1. Introduction to Apache Spark

Hadoop vs Spark

Apache Hadoop and Apache Spark are both used for **Big Data processing**, but they differ in architecture, speed, and ease of use.

Feature	Hadoop (MapReduce)	Apache Spark
Processing Speed	Slow (disk-based processing)	Fast (in-memory processing)
Ease of Use	Complex, requires extensive coding	Simple, supports multiple languages
Processing Modes	Batch processing only	Batch, streaming, SQL, ML, and graph processing
Fault Tolerance	Replication-based	RDD lineage-based recovery
Resource Management	Uses YARN	Can use YARN, Kubernetes, or Mesos

HDFS (Hadoop Distributed File System)

- A distributed file system that stores data across multiple nodes.
- Stores large datasets efficiently and allows parallel processing.

YARN (Yet Another Resource Negotiator)

- Manages cluster resources efficiently.
- **Alternatives:** Docker and Kubernetes for containerised environments.

2. Limitations of MapReduce

MapReduce is **not suitable for real-time or interactive processing** due to several reasons:

1. High Latency

- Writes intermediate data to disk, increasing **I/O overhead**.
- **Slow processing** due to disk dependency.
- Inefficient for **real-time analytics**.

2. Complex & Boilerplate Code

- Requires **separate Mapper and Reducer classes**.
- Hard to maintain and debug.

3. Only Supports Batch Processing

- **Cannot handle streaming data.**
- Inefficient for **real-time fraud detection, log processing, or monitoring.**

4. Rigid Execution Flow

- Strict **Map → Reduce** flow.
- **Difficult to implement custom workflows.**

5. No Interactive Mode & Limited Job Monitoring

- Jobs must **fully execute** before results are visible.
- **Debugging & optimization are difficult.**

3. Apache Spark: Overview & Features

What is Apache Spark?

- **An open-source distributed computing system** for processing large-scale data efficiently.
- **Much faster than Hadoop** due to **in-memory processing**.

Processing Capabilities

- **Batch Processing** – Like MapReduce but much faster.
- **Real-time Streaming** – Processes real-time data.
- **SQL Queries** – Query structured data efficiently.

- **Machine Learning (MLlib)** – Built-in ML support.
- **Graph Processing (GraphX)** – For graph-based computations.

Characteristics of Apache Spark

Feature	Description
In-Memory Processing	Reduces disk I/O, increasing speed.
Ease of Use	Supports **Scala, Python, Java, R**.
Unified Framework	Handles **batch, streaming, ML, and graph processing**.
Storage Flexibility	Works with **HDFS, Amazon S3, NoSQL, SQL databases**.