

# Chapter 1: Introduction to Big Data & Hadoop

---

By Ms. Tina D'abreo

# Content

- Introduction to Big Data
- Distributed file system
- Big Data characteristics, Drivers, types of Big Data
- Traditional vs. Big Data business approach
- Case Study of Big Data Solutions
- Big Data Applications
- Societal and Ethical issues associated with the use of big data analytics
- The key privacy issues.

## Introduction to Big Data

- The term Big Data refers to huge collections of data that are structured and unstructured which can't be processed or analyzed using traditional processes or tools
- 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.
- These massive volumes of data can be used to address business problems you wouldn't have been able to tackle before.
- Data may be sourced from servers, customer profile information, order and purchase data, financial transactions, ledgers, search history, and employee records.

## Introduction to Big Data - Big Data Principles

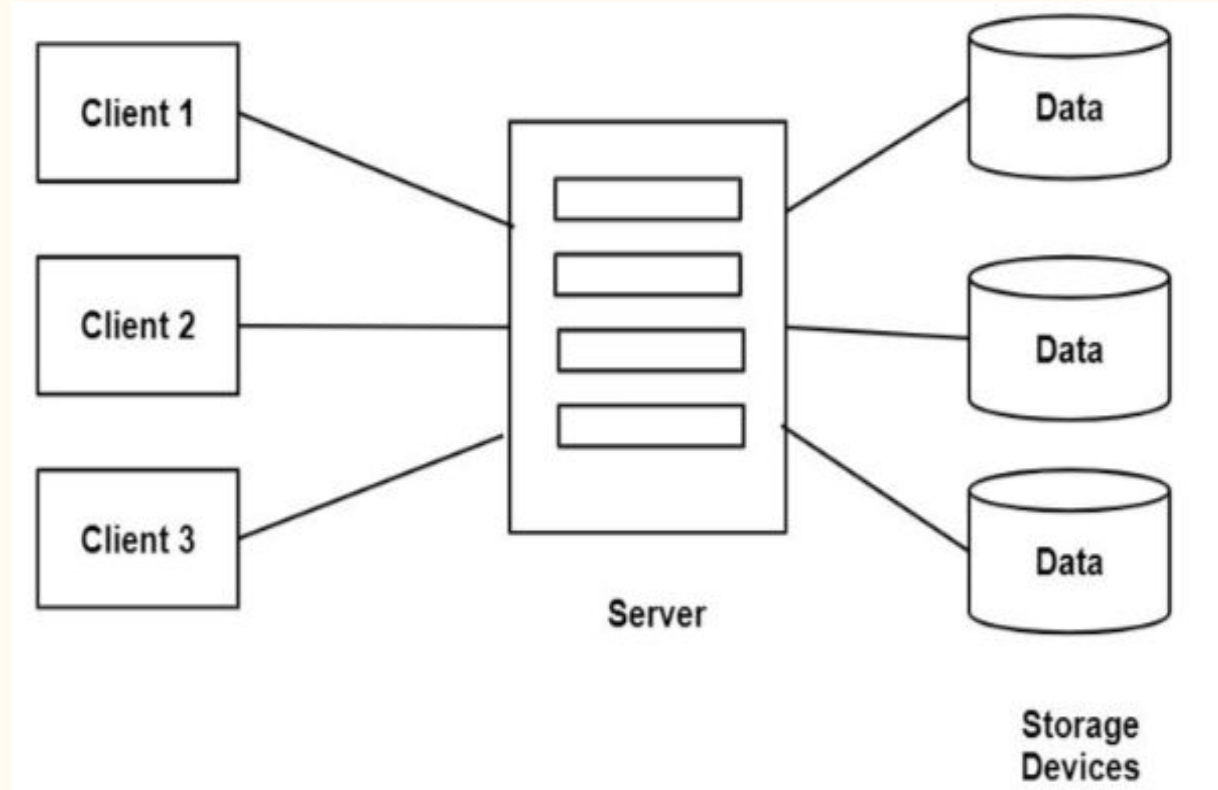
- Big Data solutions are ideal for analyzing not only raw structured data, but semistructured and unstructured data from a wide variety of sources.
- Big Data solutions are ideal when all, or most, of the data needs to be analyzed versus a sample of the data; or a sampling of data isn't nearly as effective as a larger set of data from which to derive analysis.
- Big Data solutions are ideal for iterative and exploratory analysis when business measures on data are not predetermined.

## Distributed file system

Distributed File System (DFS) is a file system that enables clients to access file storage from multiple hosts through a computer network as if the user was accessing local storage.

Files are spread across multiple storage servers and in multiple locations, which enables users to share data and storage resources.

