Setting up of the Hadoop cluster:

Here you will learn how to successfully install Hadoop and configure the clusters which could range from just a couple of nodes to even tens of thousands over huge clusters. So for that, first you need to install Hadoop on a single machine. The requirement for that is you need to install Java if you don't have it already on your system.

Getting Hadoop to work on the entire cluster involves getting the required software on all the machines that are tied to the cluster. As per the norms one of the machines is associated with the Name Node and another is associated with the Resource Manager. The other services like The MapReduce Job History and the Web App Proxy Server can be hosted on specific machines or even on shared resources as per the requirement of the task or load. All the other nodes in the entire cluster with have the dual nature of being both the Node Manager and the Data Node. These are collectively termed as the slave nodes.

Getting Hadoop to work in the non-secure mode

The Java configuration of Hadoop has two important files:

- Read-only default configuration -core-default.xml, hdfs-default.xml, yarn-default.xml and mapred-default.xml.
- Site-specific configuration -etc/hadoop/core-site.xml, etc/hadoop/hdfs-site.xml, etc/hadoop/yarn-site.xml and etc/hadoop/mapred-site.xml.

It is possible to manage the Hadoop scripts in the bin/ directory of the distribution, by setting site-specific values via the etc/hadoop/hadoop-env.sh and etc/hadoop/yarn-env.sh.

For the Hadoop cluster configuration you first need to create the ecosystem in which the Hadoop daemons can execute and also the needed parameters for configuration.

The various daemons of Hadoop Distributed File System are listed below:

- NodeManager
- ResourceManager
- WebAppProxy
- NameNode
- SecondaryNameNode
- DataNode
- YARN daemons

The Hadoop Daemons configuration environment

To get the Hadoop daemons' the right site-specific customization the administrators need to use the etc/hadoop/hadoop-env.sh or the etc/hadoop/mapred-env.sh and etc/hadoop/yarn-env.sh scripts. The JAVA-HOME should be specified appropriately so that it is defined in the right manner on every remote node.

Configuration of the individual daemons

The list of Daemons along with the relevant environment variable

NameNode –HADOOP-NAMENODE-OPTS
DataNode – HADOOP-DATANODE-OPTS
Secondary NameNode – HADOOP-SECONDARYNAMENODE-OPTS
ResourceManager – YARN-RESOURCEMANAGER-OPTS
NodeManager – YARN-NODEMANAGER-OPTS
WebAppProxy – YARN-PROXYSERVER-OPTS
Map Reduce Job History Server – HADOOP-JOB-HISTORYSERVER-OPTS

Customization of other important configuration parameters:

- **HADOOP-PID-DIR** – the process ID files of the daemons is contained in this directory.
- **HADOOP-LOG-DIR** – the log files of the daemons are stored in this directory.
- **HADOOP-HEAPSIZE / YARN-HEAPSIZE** – the heapsize is measured in MB and if you have the variable that is set to 1000 then automatically the heap is also set to 1000 MB. By default it is set to 1000.