Unit-1

BIG DATA:

Introduction:

Big data refers to large and complex data sets that can't be easily managed, processed, or analyzed using traditional data processing tools. It encompasses vast amount of structured, semi-structured, and unstructured data generated from various source such as social media, sensors, transactions, Machine generated data and more.

The growth of global data creation has been exponential over the past decades.

- By 2010, the total amount of data created worldwide reached approximately 2 zettabytes (ZB).
- This figure increased to 5 ZB in 2011 and 9 ZB in 2013.
- By 2020, the global data creation had surged to 64.2 ZB.
- In 2024, the total amount of data created, captured, copied, and consumed globally was approximately 147 to 149 zettabytes. This equates to around 402.74 million terabytes of data generated each day. Projections indicate that by the end of 2025, the total data creation will reach around 181 ZB.

		Factor binario
Bytes	B	$2^0 = 1$
KiloBytes	Kb	$2^{10} = 1024$
MegaBytes	Mb	$2^{20} = 1\,048\,576$
GigaBytes	Gb	$2^{30} = 1\,073\,741\,824$
TeraBytes	Tb	$2^{40} = 1\,099\,511\,627\,776$
PetaBytes	Pb	$2^{50} = 1\,125\,899\,906\,842\,624$
ExaBytes	Eb	$2^{60} = 1\,152\,921\,504\,606\,846\,976$
ZettaBytes	Zb	$2^{70} = 1180591620717411303424$
YottaBytes	Yb	$2^{80} = 1208925819614629174706176$

Advantages Of Big Data:

1. Business Growth
2. Improved Customer Services
3. better operational efficiency
4. Risk Management

Types of Big Data:

1. Structured Data:

Definition: Structured data is highly organized and follows a fixed schema with a preformatted data, typically stored in tabular formats with rows and columns. Each data element has a clear and defined meaning.

Examples:

- Customer records in a database (name, age, email, etc.)
- Bank transactions

2. unstructured Data:

Data that is not organized in a pre- defined record format is called unstructured data. It includes audio and video files graphics, text documents, social media posts, satellite images, etc.

Examples:

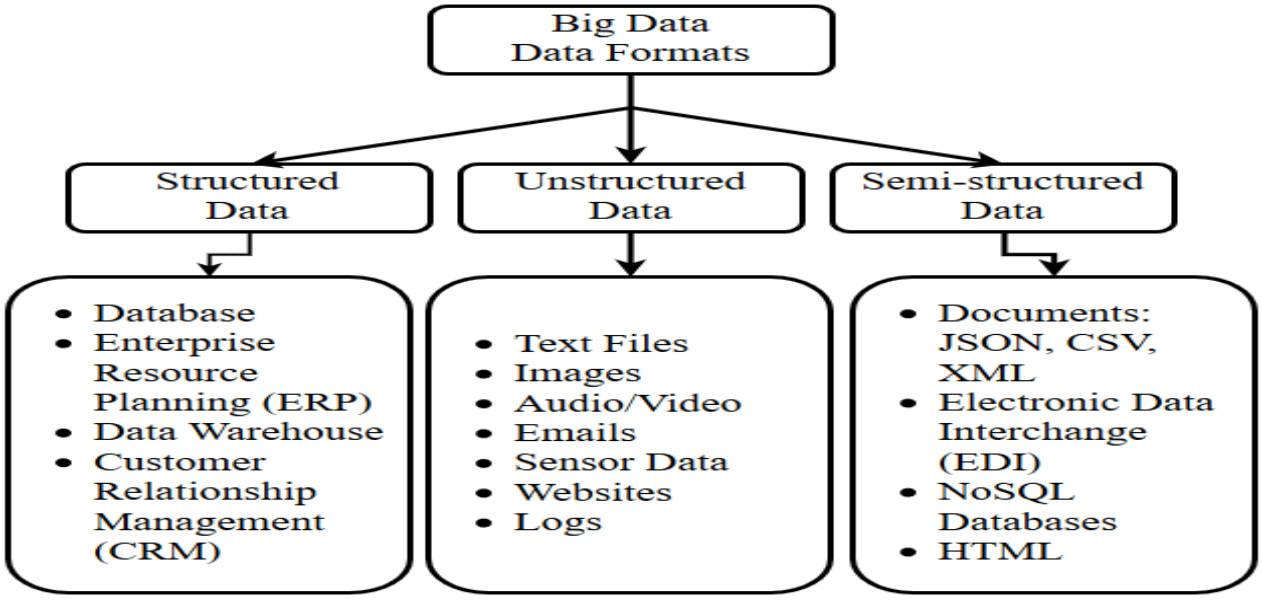
- Social media posts (tweets, comments, videos)
- Emails and messages
- Audio recordings and images

3. Semi- Structure:

Data that have no well-defined structure but maintain internal tags or markings to separate data elements are called semi-structured data. Include email document, HTML page etc.

