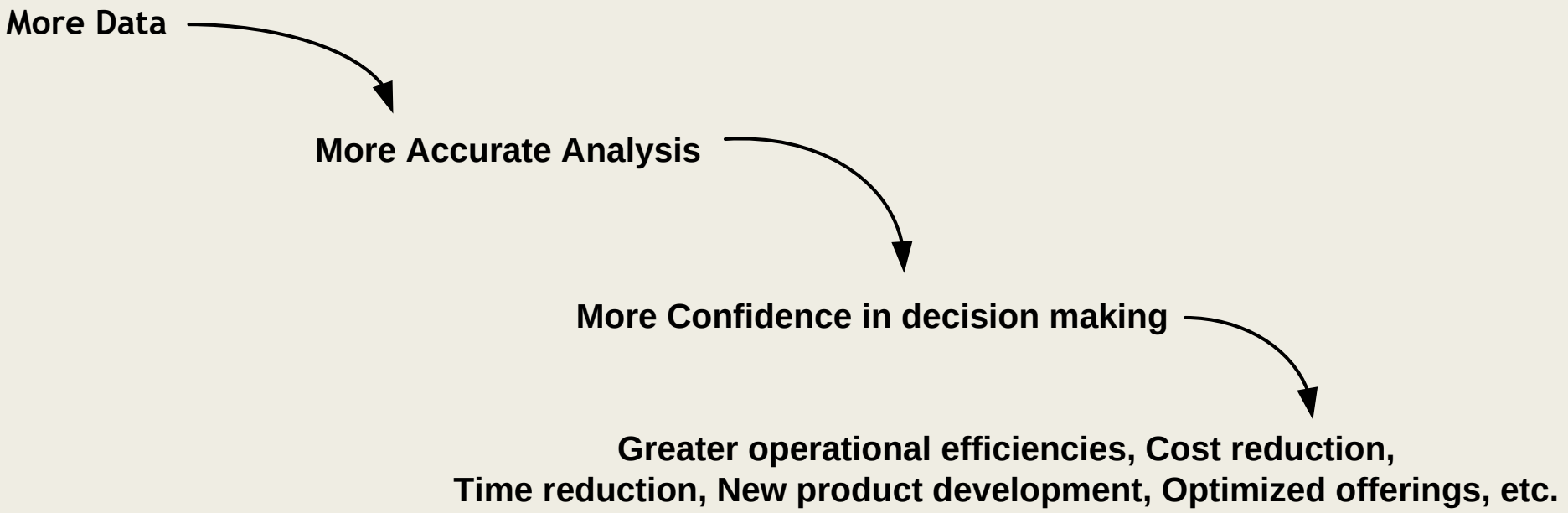
Microsoft - 6Vs Definition

- *Variability refers to the complexity of data set. In comparison with "Variety" (or different data format), it means the number of variables in data sets.*
- *Visibility emphasize that you need have a full picture of data in order to make informative decision.*

