**Introduction to Hadoop**

* Hadoop Distributed File System
* Map Reduce Programming Model
* Hadoop Common Utilities

**Advantages of Hadoop**

* Accessible
* Robust
* Fault-tolerant
* Simple
* Reliable
* Scalable

**Hadoop Distributed File System**

This below diagram depicts the HDFS. This one example of loading data file into HDFS. Here the Name node is responsible for splitting the file into Blocks (configurable size 64 MB by default) and maintaining the Meta data information bout the these blocks. Name node directs all the data nodes to interact with clients to low level I/O (reading/writing) tasks. The default replication factor is 3 (This is also configurable). All the rectangle boxes are Data nodes.

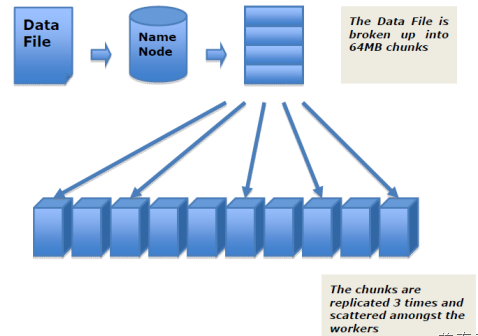


Illustration : Hadoop Distributed File System (HDFS)

**Map Reduce Programming Model:**

In Map Reduce programming Model, we have three phases

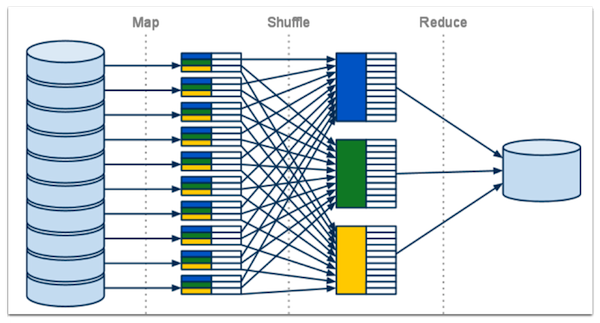
1. Map
2. Shuffle
3. Reduce

**Map:** In this phase, Data reading from input and applying filter and transformation. This is done by developer.

**Shuffle:** In this phase, All the map phase emitted data is copying, group and sorting is done. This is done by framework.

**Reduce:** In this phase, The aggregation of data is done. This is done by developer.

The following diagram depicts the flow of typical Map Reduce Program. In this diagram, we have several splits all are going to be individual map tasks (10 maps) , all maps emitting three colors. Once map process is done, in shuffle phase all the map results are going to be group and sorting and finally sending to three reducers. Here all the values of particular color is going to be the same reducer, finally aggregation process is happening on these values of every key. This applicable all map reduce programs.



**Some of the On line Resources for Hadoop**

* Apache Website and Wiki
* Cloudera
* Map R
* Horton works
* Yahoo Developer Networks
* IBM Developer Networks