Examples:

- XML and JSON files
- Log files from servers
- Sensor data from IoT devices



Note:

Example: The following JSON and XML examples both define an employees object, with an array of 3 employees:

JSON Example

```
{"employees":[
  { "firstName":"John", "lastName":"Doe" },
  { "firstName":"Anna", "lastName":"Smith" },
  { "firstName":"Peter", "lastName":"Jones" }
]}
```

XML Example

```
<employees>
  <employee>
    <firstName>John</firstName> <lastName>Doe</lastName>
  </employee>
  <employee>
    <firstName>Anna</firstName> <lastName>Smith</lastName>
  </employee>
  <employee>
    <firstName>Peter</firstName> <lastName>Jones</lastName>
  </employee>
</employees>
```

Characteristics of Big Data:

Big data is defined by five key characteristics, often called the 5V'S of Big Data:

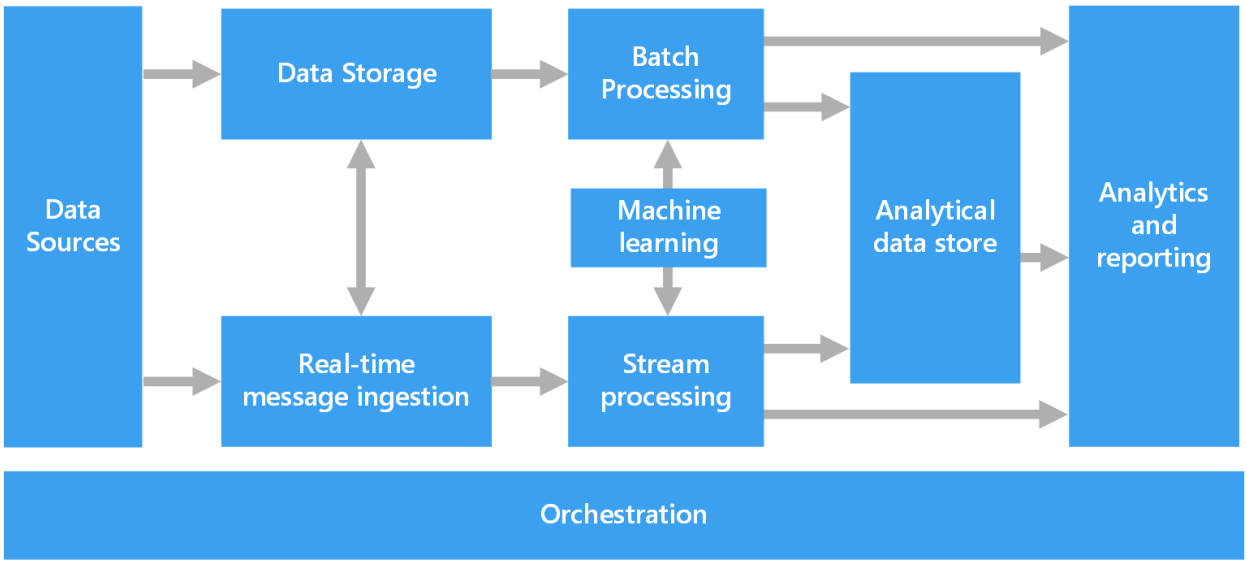
1. Volume:
   - ❖ refers to the massive amount of data generated every second.

- Data is collected from social media, IOT sensors, financial transaction etc.
- Example: As of 2025, Facebook generates approximately 5 petabytes (PB) of data daily.