Big Data: Result of three computing trends

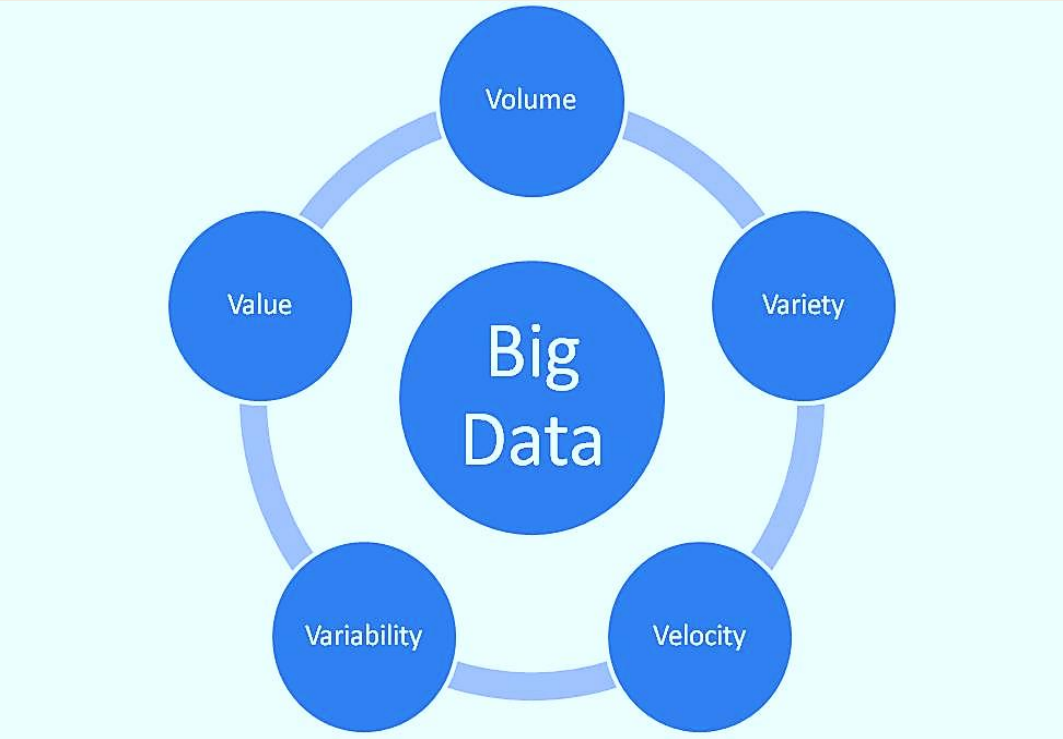


Why Big Data?

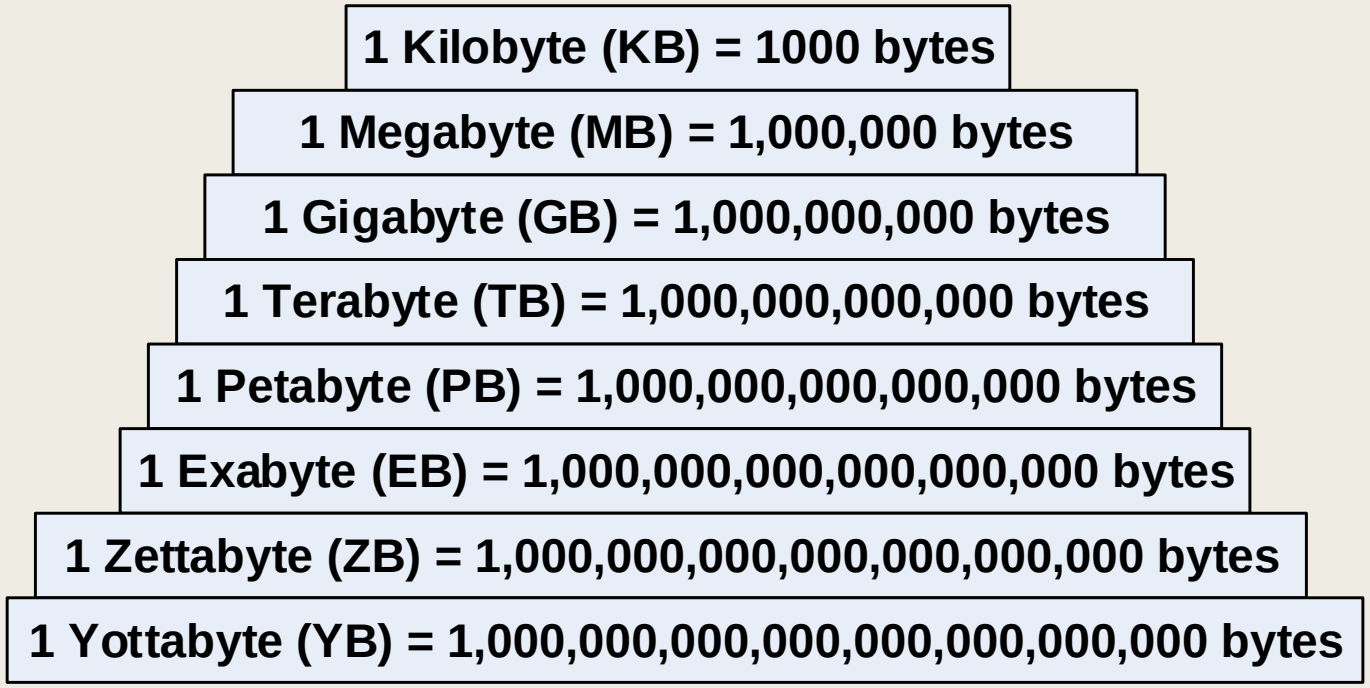

More Data
More Accurate Analysis
More Confidence in decision making
Greater operational efficiencies, Cost reduction,
Time reduction, New product development, Optimized offerings, etc.

