**1. Cluster and Hadoop cluster**

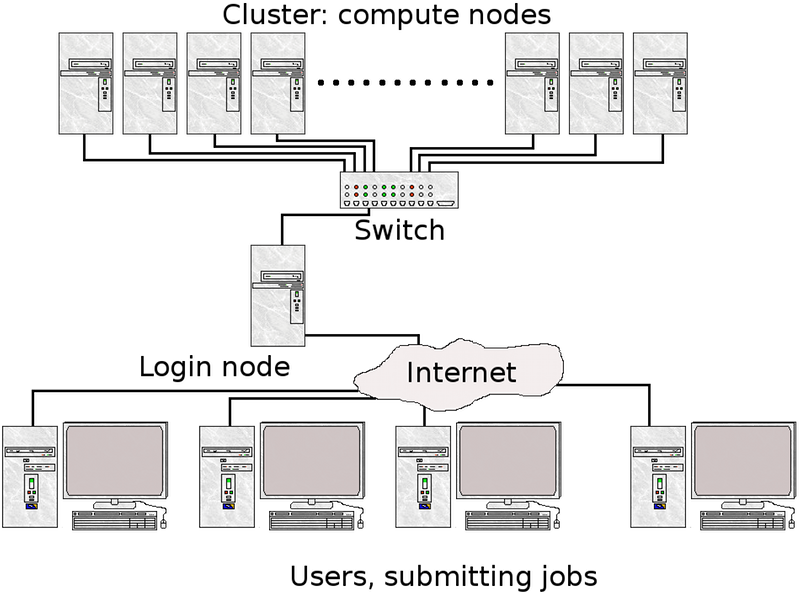
**Cluster:**

A ***cluster*** consists of a set of loosely or tightly connected computers that work together so that, in many respects, they can be viewed as a single system. Unlike grid computers, clustershave each node set to perform the same task, controlled and scheduled by software.

The components of a cluster are usually connected to each other through fast [local area networks](https://en.wikipedia.org/wiki/Local_area_network) ("LAN"), with each *node* (computer used as a server) running its own instance of an [operating system](https://en.wikipedia.org/wiki/Operating_system).

They are usually deployed to improve performance and availability over that of a single computer, while typically being much more cost-effective than single computers of comparable speed or availability.

Although a cluster may consist of just a few personal computers connected by a simple network, the cluster architecture may also be used to achieve very high levels of performance.



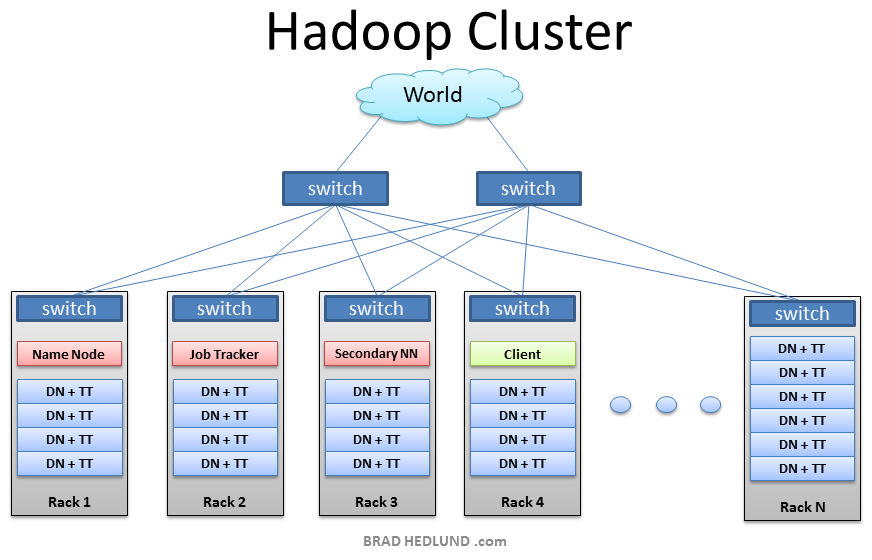
**Hadoop cluster:**

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.

Hadoop clusters are comprised of three different node types: master nodes, worker nodes, and client nodes. Understanding the different node types will help you plan your cluster, and configure the appropriate number and type of nodes when creating a cluster.

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 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.



**2. Rack and its arrangement in a Hadoop cluster:**

A **Hadoop** Cluster is a collection of racks. A **rack** is a collection of 30 or 40 nodes that are physically stored close together and are all connected to the same network switch. Network bandwidth between any two nodes in **rack** is greater than bandwidth between two nodes on different racks.

In large clusters of Hadoop, in order to improve network traffic while reading or writing HDFS files, NameNode chooses data nodes which are on the same rack or a nearby to read or write request.

NameNode achieves this rack information by maintaining rack ids of each data node. This concept of choosing closer data nodes based on racks information is called Rack Awareness in Hadoop.

NameNode tries to place replicas of block on multiple racks for improved fault tolerance. Hadoop lets the cluster administrators decide which rack a node belongs to through configuration variable net.topology.script.file.name. When this script is configured, each node runs the script to determine its rack id. A default installation assumes all the nodes belong to the same rack.

