
UNIT 1 — Introduction to Big Data & Hadoop

BDA

Open-source framework to process large amounts of data.

Big Data Analytics → Field of Big Data (2005)

“Data is like crude oil” — analogy (raw data → valuable insights).

Analytics Types

Analytics refers to systematic computational analysis of data.

4 Types

1. **Descriptive Analytics** — What happened?
2. **Diagnostic Analytics** — Why did it happen?
3. **Predictive Analytics** — What will happen?
4. **Prescriptive Analytics** — What should we do?

Types of Digital Data

1) Structured Data

- Organized in rows & columns
- Standard format
- Easy to store
- Example: Relational databases, Excel

---

2) Semi-Structured Data

- Doesn't follow strict schema
- Has tags / markers
- Examples: XML, JSON

---

3) Unstructured Data

- No predefined format
- Examples: Text, Images, Videos
- Any type of raw data formats

Big Data

Extremely large datasets that cannot be processed using traditional tools.

Characteristics — 5 V's

1. **Volume** — Huge amount of data
2. **Velocity** — Speed of data generation & processing
3. **Variety** — Different data types (structured / unstructured)
4. **Veracity** — Data accuracy & reliability
5. **Value** — Extract meaningful insights

Applications of Big Data

- Healthcare
- Finance
- Retail
- Social Media
- IoT
- Space Research

Apache Hadoop Framework

Core Components

1) HDFS (Hadoop Distributed File System)

- Stores large files
- Distributed storage system

2) MapReduce

- Programming model for distributed data processing

3) YARN (Yet Another Resource Negotiator)

- Resource management
- Job scheduling

RDBMS vs Big Data

RDBMS

- Small data
- Structured
- Stored in rows & columns

Big Data

- All data types
- Large scale
- Uses Hadoop

HDFS Architecture

- Large data split & stored across nodes
- Master Node → NameNode

- Slave Nodes → DataNodes
- Data stored in local disks

Block Size Example

- 128 MB per block (HDFS)
- Oracle example → 64 KB

Cluster

- Group of similar data nodes
- Adds extra storage blocks
- DataNodes placed together

Replication in HDFS

- Replication factor: **3**
- Copies of data stored for reliability

Propagation Delay

Propagation delay = Distance / Velocity

Flow of Read Operation

1. Client accesses NameNode
2. NameNode searches nearest DataNode
3. Based on:
    - Distance
    - Speed
    - Processing
    - Easy access

4. Data block returned to client

Write Operation

- Changes made in one DataNode
- Replicated to other copies
- Parallel updates occur

FSImage & EditLog

FSImage

- Stores metadata snapshot
- DataNode locations & replicas

EditLog

- Stores recent changes
- Used during write operations

MapReduce (Data Processing)

Steps

1. **Mapping / Splitting**
    - Convert to key-value format
    - Example: (Key, Value)
2. **Shuffle & Sorting**
    - Organize ascending / descending
3. **Reduce**
    - Aggregates results
    - Represents count directly

Functions in MapReduce

1) Mapper (Transformation)

- Transfers data from one form to another
- Performs splitting

2) Reducer (Aggregation)

- Combines values
- Produces final output

Hadoop Streaming Flow

Input Reader / Format
→ Key-Value Pairs
→ Mapper Stream
→ Intermediate Key-Value
→ Reduce Stream
→ Output Format

Key-Value Pair Rule

- Key → Always unique
- Values → Can be multiple

SQL Commands

DDL (Data Definition Language)

- Create
- Alter
- Drop
- Truncate

DML (Data Manipulation Language)

- Insert
- Update
- Delete
- Select