Big Data Characteristics

- Volume
- Velocity
- Variety
- Veracity
- Value



Volume - A Mountain of Data



Where Does This Data get Generated?

- There are a multitude of sources for big data.
- An XLS, a DOC, a PDF. etc. is unstructured data;
- A video on YouTube, a chat conversation on Internet Messenger, a customer feedback form on an online retail website is unstructured data;
- A CCTV coverage, a weather forecast report is unstructured data too.

1. Typical Internal Data Sources

- Data present within an organizations firewall. It is as follows:
  - Data storage: File systems, SQL (RDBMSs - Oracle, MS SQL Server, DB2, MySQL, PostgreSQL, etc.), NoSQL (MongoDB, Cassandra, etc.), and so on.
  - Archives: Archives of scanned documents, paper archives, customer correspondence records, patients' health records, students admission records, students' assessment records, and so on.

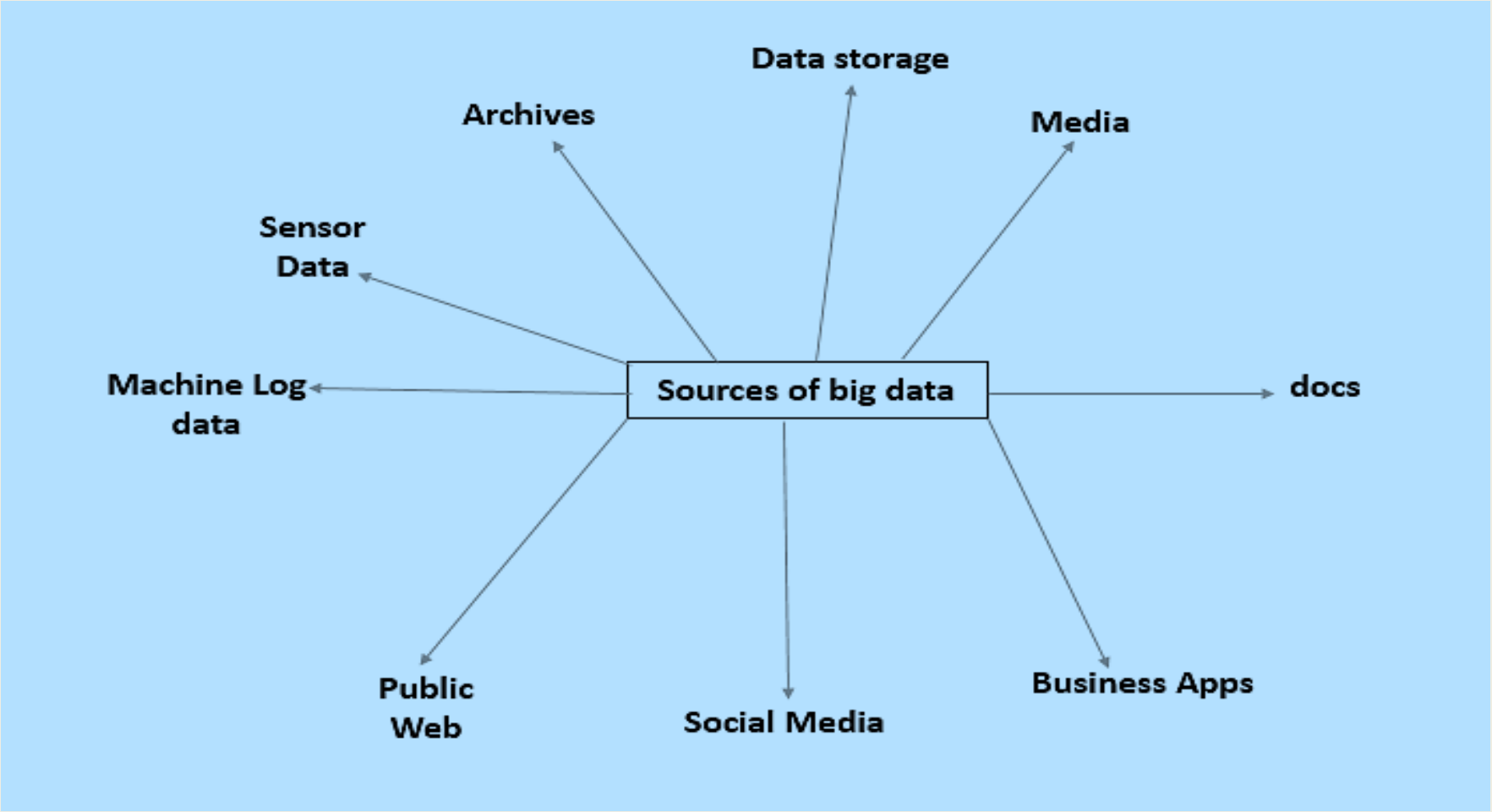
2. External Data Sources

- Data residing outside an organization's firewall. It is as follows:
  - **Public Web:** Wikipedia, weather, regulatory, compliance, census, etc.

3. Both (Internal and External) Data Sources

- **Sensor data:** Car sensors, smart electric meters, office buildings, air conditioning units, refrigerators, and so on.
- **Machine log data:** Event logs, application logs, Business process logs, audit logs, clickstream data, etc.
- **Social media:** Twitter, blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
- **Business apps:** ERP, CRM, HR, Google Docs, and so on.
- **Media:** Audio, Video, Image, Podcast, etc.
- **Docs:** Comma separated value (CSV), Word Documents, PDF, XLS, PPT, and so on.

