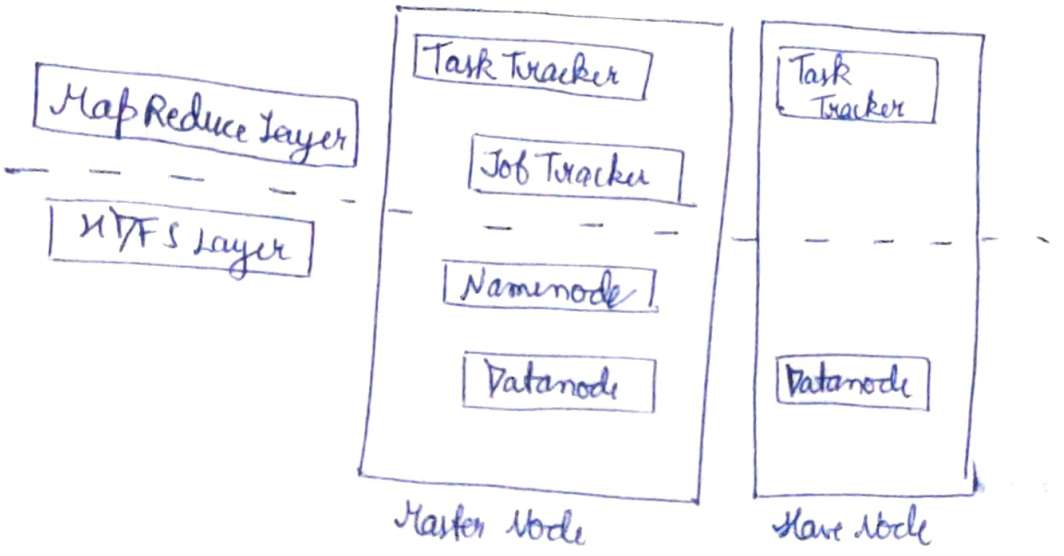
Apache Hadoop is an open source software framework used to develop data processing application which are executed in distributed computing environment.

It run on large data sets distributed across cluster of comodity computer.

It consist of two sub projects:-

- Hadoop Map Reduce : Map Reduce is a computational model and software framework for writing applications which are run on Hadoop. These programs are capable of processing enormous data in parallel on large clusters of computation nodes.



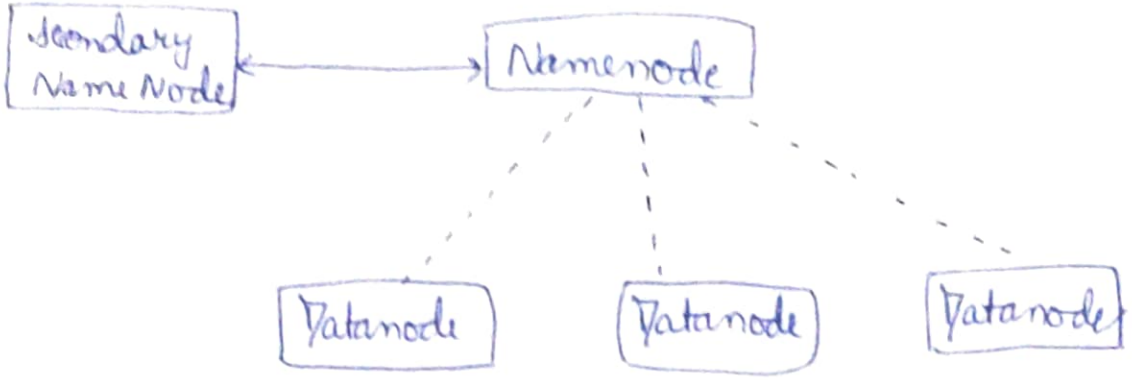
HDFS stands for Hadoop Distributed File System. It is the primary data storage system used by Hadoop application. It employs a namenode and datanode architecture to implement a distributed file system that provides high performance access to data across highly scalable Hadoop clusters. All data stored on Hadoop is stored in a distributed manner across a cluster of machine. The properties of HDFS are as follows:-

- Huge volume :- Being a distributed file system, it is highly capable of storing petabytes of data without any glitches.

- Data Access :- It is based on the philosophy that the most effective data processing pattern is write-once, the read many times pattern
- Cost-effative :- HDFS runs on a cluster of commodity hardware. These are the inexpensive machines that can be bought from any vendor.

The main components of HDFS are :-

- Name Node
- Data Node



Name Node :-

- It is a Master daemon
- It maintains and manages datanodes
- It records metadata eg location of blocks of data stored, the size of files, permission, etc.
- Receives heartbeat and block report from all the data nodes.

Data Node:-

- It is the slave daemon
- It stores the actual data
- It serves read and write request from the client.
- It is responsible for storing, replicating, deletion, etc of blocks.