CLUSTER:

In a computer system, a cluster is a group of servers and other resources that act like a single system and enable high availability and, in some cases, load balancing and parallel processing.

HADOOP CLUSTER:

 A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment.

 Such clusters run Hadoop's open source distributed processing software on low-cost commodity computers.

 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.

 If a cluster's processing power is overwhelmed by growing volumes of data, additional cluster nodes can be added to increase throughput. Hence, Scalable

 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, termed as replication .

 The Master nodes oversee the two key functional pieces that make up Hadoop:

1.storing lots of data (HDFS)

2. running parallel computations on all that data (Map Reduce).

 The Name Node oversees and coordinates the data storage function (HDFS), while the Job Tracker oversees and coordinates the parallel processing of data using Map Reduce.

 Slave Nodes make up the vast majority of machines and do all the dirty work of storing the data and running the computations.

 Each slave runs both a Data Node and Task Tracker daemon that communicate with and receive instructions from their master nodes.

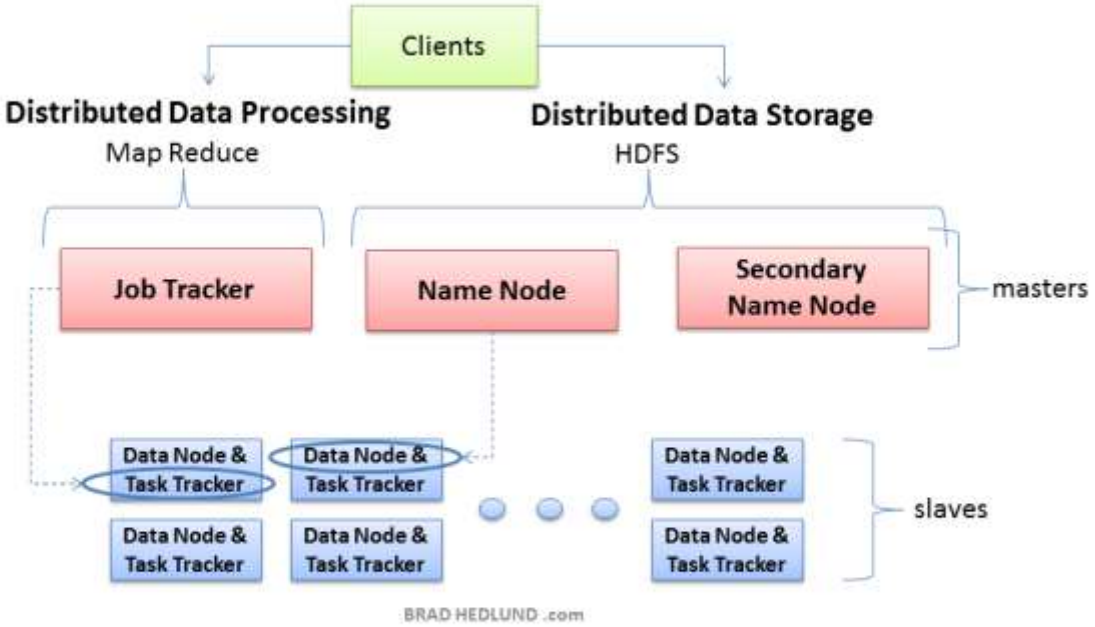
 The Task Tracker daemon is a slave to the Job Tracker, the Data Node daemon a slave to the Name Node.

DataNodes send heartbeat signal periodically to the NameNode. Thus, NameNode keeps track of all the DataNodes in the hadoop cluster.

Similarly, JobTracker assigns the task to several TaskTrackers running on the DataNodes. They in turn keep track of the process as TaskTrackers periodically sends signal.

If JobTracker detects a failure, it reassigns the task to another TaskTracker residing on another DataNode(THE NODE CLOSER TO THE FAILURE NODE).

Hadoop Server Roles



Hadoop Cluster