


Department of CSE

COURSE NAME: DBMS
COURSE CODE: 23AD2102R

Topic: Big Data Concepts

Session - 2

AIM OF THE SESSION



To familiarize students with the basic concept of BigData

INSTRUCTIONAL OBJECTIVES



This Session is designed to: discuss and study the concepts of BigData
Distributed Storage and Processing Framework (Hadoop)

LEARNING OUTCOMES



At the end of this session, you should be able to: understand Hadoop Framework

What is Big Data?

- Data which are very large in size is called Big Data.
- Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^{15} byte size is called Big Data.
- It is stated that almost 90% of today's data has been generated in the past 3 years.

Memory unit	Description
Kilo Byte	1 KB = 1024 Bytes
Mega Byte	1 MB = 1024 KB
Giga Byte	1 GB = 1024 MB
Tera Byte	1 TB = 1024 GB
Peta Byte	1 PB = 1024 TB
Hexa Byte	1 EB = 1024 PB
Zetta Byte	1 ZB = 1024 EB
Yotta Byte	1 YB = 1024 ZB
Bronto Byte	1 Bronto Byte = 1024 YB
Geop Byte	1 Geo Byte = 1024 Bronto Bytes

Sources of Big Data

These data come from many sources like

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

The 5 Vs of Big Data ?

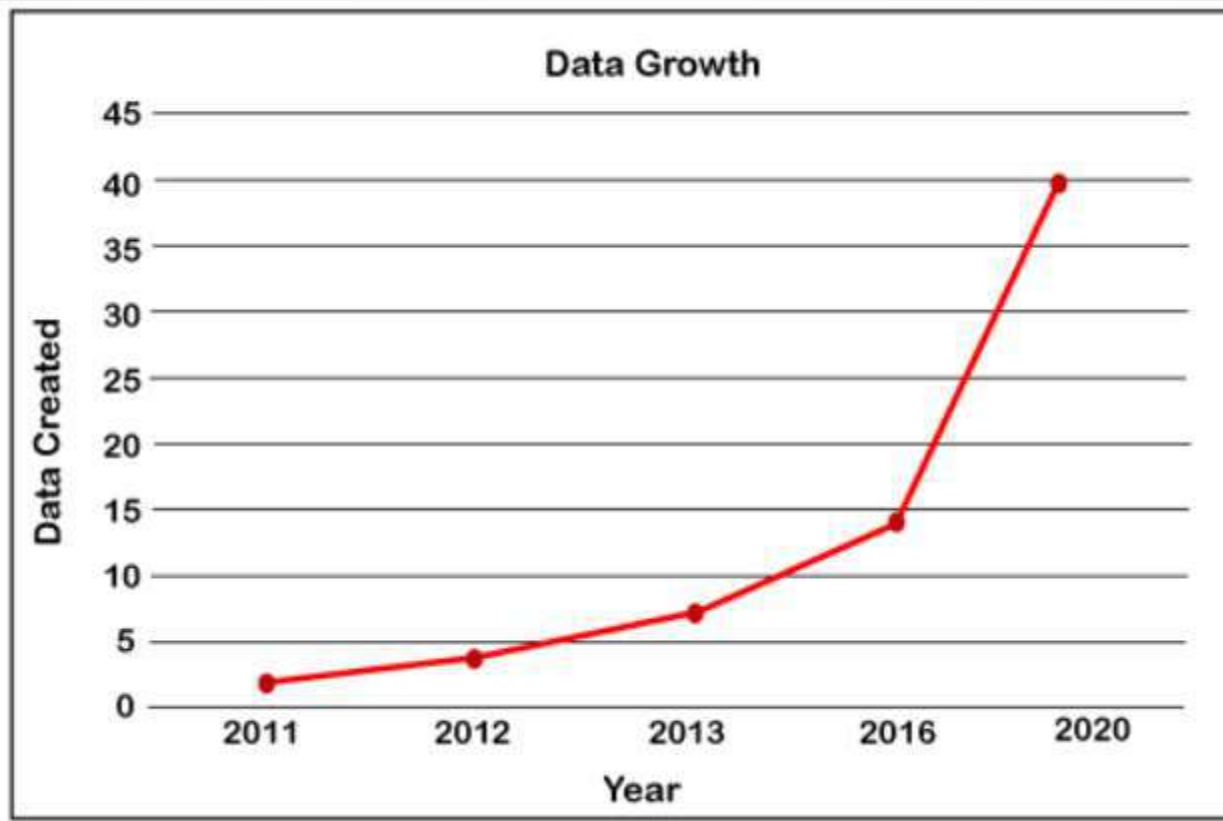


Big data is a collection of data from many different sources and is often describe by five characteristics: **volume, value, variety, velocity, and veracity**

- **Volume:** The amount of data
- **Velocity:** The speed at which data is generated and processed
- **Variety:** The different types of data
- **Veracity:** The accuracy and quality of the data
- **Value:** The value of the data

