Hadoop is an open-source framework to store and process Big Data in a distributed environment. It contains two modules, one is Map Reduce and another is Hadoop Distributed File System (HDFS).

1.HDFS

Hadoop Distributed File System is a part of Hadoop framework, used to store and process the datasets. It provides a fault-tolerant file system to run on commodity hardware.

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

Features of HDFS

* It is suitable for the distributed storage and processing.
* Hadoop provides a command interface to interact with HDFS.
* The built-in servers of namenode and datanode help users to easily check the status of cluster.
* Streaming access to file system data.
* HDFS provides file permissions and authentication.

HDFS follows the master-slave architecture and it has the following elements.

### Namenode

The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is a software that can be run on commodity hardware. The system having the namenode acts as the master server and it does the following tasks:

* Manages the file system namespace.
* Regulates client’s access to files.
* It also executes file system operations such as renaming, closing, and opening files and directories.

### Datanode

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system.

* Datanodes perform read-write operations on the file systems, as per client request.
* They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

2. Hadoop Cluster

A [Hadoop](http://searchcloudcomputing.techtarget.com/definition/Hadoop) cluster is a special type of computational [cluster](http://searchexchange.techtarget.com/definition/cluster) designed specifically for storing and analyzing huge amounts of [unstructured data](http://searchbusinessanalytics.techtarget.com/definition/unstructured-data) in a [distributed computing](http://whatis.techtarget.com/definition/distributed-computing) environment.

Such clusters run Hadoop's [open sourc](http://searchenterpriselinux.techtarget.com/definition/open-source)e distributed processing software on low-cost [commodity computers](http://whatis.techtarget.com/definition/commodity-computer). Typically one machine in the cluster is designated as the NameNode and another machine the as JobTracker; these are the masters. The rest of the machines in the cluster act as both DataNode and TaskTracker; these are the slaves. Hadoop clusters are often referred to as "shared nothing" systems because the only thing that is shared between nodes is the network that connects them.

Hadoop clusters are known for boosting the speed of data analysis applications. They also are highly scalable: If a cluster's processing power is overwhelmed by growing volumes of [data](http://searchdatamanagement.techtarget.com/definition/data), additional cluster nodes can be added to increase throughput. Hadoop clusters also are highly resistant to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails.

As of early 2013, [Facebook](http://whatis.techtarget.com/definition/Facebook) was recognized as having the largest Hadoop cluster in the world. Other prominent users include [Google](http://searchcio-midmarket.techtarget.com/definition/Google), [Yahoo](http://whatis.techtarget.com/definition/Yahoo) and [IBM](http://searchitchannel.techtarget.com/definition/IBM-International-Business-Machines).

### 3. Block

Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block. The default block size is 64MB in 1.x version ,while 2.x versions the default size is 128mb , but it can be increased as per the need to change in HDFS configuration.