JUNAID GIRKAR

60004190057

TE COMPS A4

**EXPERIMENT - 2**

**AIM**: Installation of Hadoop on a single node cluster.

**THEORY**:

# **What is Hadoop**

Hadoop is an open source framework from Apache and is used to store process and analyse 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.

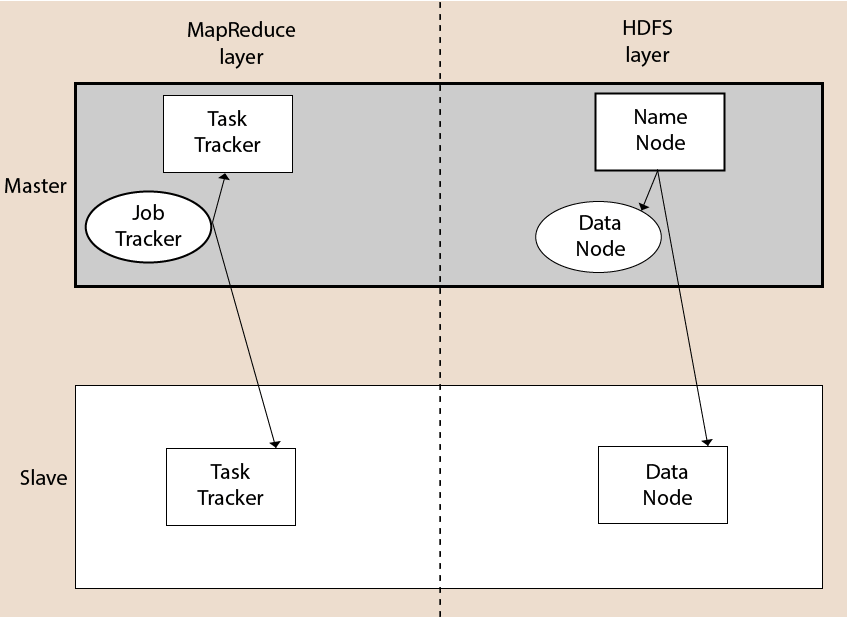
## **Modules of Hadoop**

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.

### **NameNode**

* It is a single master server exist in the HDFS cluster.
* As it is a single node, it may become the reason of single point failure.
* It manages the file system namespace by executing an operation like the opening, renaming and closing the files.
* It simplifies the architecture of the system.

### **DataNode**

* The HDFS cluster contains multiple DataNodes.
* Each DataNode contains multiple data blocks.
* These data blocks are used to store data.
* It is the responsibility of DataNode to read and write requests from the file system's clients.
* It performs block creation, deletion, and replication upon instruction from the NameNode.

### **Job Tracker**

* The role of Job Tracker is to accept the MapReduce jobs from client and process the data by using NameNode.
* In response, NameNode provides metadata to Job Tracker.

### **Task Tracker**

* It works as a slave node for Job Tracker.
* It receives task and code from Job Tracker and applies that code on the file. This process can also be called as a Mapper.

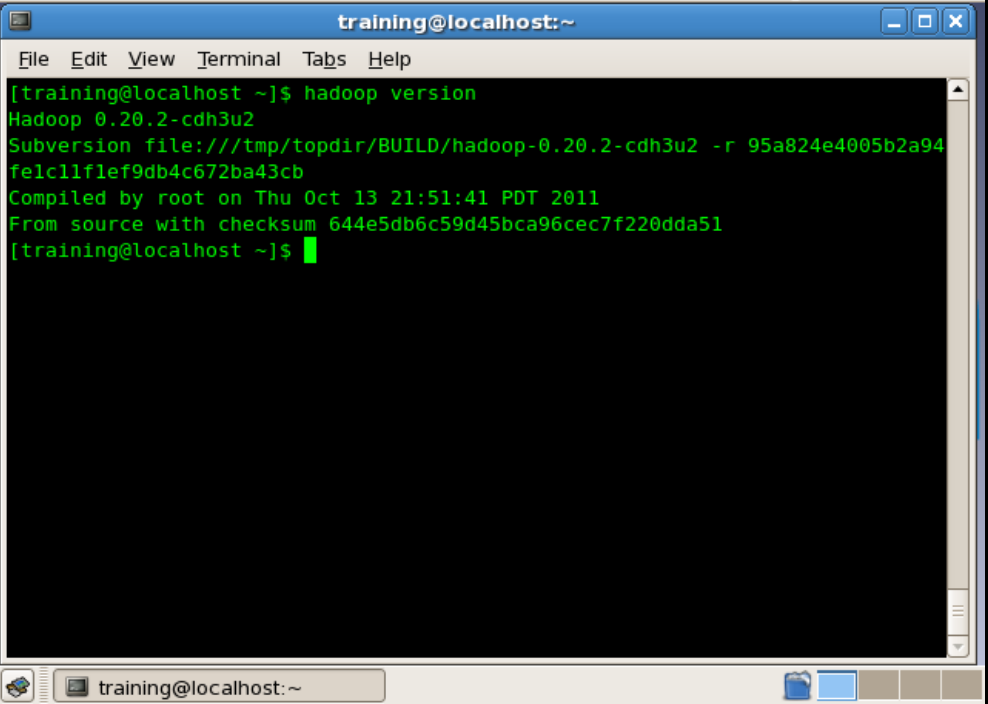
## **MapReduce Layer**

The MapReduce comes into existence when the client application submits the MapReduce job to Job Tracker. In response, the Job Tracker sends the request to the appropriate Task Trackers. Sometimes, the TaskTracker fails or time out. In such a case, that part of the job is rescheduled.

## **Advantages of Hadoop**

* **Fast:** In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
* **Scalable:** Hadoop cluster can be extended by just adding nodes in the cluster.
* **Cost Effective:** Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
* **Resilient to failure:** HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.

**IMPLEMENTATION**:



**CONCLUSION**: We learnt about big data tools like Hadoop and implemented it in cloudera.