Hadoop Architecture, HDFS Features & YARN vs MapReduce – Detailed Answers

1. Architecture of Hadoop

Hadoop follows a master–slave architecture designed for distributed storage and processing of large datasets. It consists of two major layers: Hadoop Distributed File System (HDFS) for storage and YARN for resource management and job scheduling.

In the storage layer, HDFS stores data across multiple nodes using a distributed file system. In the processing layer, Hadoop uses MapReduce or other processing frameworks to process large data in parallel.

The master nodes manage the cluster, while slave nodes store data and perform computation. This architecture provides scalability, fault tolerance, and high availability.

2. Features of Hadoop HDFS

- Distributed Storage – Data is stored across multiple machines.
- Fault Tolerance – Data is replicated automatically.
- High Throughput – Optimized for large data access.
- Scalability – Supports thousands of nodes.
- Data Locality – Processing happens near data storage.
- Cost Effective – Uses commodity hardware.
- Reliability – Ensures data availability during failures.

3. HDFS Components

- NameNode – Manages metadata and file system namespace.
- DataNode – Stores actual data blocks.
- Secondary NameNode – Performs checkpointing.
- Standby NameNode – Provides high availability.
- JournalNode – Stores edit logs.
- Zookeeper – Coordinates HA services.

4. Difference between YARN and MapReduce

YARN (Yet Another Resource Negotiator) is a cluster resource management system, while MapReduce is a programming model used for data processing.

- YARN manages cluster resources; MapReduce processes data.
- YARN supports multiple processing frameworks; MapReduce supports only batch processing.
- YARN improves scalability; MapReduce has limited scalability.
- YARN separates resource management from processing logic.

- MapReduce is slower for real-time processing.

5. Notes

A. Data Reliability

Data reliability in Hadoop ensures that data remains accurate and accessible even during failures. HDFS achieves reliability through replication and continuous monitoring of data blocks.

B. Replication

Replication is the process of storing multiple copies of data blocks across different DataNodes. This ensures data availability and fault tolerance in case of node failure.

C. Fault Tolerance

Fault tolerance allows Hadoop to continue operating even when hardware or network failures occur. HDFS automatically detects failures and recovers data using replicated blocks.