Day-3 Hadoop

08 May 2024 09:50

➢ What is Hadoop

- Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume.

- Hadoop is written in Java and is not OLAP (online analytical processing)

- It is used for batch/offline processing.

- It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more.

- Moreover it can be scaled up just by adding nodes in the cluster.

➢ Modules of Hadoop

Modules of Hadoop

1. **HDFS:** Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
2. **Yarn:** Yet another Resource Negotiator is used for job scheduling and manage the cluster.
3. **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
4. **Hadoop Common:** These Java libraries are used to start Hadoop and are used by other Hadoop modules.

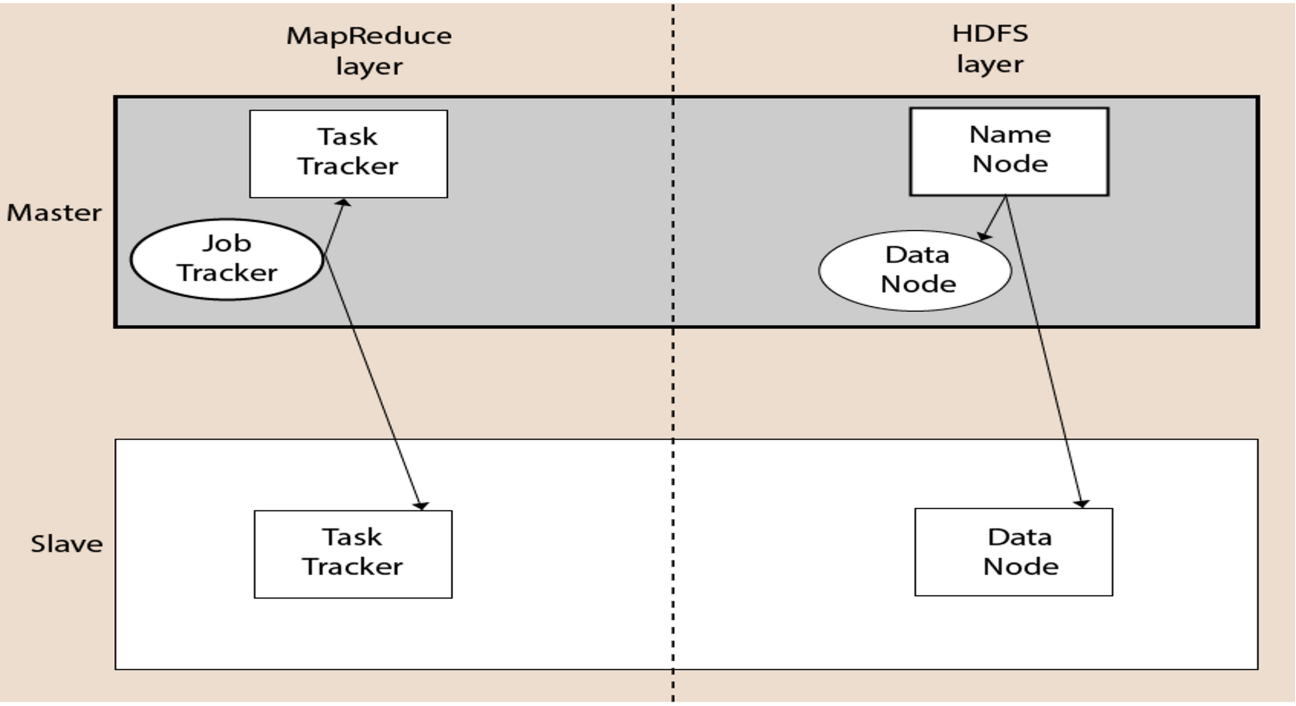
➢ Hadoop Architecture

- The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System).

- The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

- A Hadoop cluster consists of a single master and multiple slave nodes.

- The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.



➢ Hadoop Distributed File System

- The Hadoop Distributed File System (HDFS) is a distributed file system for Hadoop.

- It contains a master/slave architecture.

- This architecture consist of a single NameNode performs the role of master, and multiple DataNodes performs the role of a slave.

- Both NameNode and DataNode are capable enough to run on commodity machines.
- The Java language is used to develop HDFS. So any machine that supports Java language can easily run the NameNode and DataNode software.