CLUSTER AND 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)

COMPONENTS OF HADOOP 1.X:

Hadoop 1.x Major Components components are: HDFS and MapReduce. They are also know as “Two Pillars” of Hadoop 1.x.

**HDFS:**  
HDFS is a Hadoop Distributed FileSystem, where our BigData is stored using Commodity Hardware. It is designed to work with Large DataSets with default block size is 64MB (We can change it as per our Project requirements).

HDFS component is again divided into two sub-components:

1. Name Node

Name Node is placed in Master Node. It used to store Meta Data about Data Nodes like “How many blocks are stored in Data Nodes, Which Data Nodes have data, Slave Node Details, Data Nodes locations, timestamps etc” .

1. Data Node

Data Nodes are places in Slave Nodes. It is used to store our Application Actual Data. It stores data in Data Slots of size 64MB by default.

**MapReduce:**  
MapReduce is a Distributed Data Processing or Batch Processing Programming Model. Like HDFS, MapReduce component also uses Commodity Hardware to process “High Volume of Variety of Data at High Velocity Rate” in a reliable and fault-tolerant manner.

MapReduce component is again divided into two sub-components:

1. Job Tracker

Job Tracker is used to assign MapReduce Tasks to Task Trackers in the Cluster of Nodes. Sometimes, it reassigns same tasks to other Task Trackers as previous Task Trackers are failed or shutdown scenarios.

Job Tracker maintains all the Task Trackers status like Up/running, Failed, Recovered etc.

1. Task Tracker

Task Tracker executes the Tasks which are assigned by Job Tracker and sends the status of those tasks to Job Tracker.