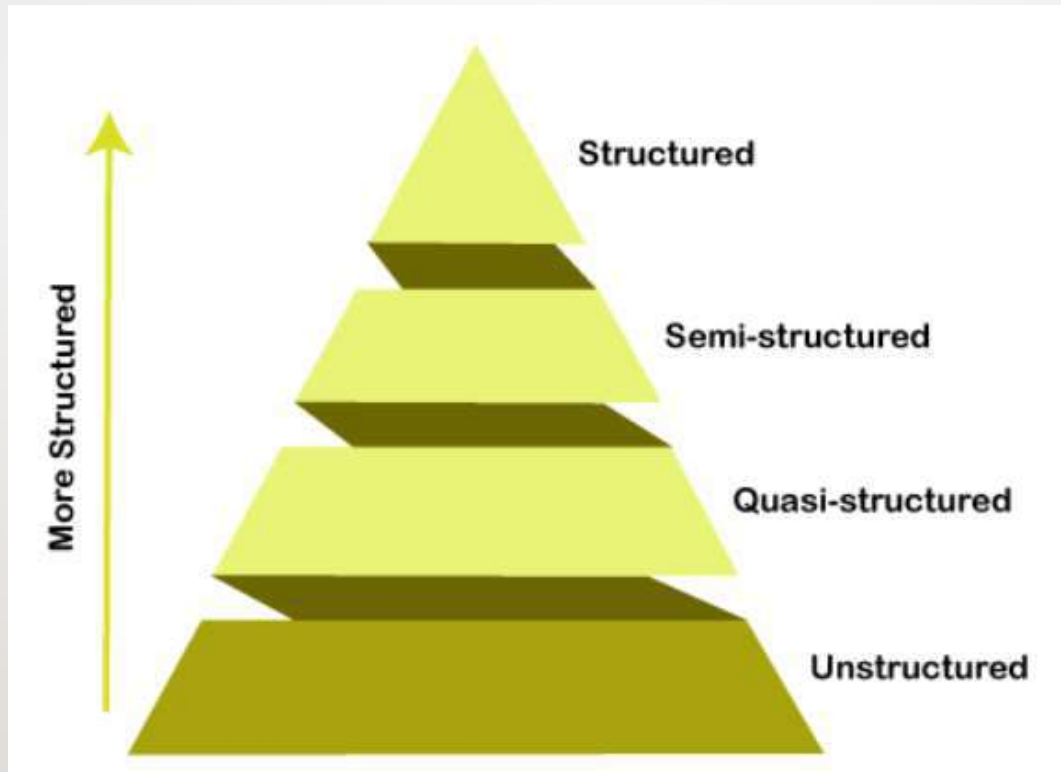
Volume

- The name Big Data itself is related to an enormous size.
- Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions**, and many more.
- **Facebook** can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.



Variety

- Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources.
- Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos**, etc.



The data is categorized as below:

1. **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.
2. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.
3. **Unstructured Data:** All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.

Veracity

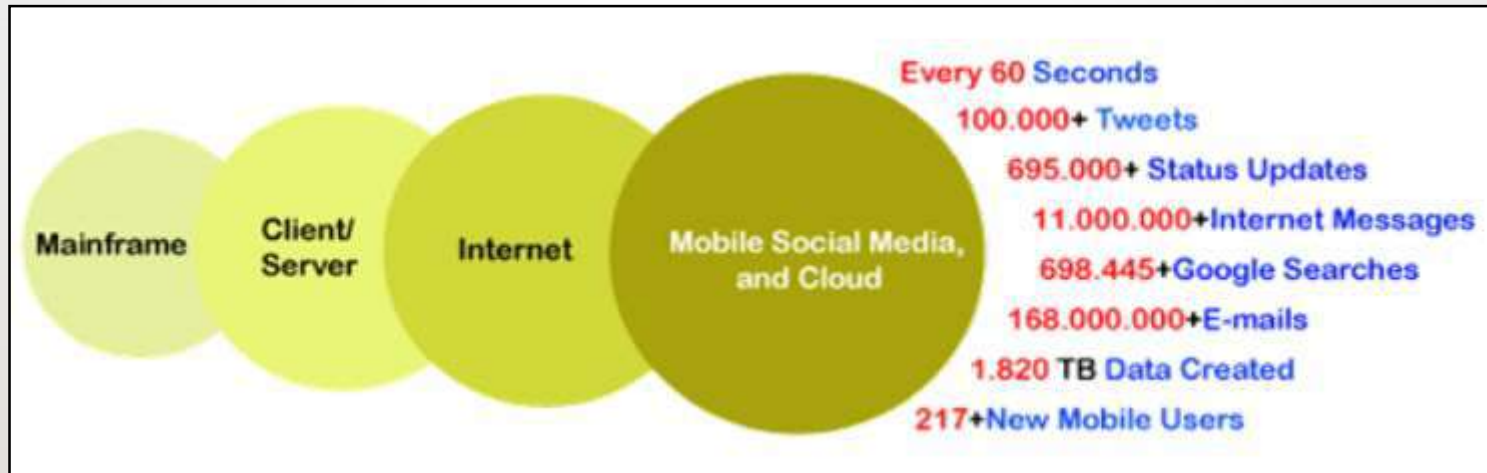
- Veracity means how much the data is reliable.
- It has many ways to filter or translate the data.
- Veracity is the process of being able to handle and manage data efficiently.
- Big Data is also essential in business development.
- For example, **Facebook posts** with hashtags.

Value

- Value is an essential characteristic of big data.
- It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process, and also analyze**.

Velocity

- Velocity creates the speed by which the data is created in **real-time**.
- It contains the linking of incoming **data sets speeds, rate of change,** and **activity bursts**.
- The primary aspect of Big Data is to provide demanding data rapidly.
- **Big data** velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.**



Use case

An e-commerce site XYZ (having 100 million users) wants to offer a gift voucher of 100\$ to its top 10 customers who have spent the most in the previous year. Moreover, they want to find the buying trend of these customers so that company can suggest more items related to them.

Issues

Huge amount of unstructured data which needs to be stored, processed and analyzed.

Solution

Storage: This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.

Processing: Map Reduce paradigm is applied to data distributed over network to find the required output.

Analyze: Pig, Hive can be used to analyze the data.

Cost: Hadoop is open source so the cost is no more an issue.