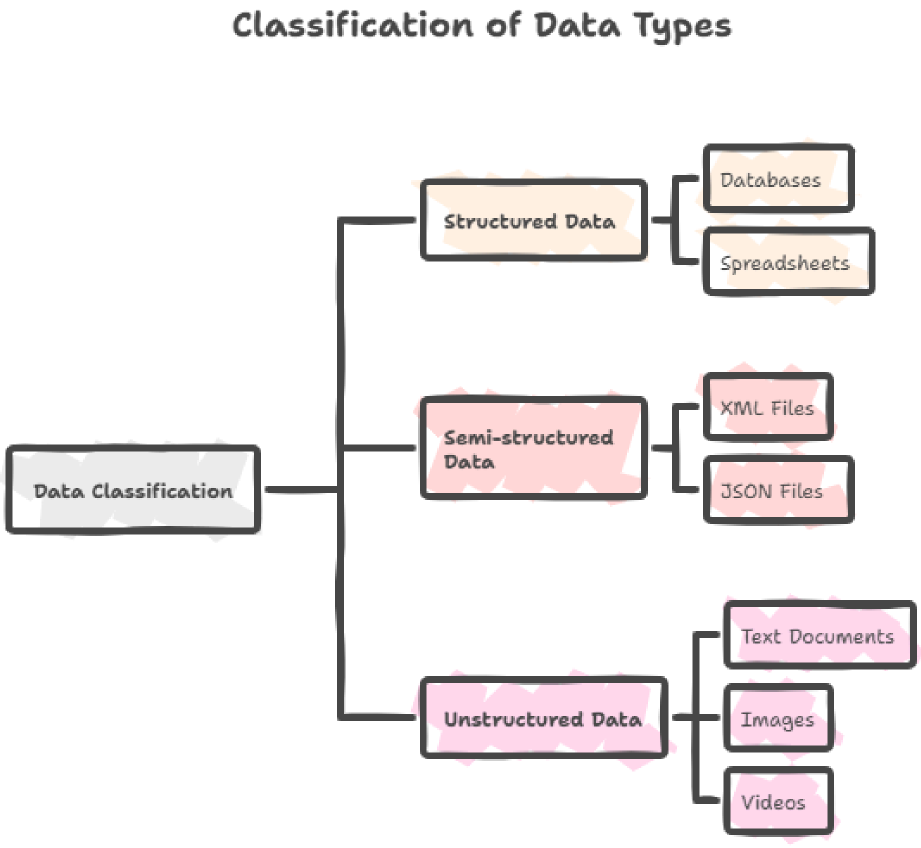
Unit - I Introduction to Big Data Analytics and Data Architecture

1.1 Classification of Data : Structured, Semi-structured and Unstructured

In the era of Big Data, data is generated from multiple sources — social media, sensors, IoT devices, etc. This data can be categorized based on its requirement.

The three major types of data are:

1. Structured Data
2. Semi-structured Data
3. Unstructured Data



1. Structured Data

Structured data is data that is organized in a format — typically stored in tables (rows and columns) or database systems (like RDBMS).

Characteristics:

- Highly organized and formatted.
- Stored in relational databases (SQL-based).
- Easy to search using SQL queries.

Examples:

- Employee database (Name, ID, Salary, Department)
- Sensor readings stored in tables

Tools/Technologies:

- RDBMS (MySQL, Oracle)

Format:

ID	Name	Age
1	John	28

2. Semi-structured Data

Semi-structured data does not follow tabular format but contains elements like tags, markers, or key-value pairs that make it easier to analyze than unstructured data.

Characteristics:

- Has a **flexible schema**.
- Data is self-describing (uses tags or identifiers).

- Not easily stored in traditional relational databases.
- Often stored in NoSQL databases.

