Hadoop JobTracker, TaskTracker, Installation & Configuration – Detailed Answers

1. Difference between JobTracker and TaskTracker

JobTracker and TaskTracker are components of Hadoop MapReduce version 1 (MRv1). They work together to process data in a distributed environment.

The JobTracker is the master daemon responsible for managing MapReduce jobs. It receives job requests from clients, splits jobs into tasks, schedules tasks, monitors execution, and handles failures.

The TaskTracker is a slave daemon that runs on worker nodes. It executes the individual map and reduce tasks assigned by the JobTracker and reports progress back to the JobTracker.

Key Differences:
• JobTracker runs on the master node, TaskTracker runs on slave nodes.
• JobTracker schedules and monitors jobs, TaskTracker executes tasks.
• JobTracker handles failures, TaskTracker reports failures.
• JobTracker manages resources indirectly, TaskTracker manages task execution locally.

2. Steps to Install Hadoop

Installing Hadoop involves setting up Java, configuring environment variables, and configuring Hadoop configuration files. Below are the general steps:

- Install Java (JDK) and verify using java -version.
- Download Hadoop from the official Apache website.
- Extract the Hadoop tar file.
- Configure environment variables in ~/.bashrc.
- Edit Hadoop configuration files (core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml).
- Configure SSH for password-less login.
- Format the NameNode.
- Start Hadoop services (start-dfs.sh and start-yarn.sh).
- Verify Hadoop installation using jps command.

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. Resource Manager

The ResourceManager is a core component of YARN (Yet Another Resource Negotiator). It is responsible for managing cluster resources and scheduling applications.

It allocates resources to applications, monitors NodeManagers, and ensures efficient utilization of CPU and memory across the cluster.

5. Function of ~/.bashrc File

The ~/.bashrc file is a shell configuration file executed whenever a new terminal session is started. It is used to define environment variables, aliases, and shell settings.

In Hadoop installation, ~/.bashrc is used to set JAVA_HOME, HADOOP_HOME, and update the PATH variable so Hadoop commands can be executed from anywhere.