Hadoop is an open-source java-based software framework sponsored by the Apache Software Foundation for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

It provides storage for big data at reasonable cost. Hadoop process big data in a single place as in a storage cluster doubling as a compute cluster.

Hadoop Architecture and Components:

Apache Hadoop consist of two major parts:

1. Hadoop Distributed File System (HDFS)
2. MapReduce

1. Hadoop Distributed File System:

HDFS is a file system or storage layer of Hadoop. It can store data and can handle very large amount of data.

When capacity of file is large then it is necessary to partition it. And the file systems manage the storage across a network of machine are called distributed file systems.

An HDFS cluster has two types of node operating in a master-worker pattern- Name Node and No. of Data Nodes.

Hadoop keep data safe by duplicating data across nodes.

2. MapReduce:

MapReduce is a programming framework. It organize multiple computers in a cluster in order to perform the calculations. It takes care of distributing the work between computers and putting results together.

Hadoop works in a Master-Worker / Master-slave fashion:-

1. Master:
Master contains Name node and Job tracker components.

1. Name node: It holds information about all the other nodes in the Hadoop Cluster, files in the cluster, blocks of files, their locations etc.
2. Job tracker: It keeps track of the individual tasks assigned to each of the nodes and coordinates the exchange of information and result.

2. Worker:
Worker contains Task tracker and Data node components.

1. Task Tracker: It is responsible for running the task assigned to it.
2. Data node: It is responsible for holding the data.

Other components of Hadoop architecture are : -Chukwa, Hive, HBase, Mahoutetc.

Characteristics of Hadoop:

1. Hadoop provides a reliable shared storage(HDFS) and analysis system (Map Reduce).
2. Hadoop is highly scalable. It can contain thousands of servers.

3. Hadoop works on the principles of write once and read multiple times.
4. Hadoop is highly flexible, can process both structured as well as unstructured data.