If organizations need to scale up their infrastructure, they can add more storage nodes to the DFS.

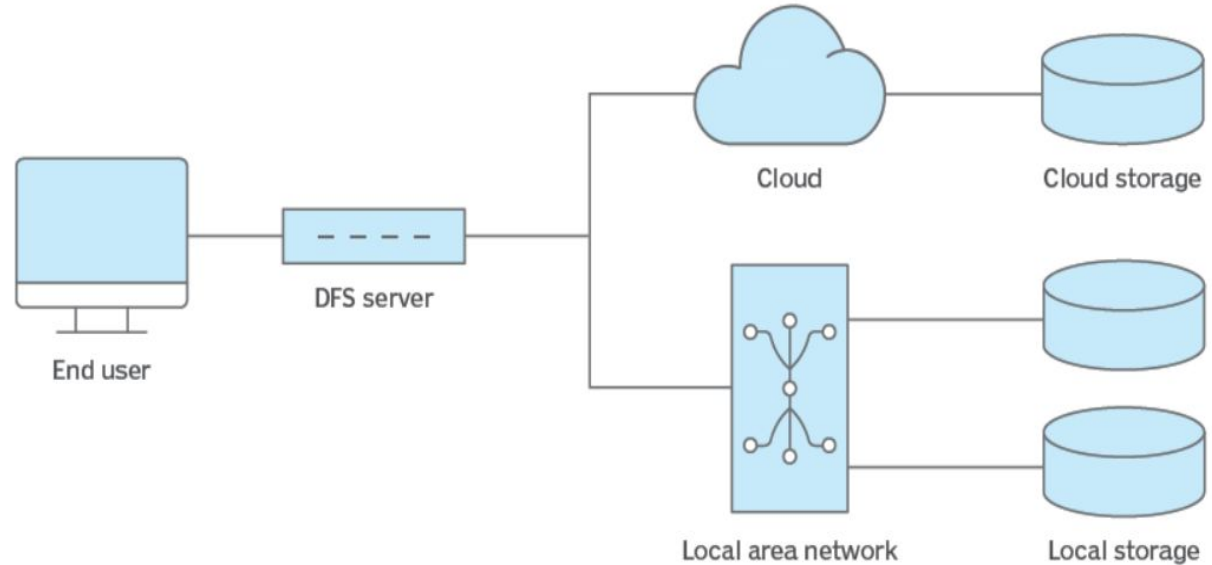


## Distributed file system

DFS clusters together multiple storage nodes and logically distributes data sets across multiple nodes that each have their own computing power and storage. The data on a DFS can reside on various types of storage devices.

DFS is located on a collection of servers, mainframes or a cloud environment over a local area network (LAN)