Examples:

- JSON (JavaScript Object Notation)
- XML (eXtensible Markup Language)
- Email (has structured headers + unstructured body)
- Web logs

Tools/Technologies:

- NoSQL databases (MongoDB, Cassandra)
- Hadoop ecosystem (HDFS, Hive)

Format:

```
{
 "id": 1,
 "name": "John",
 "designation": ["Developer", "Analyst"]
}
```

3. Unstructured Data

Unstructured data is data that does not have a predefined structure. It cannot be easily organized into rows and columns and is typically text-heavy or multimedia in nature.

Characteristics:

- No fixed schema or format.
- Difficult to store in traditional databases.

- Requires special tools for processing and analysis (like NLP, text mining).

Examples:

- Text documents (PDFs, Word files)
- Images, videos, audio files
- Social media posts, emails, chat messages
- Sensor and IoT data streams

Tools/Technologies:

- Hadoop, Spark

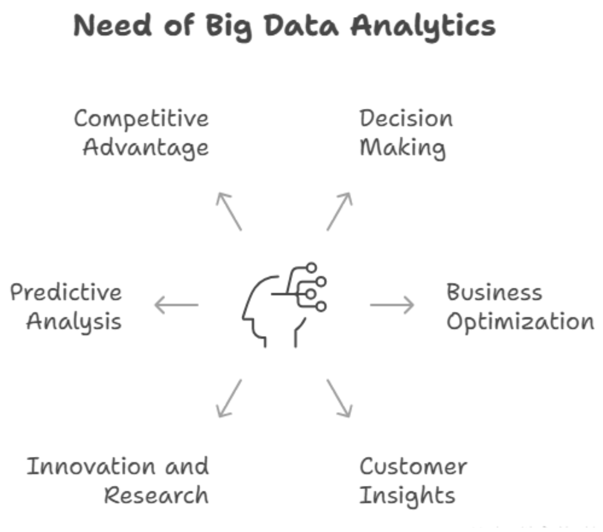
1.2 Introduction :Big Data Definitions, Need of Big Data

Definition of Big Data

Big Data is the ability to derive value from extremely large datasets.

Need for Big Data

1. Decision Making
2. Business Optimization
3. Customer Insights
4. Innovation and Research
5. Predictive Analysis
6. Competitive Advantage



1.3 Big Data Characteristics : Volume, Velocity, Variety, Veracity

1. Volume

- M**assive amount of data** generated from various sources such as social media, sensors, transactions, and devices.
- The scale of data can range from terabytes to petabytes and beyond.
- **Example:** Facebook generates over 4 petabytes of data per day from user activity.

2. Velocity

- Refers to the **speed at which data is generated, processed, and analyzed**.
- Modern systems require real-time or near-real-time processing to derive insights quickly.
- **Example:** Stock market data, online transaction records, or streaming services like Netflix require rapid data processing.

3. Variety

- Refers to the **different types of data** available for analysis.
- Data can be:
  - **Structured** (e.g., databases, spreadsheets)
  - **Unstructured** (e.g., emails, videos, social media posts)
  - **Semi-structured** (e.g., XML, JSON files)
- Handling diverse data types requires advanced tools and techniques.

4. Veracity

- Refers to the **trustworthiness, quality, and accuracy** of data.
- High veracity data ensures that the insights derived are reliable and meaningful.
- Challenges include **incomplete, inconsistent, or noisy data**.

- **Example:** Social media opinions or sensor readings may be inaccurate and need careful validation.