2. Velocity:
   - Refers to the speed at which data is generated, collected, and processed.
   - Example: google processes Approximately 158,500 searches per second.
3. Variety:
   - Big data comes in different formats.
   - Structure
   - Unstructured
   - Semi-structured
4. Veracity
   - Not all data is correct; some data can be noisy, incomplete, or misleading.
   - Ensuring high quality, reliable data is essential for good decision making.
   - Example: fake news in social media vs verified news from government sources.
5. Value:
   - Extracting value from big data involves analyzing and interpreting the data to uncover patterns, trends, correlations, and actionable insights that can lead to better decision-making, innovation, and competitive advantage.
   - extracting useful information and insights to make better business decisions.
   - Example : Netflix use data to recommend personalized shows to user.

Big Data Architecture:

Big Data Architecture is the blueprint that outlines how large volumes of data are **collected, processed, stored, and accessed**. It includes all the components, tools, and processes required to handle data that is too large, fast, or complex for traditional systems.



Components of Big Data Architecture

1. Data Sources

- These are the origins of data, which could be:
  - Structured (e.g., databases)
  - Semi-structured (e.g., JSON, XML)
  - Unstructured (e.g., videos, images, social media, logs)

2. Data Storage:

Data storage is used to **store and manage large amounts of data** in a big data system.

- Big data includes structured, semi-structured, and unstructured data.
- Traditional databases may not handle big data well because they can't scale easily.

So, we use **distributed file systems** that can store huge amounts of data in different formats. This kind of storage is often called a **data lake**.

Examples:

- Azure Data Lake Storage

3. Real-time Message Ingestion:

Real-time message ingestion in big data involves capturing and processing data as it is generated, like from sensors, logs, social media, or IoT devices. This system helps businesses quickly gather insights, detect issues, and take immediate action. It's essential for handling fast-moving data and making timely decisions.

4.Batch Processing:

Batch processing is the method of **processing large volumes of data in groups (batches)** at scheduled times — **not in real-time**.

Key Features:

- Processes data in **bulk**, not one-by-one.
- Runs at **specific time intervals** (e.g., hourly, nightly).
- Handles **huge datasets** efficiently.
- Doesn't need **immediate output**.

5.Machine learning :

Machine learning uses prepared data (from batch or stream processing) to build models that predict outcomes or classify data. These models are trained on large datasets and can then be used to analyze new data and make predictions.

6.Stream Processing:

Stream Processing is a way of processing data continuously in real-time as it arrives — like a live stream of information. Instead of waiting for all data to arrive (like in batch processing), it handles each piece of data immediately when it is received

7.Analytical Data Store:

An Analytical Data Store (ADS) is a special database used in big data analytics to handle complex queries and large data sets. It supports tasks like data exploration, reporting, and advanced analysis, making it crucial for business intelligence and analytics.

8. **Analysis and Reporting:**

This is the part of the framework where software inspects the analyzed data for insights, patterns, and trends. Next, these results transfer to the reporting mechanism, preparing them for human viewing. You can then utilize this information to make more effective decisions for your business.

9. **Orchestration :**

Big data solutions usually involve repeating data processing steps in workflows. These steps include transforming data, moving it between different sources and destinations, loading the processed data into an analytical system, or directly sending the results to a report or dashboard.

Cloud Computing:

Cloud computing is a that allows users to store, manage and process data over the internet instead of using local computers or servers.

Example:

When you save files on google drive, watch movies, on Netflix, or send emails through Gmail, you are using cloud computing. The "Cloud" is just a network of powerful computers and servers located elsewhere that you can access online whenever you need.

Types of cloud computing services:

Cloud computing services are broadly categorized into three main types based on the level of control, flexibility, and management required:

1. software as a Service (Saas)
2. platform as Service (Paas)

3. Infrastructure as a service (IaaS)

1. **Software as a service (saas):**

It is a way to use software over the internet without installing or maintaining it on your computer. You simply access it via a web browser, saving a web browser, saving time and costs on hardware and software management. Saas is often called web-based on-demand software and works on a pay-as-you-go basis. It is generally used by end users.

Advantage:

1. cost effective

2. Reduced time

3. Accessibility.

2. Paas (platform as a service)

- ❖ It provide a platform & environment. (ie runtime environment) to allow developers to build applications & services over the internet
- ❖ paas Service are hosted in the cloud & accessed by users Via web browsers.
- ❖ It include operating systems, development tools, and databases, so developers don't need to worry about infrastructure.
- ❖ Example:

(i) Google App Engine - Build and deploy web application.