## Distributed file system architecture

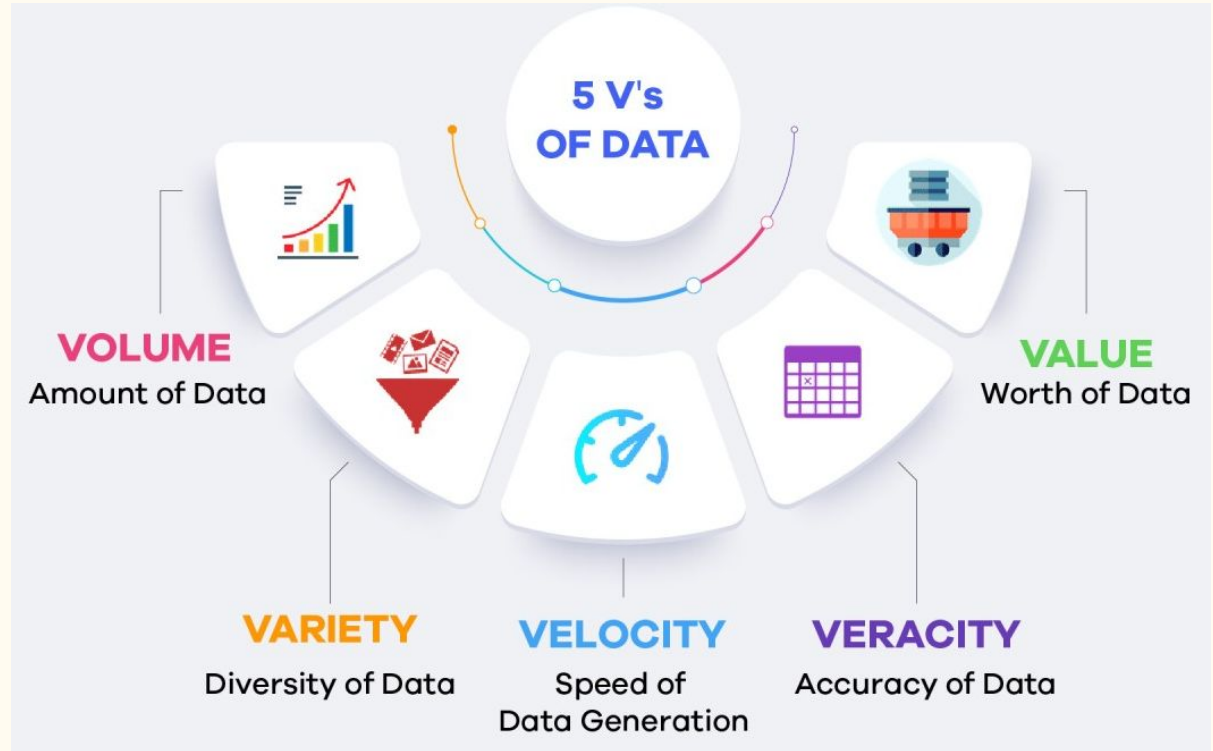


## Various Units of Memory - Data Size Storage

Name	Equal to:	Size in Bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte	1,024 bytes	1,024
Megabyte	1,024 kilobytes	1,048,576
Gigabyte	1,024 megabytes	1,073,741,824
Terrabyte	1,024 gigabytes	1,099,511,627,776
Petabyte	1,024 terrabytes	1,125,899,906,842,624
Exabyte	1,024 petabytes	1,152,921,504,606,846,976
Zettabyte	1,024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1,024 zettabytes	1,208,925,819,614,629,174,706,176

## Big Data characteristics - 5 V's of Big Data Drivers, types of Big Data

- **Volume-** Huge amount of data generated from various sources
- **Variety-** Different sources of data and its nature
- **Velocity-** speed at which data is generated and speed. Data processing is related to this increased speed of generation

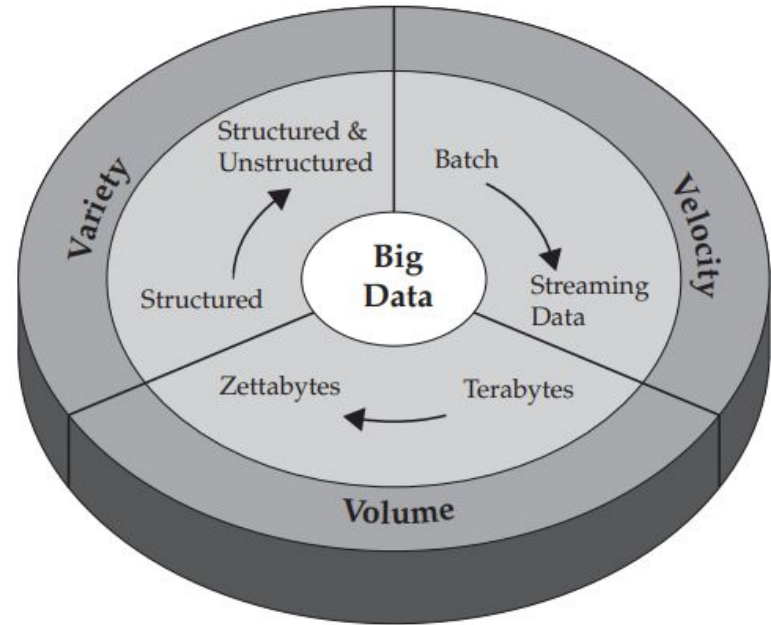




## Big Data characteristics - 5 V's of Big Data Drivers, types of Big Data

Characterised By IBM

- **Veracity**- It defines the degree of *trustworthiness* of the data. It denotes data inconsistency as well as data uncertainty.
- **Value**- has to be reliable and useful.
- volume is useful in determining whether a collection of data is Big Data or not.
- variety of data is crucial for its storage and analysis.