1.4 Big Data Types

Batch Data

- Collected and processed over time
- Example: Monthly sales reports

Real-Time Data

- Generated and processed instantly
- Example: Stock market data

Historical Data

- Stored past data used for analysis
- Example: Old customer records

Streaming Data

- Continuous flow of data
- Example: IoT sensor data

1.5 Big Data Processing Architecture Design

Data Sources

- Generate data
- Examples: sensors, social media, databases

Real-Time Message Ingestion

- Collects streaming data
- Tools: Kafka, Flume

Data Storage

- Stores large-scale data
- Examples: HDFS, cloud storage, NoSQL

Batch Processing

- Processes data in batches
- Tools: Hadoop, Spark

Stream Processing

- Processes data in real time
- Tools: Spark Streaming, Flink

Analytical Data Store

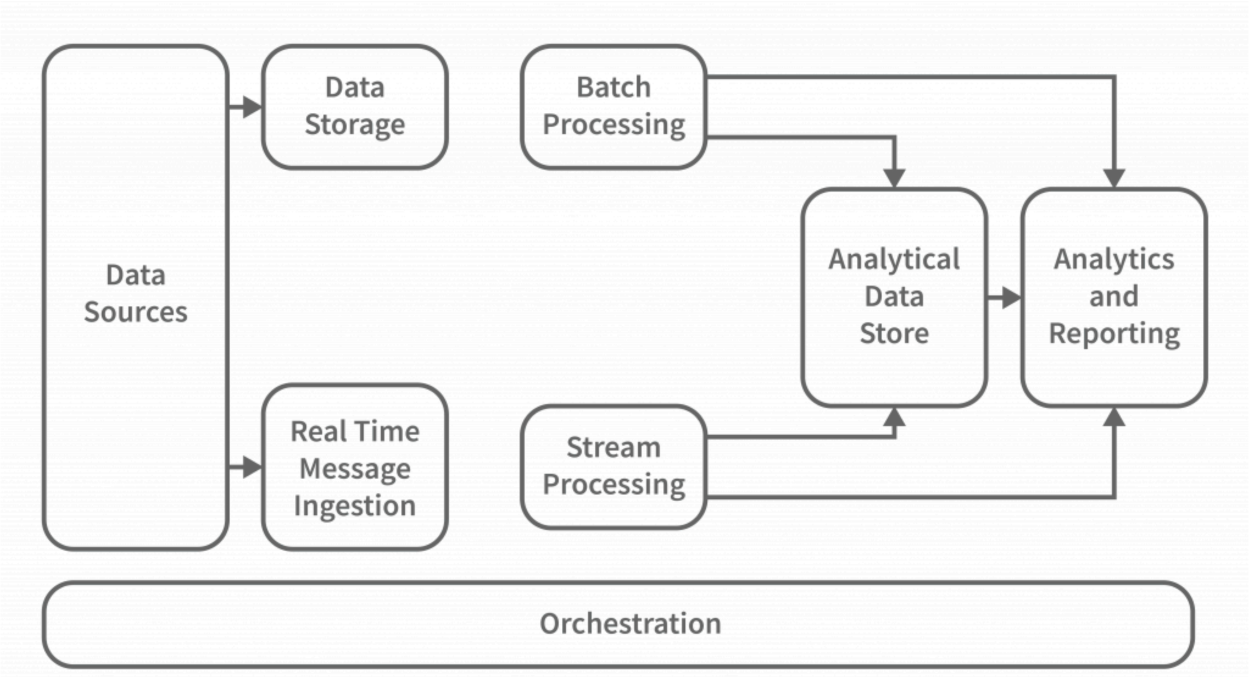
- Stores processed data for analysis
- Examples: Data warehouse, OLAP