Sources of Big Data

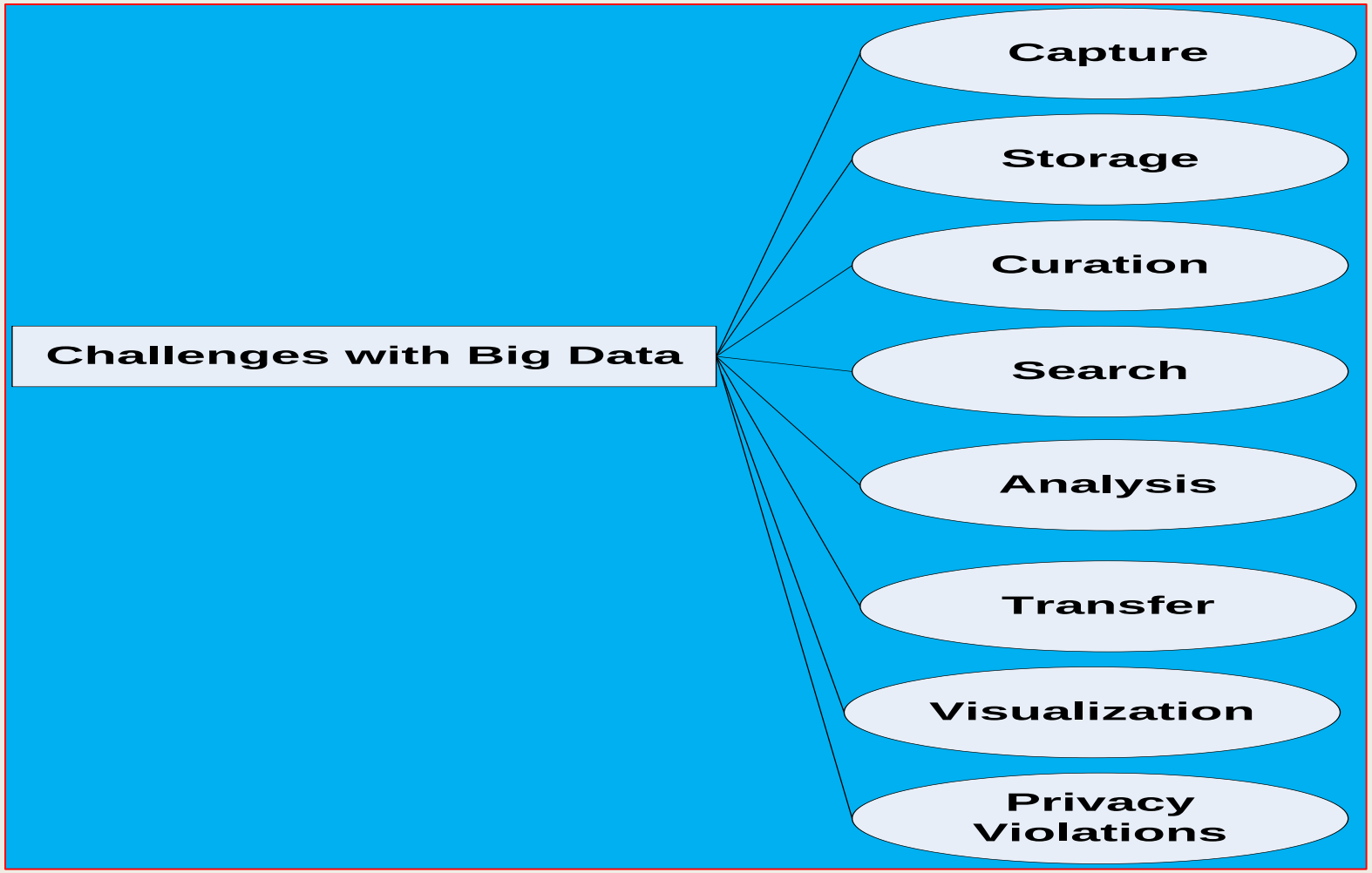


Lecture 3

Challenges with Big Data

- Usefulness
- Cloud computing and virtualization
- Retention
- Scarcity of Data Scientist
- Data Visualization
- Storage Capacity

Challenges with Big Data



Traditional BI Versus Big Data Approach

Comparison of Objectives	Business Intelligence	Big Data
Purpose	The purpose of Business Intelligence is to help the business to make better decisions. Business Intelligence helps in delivering accurate reports by extracting information directly from the data source.	The main purpose of Big Data is to capture, process, and analyze the data, both structured and unstructured to improve customer outcomes.

Traditional BI Versus Big Data Approach

Comparison of Objectives	Business Intelligence	Big Data
EcoSystem / Components	Operation systems, ERP databases, Data Warehouse, Dashboard etc.	Hadoop, Spark, R Server, hive, HDFS etc.
Characteristics/ Properties	Below are the six features of Business Intelligence Location intelligence, Executive Dashboards, "what if" analysis, Interactive reports, Metadata layer, and Ranking reports	Big data can be described by some characteristics such as Volume, Variety, Variability, Velocity, and Veracity.

Case Study of Big Data Solutions

- A few of the domain that can benefit from Big Data Analytics are mentioned below:
- Insurance companies can understand the likelihood of fraud by accessing the internal and external data while processing the claims. This will help them to speed up the handling of the simple claims and analyze the complex or fraudulent ones.
- Manufacturers and distributors can be benefitted by realizing supply chain issues earlier so that they can take decisions on different logistical approaches to avoid the additional costs associated with material delays, overstock or stock-out conditions.

Case Study of Big Data Solutions

- Firms such as hotels, telecommunications companies, retailers and restaurants that serve customers likely to have better clarity on customer needs to build a strong customer base and loyalty.