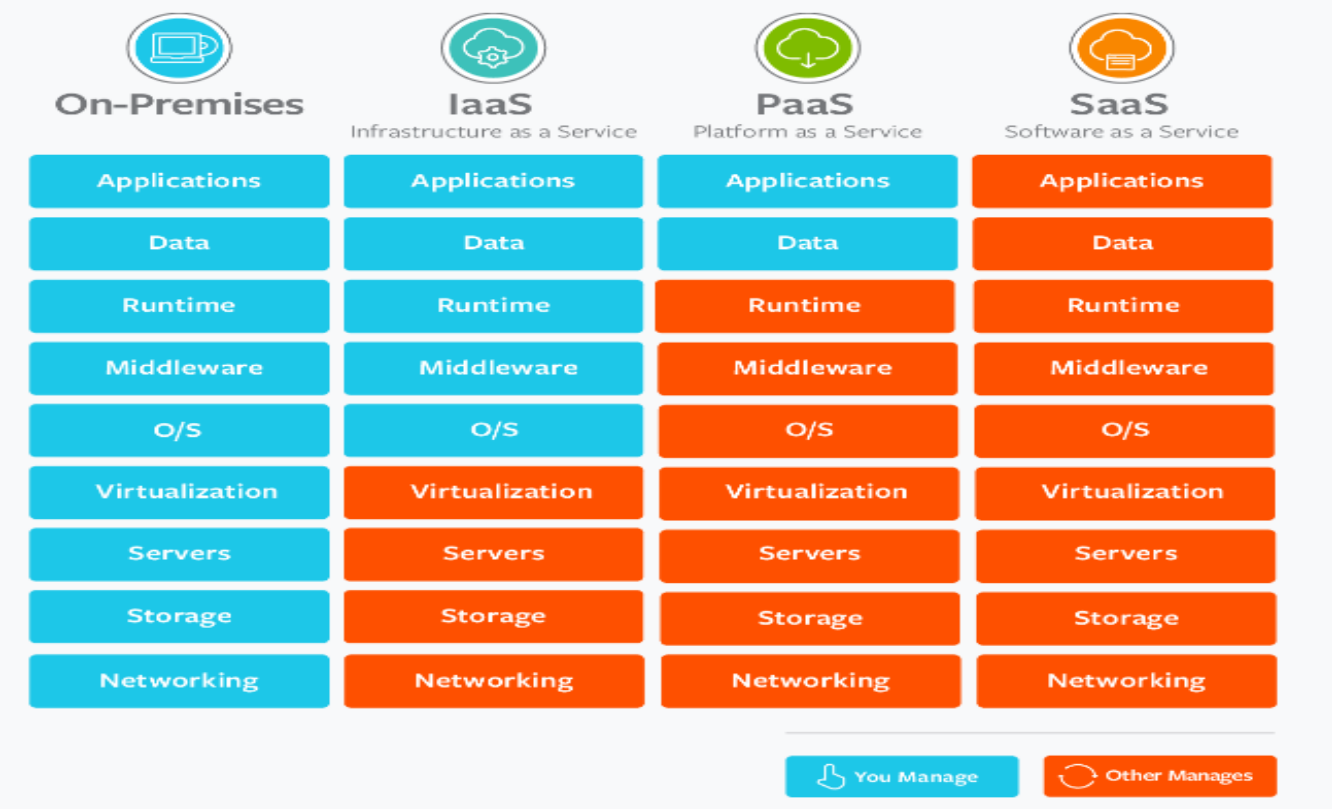
(ii) micro soft Azure APP Services-develop cloud based application.

3. infrastructure as a Service (Iaas):

- ❖ Iaas provides computing infrastructure (virtual machines, storage, and networks) on demand via the Cloud.

- you don't need to buy and maintain physical servers, hard drives, or network Cables.

Example: Imagine you need a powerful computer for a sort time. Instead of buying one, you rent a virtual computer or google Cloud and access it over the internet.



Types of Cloud Computing.

Cloud Computing is divided into four main types based on how the cloud services are deployed.

1. Public Cloud:

- A cloud environment available to the general public over the internet.
- Providers like Amazon web services (AWS), Microsoft Azure, and google cloud platforms (GCP) own and manage the infrastructure.