# Types of Big Data

## Structured

Pre-defined data models like databases

Usually text only

Easy to search and filter

Examples:  
Dates, phone numbers, transaction information



## Semi-Structured

Both structured & unstructured qualities

Considerably easier to analyze than unstructured data

Examples:  
Emails, CSV files, JSON files



## Unstructured

No Pre-defined data models

Difficult to search through

Usually stored as different types of files

Examples:  
Social media data, audio files, images



# Data growth

## 1 How much data is generated every minute?

Source: Domo

 **41,666,667**

messages shared  
by WhatsApp users

 **1,388,889**

video / voice calls made  
by people worldwide

 **404,444**

hours of video streamed  
by Netflix users

 **347,222**

stories posted by Instagram users

 **150,000**

messages shared by Facebook users

 **147,000**

photos shared by Facebook users

## 2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista



## 3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

 **2 TRILLION**

searches on Google by the end of 2021

 **1.134 TRILLION MB**

volume of data created every day

 **3,026,626**

emails sent every second, 67% of which are spam

 **278,108 PETABYTES**

global IP data per month by the end of 2021

 **230,000**

new malware versions created every day

 **82%**

share of video in total global internet  
traffic at the end of 2021

# Data growth

Category	Proportion of Internet Data Traffic
Video	53.72%
Social	12.69%
Gaming	9.86%
Web browsing	5.67%
Messaging	5.35%
Marketplace	4.54%
File sharing	3.74%
Cloud	2.73%
VPN	1.39%
Audio	0.31%

Year	Data Generated	Change Over Previous Year	Change Over Previous Year (%)
2020*	64.2 zettabytes	↑ 23.2 zettabytes	↑ 56.59%
2021*	79 zettabytes	↑ 14.8 zettabytes	↑ 23.05%
2022*	97 zettabytes	↑ 18 zettabytes	↑ 22.78%
2023*	120 zettabytes	↑ 23 zettabytes	↑ 23.71%
2024*	147 zettabytes	↑ 27 zettabytes	↑ 22.5%
2025*	181 zettabytes	↑ 34 zettabytes	↑ 23.13%

# Big Data Approach vs Traditional Data Systems

## Volume

Involves massive data volumes, often in the petabytes or more, generated at high speeds from various sources like sensors and social media.



Involves manageable data sizes generated relatively steadily from structured sources like databases.

## Variety

Comprises diverse data types, including structured, semi-structured, and unstructured data from various sources.



Primarily structured data with predefined formats suitable for conventional database Processing.

## Velocity

Big data is generated and processed at high speeds in real-time or near real-time.



Traditional data is usually processed at a slower pace.

# Big Data Approach vs Traditional Data Systems

## Storage and Infrastructure

Due to its large size and real-time processing needs require, scalable storage solutions, often using cloud-based platforms.



Traditional Data Can be stored using on-premises databases or storage solutions due to its manageable size.

## Processing Techniques

Utilizes distributed computing techniques like Hadoop and Spark to process data in parallel across clusters of computers.



It relies on traditional relational database management systems that process data using a single server.

## Tools and Technologies

Employs various tools, including Hadoop, Spark, NoSQL databases (MongoDB, Cassandra), and machine learning libraries.



Uses SQL-based databases (MySQL, Oracle), spreadsheets, and established data processing tools.

# Big Data Approach vs Traditional Data Systems

## Decision-Making Speed

Big Data enables real-time decision-making based on up-to-the-minute insights from diverse and dynamic data sources.



Traditional Data supports informed decisions using historical and structured data, often in less time-sensitive contexts.

## Privacy and Security

Requires advanced security measures due to the diversity of data sources and potential risks of unauthorized access and breaches.



While security is essential, it's often more standardized and relies on established practices.

## Costs and Resource Requirements

Involve higher costs due to the need for scalable infrastructure, specialized expertise, and the management of large datasets.



Generally more cost-effective as data volumes and processing complexities are lower.



# Big Data Approach vs Traditional Data Systems

## Use Cases and Industries

Applied in IoT, social media analysis, predictive analytics, and real-time monitoring.



Used in transactional systems, reporting, and applications with relatively low data volume and complexity.

## Scalability and Future Readiness

Big data systems are designed to handle massive data growth. They are built with flexibility, capable of integrating new data sources, adapting to emerging technologies, and accommodating evolving analytical methods.



Traditional data systems might struggle to handle the explosive growth of data, especially when faced with unanticipated increases in volume. They are often less agile in adapting to changing requirements.



## Case Study of Big Data Solutions Examples

- Retail:
  - Product Development
  - Customer Experience
  - Customer Lifetime Value
- Health Care:
  - Healthcare billing analytics
  - Claim Frauds
- Banking:
  - Fraud Detection
  - Business process optimization and automation

# Big Data Applications



## **Societal and Ethical issues associated with the use of big data analytics**

- Private customer data and identity should remain private
- Shared private information should be treated confidentially
- Customers should have a transparent view
- Big Data should not interfere with human will
- Big data should not institutionalize unfair biases

## **The key privacy issues or Big Data Challenges**

- Cybersecurity and Privacy
- Data Quality
- Data Storage & Processing
- Data Interpretation and Analysis
- Evaluating and selecting Big Data Technologies
- Lack of Knowledge Professionals
- Integrating data from multiple sources
- Ethical Issues.

END

—