- Public services such as traffic, ambulance, transportations, etc. can optimize their delivery mechanisms by measuring the usage of these essential services.
- Smart-city is the buzz word today. The idea is make cities more efficient and sustainable to improve the lives of the citizens. Data related to censors, crime, emergency services, real-estate, energy, financial transactions, call details, astronomy, data.gov, customs.gov and scientific data are all used for analysis to do the improvement.

A Case Study

- What problem does Hadoop solve? Businesses and governments have a large amount of data that needs to be analyzed and processed very quickly.
- If this data is fragmented into smaller chunks and spread over many machines, all those machines process their portion of the data in parallel and the results are obtained extremely fast.
- For example, a huge data file containing feedback mails is sent to the customer service department.
- The objective is to find the number of times goods were returned and refund requested. This will help the business to find the performance of the vendor or supplier.

A Case Study

- It is a simple word count exercise. The client will load the data into the cluster (Feedback.txt),submit a job describing how to analyze that data (word count), the cluster will store the results in a new file (Returned.txt), and the client will read the results file.

Question's Answer ??

- How is traditional BI environment different from the Big Data environment?
- Share your experience as a customer on an e-commerce site. Comment on the big data that gets created on a typical e-commerce site.
- A bookmarking site permits to bookmark, review, rate and search various links on any topic. Analyze social bookmarking sites like stumbleupon.com or reddit.com to find insights.
- A Consumer Complaints site like complaintsboard.com, consumercourtforum.in allows to register complaints. Find the attributes available for analytics. Write a few queries.

References

- Douglas Laney, 3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, Meta Group, 6 Feb 2001, pp 1-4.
- Big data analytics by Radha Shankarmani, Vijayalakshmi
- Big Data Analytics_ A Hands-On Approach by Arshdeep Bahga, Vijay Medisetti