Features:

- cost effective- No need to buy hardware pay only for what you use.
- Scalable: Easily increase or decrease resource.
- Ex: Google drive, Microsoft one Drive.

2. Private Cloud: -

- cloud infrastructure is dedicated to a Single organization.

Features:

- **High security & privacy:** Resources are not Shared with other users.
- **Better performance:** since resources are not shared, performance is optimized.
- **Example:** Banking Cloud Services.

3. Hybrid Cloud:

- A mix of public and private clouds, allowing data to move between them.

Features:

- Keeps sensitive data in private cloud and scalable applications in public cloud.
- **cost-effective:** use private cloud for critical operations and public cloud for general task.
  Example: (i) Netflix - uses a private cloud for user data and a public cloud for video streaming.
  (ii) Hospitals: store Patient records in a private cloud and use Public Cloud for appointments and billing.

4. Community Cloud :-

- A cloud environment is shared between multiple organizations with similar needs.
  Features:

- **shared infrastructure:** used by organizations with Common goals (eg bank, hospitals, universities).
- **Cost Sharing:** cost is divided among the Participating organizations.
- **Example: (i) Government Cloud Services: -** Shared between different government agencies.
  (ii) Healthcare Cloud: used by hospitals and research labs to store patient data securely.

Big data used in distributed system:

Big data is extensively used in distributed systems to process and analyze massive volumes of data efficiently. Here's how big data integrates with distributed System.

1. storage and management:

- **Distributed file systems (DFS):** Big data is Stored in distributed file systems like Hadoop distributed file System (HDFS), which split large data sets across multiple nodes.
- **NOSQL databases:** System like Cassandra, MongoDB, and HBase, distributed data across multiple servers to ensure scalability and fault tolerance.

2.Processing and computation.

- **parallel processing:** frameworks like Apache Hadoop MapReduce divide task into small chunks and process them in parallel across different nodes.
- **Real time processing:** System like Apache Flink and Spark and Apache Storm process Streaming data in real time, essential for applications like Fraud detection and recommendation systems.

3. Scalability and fault Tolerance:

- Distributed system allows horizontal scaling (adding more machines) to handle growing data volume.

- ❖ Replication: data is replaced across nodes (e.g Cassandra) to prevent data loss and ensure availability.

4. Cloud-Based Big Data Solution:

- ❖ **Cloud Service:** platforms like AWS, Google Cloud, and Microsoft Azure offer distributed Big Data Services like Big Query, Azure Data Lake, and Amazone Redshift.

5. Applications of Big Data in Distributed systems:

- ❖ **social media analytics**: platforms like Facebook and twitter use distributed systems for processing petabytes of user-generated data.
- ❖ **Health Care:** Distributed Big Data system analyzes Patient records across multiple hospitals.

Development of big Data:

The term "big data" has been in use since the early 1990's. while it is unclear who first used it, John R. Mashey, a researcher at silicon Graphics, is often credited with popularizing the term.

In recent years,

- This figure increased to 5 ZB in 2011 and 9 ZB in 2013.
- By 2020, the global data creation had surged to 64.2 ZB.
- In 2024, the total amount of data created, captured, copied, and consumed globally was approximately 147 to 149 zettabytes. This equates to around 402.74 million terabytes of data generated each day. Projections indicate that by the end of 2025, the total data creation will reach around 181 ZB.

1. Early Data processing (1960s-1980s)

- Data Processing was limited to structured data in relational databases.
- storage was expensive, limiting the scale of data collection.

2. Growth of the internet and Digitalization (1990s-2000)

- The rise of the internet led to an explosion of unstructured data (emails, web pages, images, etc.)
- By Google introduced MapReduce (2004), revolutionizing large scale data processing.

3. The Era of big Data (2006s – 2020s)

- The emergence of Hadoop (2006) and its ecosystem (HDFS, Hive, pig, Spark) enabled large-scale data storage and processing.
- cloud computing platform (AWS, Azure, Google cloud) facilitated scalable data storage and analytics.

4 Modern Big Data Trends (2020s and beyond):

- **Real time processing:** Technologies like Apache Kafka and Apache Flink enable real-time data streaming.
- **Edge Computing & IOT:** Device generates massive amounts of data processed Closer to the source.
- **AI Integration:** Advanced AI models analyze Big Data for automation and decision-making