Analytics and Reporting

- Analyzes data and shows results
- Tools: Tableau, Power BI

Orchestration

- Manages and schedules workflows
- Tools: Airflow, Oozie

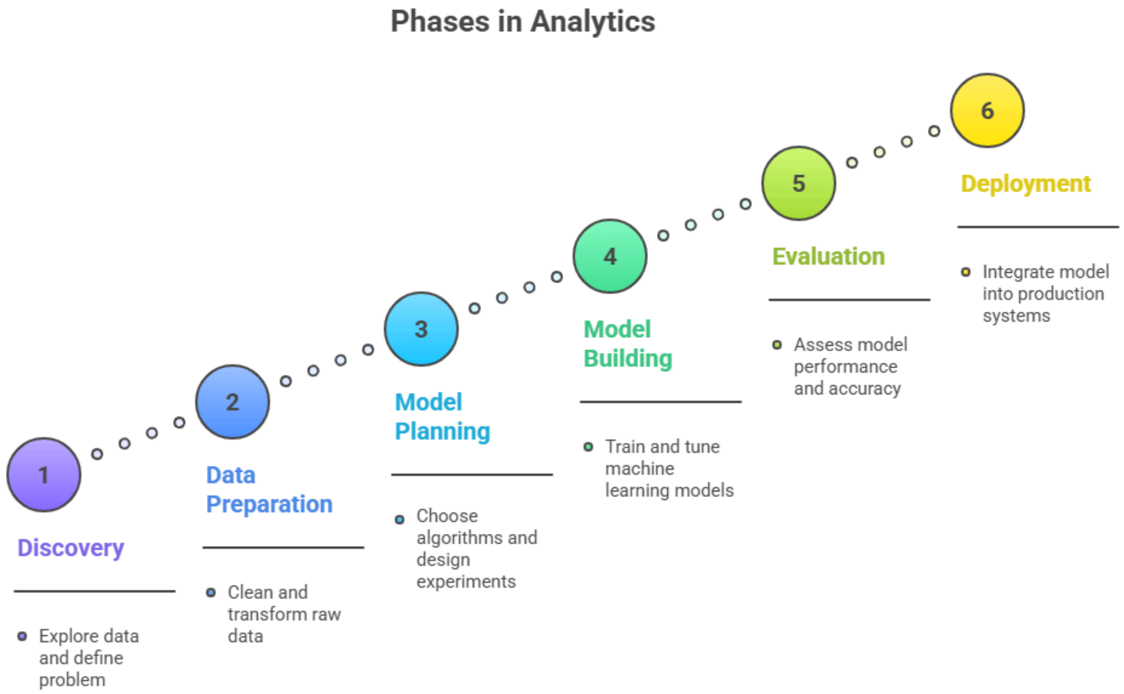


1.6 Big Data Analytics : Data analytics Definitions, Phases in Analytics

Data Analytics: Data analytics is the process of extracting, analyzing, and categorizing data to identify trends and patterns, with the aim of improving decision-making.

Phases in Analytics

1. **Discovery** – Define the problem and objectives.
2. **Data Preparation** – Collect, clean, and organize data.
3. **Model Planning** – Choose methods and tools for analysis.
4. **Model Building** – Develop and test analytical models.
5. **Evaluation** – Interpret results and validate insights.
6. **Deployment** – Apply results, monitor, and refine.


Phases in Analytics
1
Discovery
Explore data and define problem
2
Data Preparation
Clean and transform raw data
3
Model Planning
Choose algorithms and design experiments
4
Model Building
Train and tune machine learning models
5
Evaluation
Assess model performance and accuracy
6
Deployment
Integrate model into production systems

1.7 Big Data Analytics Applications : Big Data in Marketing and Sales, Big Data and Healthcare, Big Data in Medicine, Big Data in Advertising

1. Marketing & Sales – Understands customer behavior, predicts trends, personalizes marketing, and boosts sales.
2. Healthcare – Analyzes patient data, improves diagnosis, predicts outbreaks, and enhances hospital efficiency.
3. Medicine – Enables precision medicine, aids drug discovery, and improves treatment outcomes.
4. Advertising – Delivers targeted ads, optimizes placements, measures performance, and increases ROI.

Questions from previous exams:

1. Define Big data, Define Big data analytics.
2. State the various raw data sources
3. Explain Real-time analytics
4. State the significance of Big Data Analytics
5. Explain the challenges with Big Data
6. Explain analytics flow for Big data
7. Explain terminologies used in the Big Data Environment.
8. Describe Mapping analytics flow to Big data stack.
9. Describe case study of Weather Data Analysis.
10. List the various domain specific of Big Data & explain any one with example of Big Data
11. Classify the analytics process.

12. Describe data preparation of Big Data Analytics.

13. Describe the characteristics of Data.

14. Explain any Four analytics patterns.

15. Describe classification of Big Data Analytics.