**Pre-Requisites:**

* Any Linux Operating System (Ubuntu Preferable)
* Java (JDK)
* ssh
* rsync

**Installing java:**

**This is required for all three modes**

We can install java in two ways. One is using apt-get and another one download the java jdk from Oracle Website.

**Installing Java using apt-get:**

cmd> sudo apt-get update

cmd> sudo apt-cache search jdk (Will get java open jdk package)

Cmd> sudo apt-get install java-package

**Installing Java using downloaded package from Oracle:**

Download it from oracle website, choose either 32 bit or 64 bit based on your Computer Architecture.

Change the file permissions of downloaded java package

cmd> chmod -755 java package.bin

cmd> ./java-package.bin

The above command extracts all the java home directory

**Installing ssh:**

**This is required for Pseudo and Fully Distributed Modes**

cmd> sudo apt-get install ssh

Cmd> sudo service ssh start

**creating Passwordless ssh:**

Cmd> ssh-keygen -t rsa

It will ask password, here we have to press enter without entering our password

Copy the id\_rsa.pub to authorized\_keys

Cmd> cat id\_rsa.pub >> authorized\_keys

**Testing our ssh:**

Cmd> ssh username@hostname

**To get hostname:**

Cmd> hostname

To change hostname edit the file /etc/hostname file

To map ip address to dns name edit the file /etc/hosts file

**Steps for downloading and installing Hadoop:**

Download Apache Hadoop from hadoop.apache.org

Preferably hadoop-1.x.x.tar

Create bigdata under your home directory /home/username

Cmd> mkdir bigdata

Cmd> cd bigdata

**Downloading the Hadoop tar file apache site:**

Cmd> wget <http://apache.techartifact.com/mirror/hadoop/common/hadoop-1.0.4/hadoop-1.0.4.tar.gz>

Extract the tar file to hadoop-1.x.x directory

Cmd> tar -xvf hadoop-1.x.x.tar

Move extracted hadoop-1.x.x into the directory bigdata.

Cmd> mv hadoop-1.x.x bigdata/

**Change the directory permissions recursively to 755**

755 means owner has full permissions; group and rest of the world have only read and execute permissions

Cmd> chmod -R 755 hadoop-1.x.x

**Installation Modes:**

* **Local mode**
* **Pseudo Distributed Mode**
* **Distributed Mode**

In hadoop.1.x.x directory, we have one sub directory called conf

In conf directory, we have files:

* **hadoop-env.sh**
* **core-site.xml**
* **hdfs-site.xml**
* **mapred-site.xml**
* **masters**
* **slaves**

hadoop-env.sh --> For setting, hadoop environment variables

core-site.xml --> For setting, Hadoop cluster Information related configuration propertys

hdfs-site.xml --> For setting, HDFS related configuration propertys

mapered-site.xml --> For setting, Map Reduce related configuration propertys

slaves --> All domain names (IP info) of slave nodes (Data Node + Task Tracker)

masters --> Domain name of Secondary Name Node

Default content in these files is:

All the xml files are with empty configuration information

Both masters and slaves have hostname as localhost

In hadoop-env.sh --> All the default environment variables are configured.

**Local Mode:**

**In this mode, we use Local Linux file system as File System**

For running hadoop in local mode, only we have to modify hadoop-env.sh

In hadoop-env.sh, we have to Set JAVA\_HOME

Uncomment the JAVA\_HOME and replace java installation directory with our Java home:

**In Linux (Ubuntu) Location of java home:**

**/usr/lib/jvm/javapackage/**

**Pseudo distributed mode:**

For running Hadoop in pseudo distributed mode, we have to modify hadoop-env.sh.

In hadoop-env.sh, we have to Set JAVA\_HOME.

We have to modify the some important configuration propertys in core-site.xml, hdfs-site.xml, mapred-site.xml, slaves, and masters.

**Core-site.xml:**

<property>

<name>fs.default.name</name>

<value>hdfs://hostname:port</value>

</property>

**Mapred-site.xml:**

<property>

<name>mapred.job.tracker</name>

<value>hostname:port1</value>

</property>

**hdfs-site.xml:**

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.name.dir</name>

<value>path of namenode[namenode meta information directory] directory</value>

</property>

<property>

<name>dfs.data.dir</name>

<value>path of datanode[actual data location] directory</value>

</property>

**slaves:**

hostname

**masters:**

hostname

**Fully distributed mode:**

For running Hadoop in Fully distributed mode, we have to modify only slaves file.

We will use **Pseudo distributed configuration as it is**. For adding more slave machines we have to **modify conf/slaves file**. We will copy the entire **Hadoop Directory into other slave machines**. Moreover we have to **share the SSH public keys** of each machine. The **Absolute path of the Hadoop Home Directory has to same on all machines.**

Copy the entire hadoop-1.x.x directory to the same path in the slave machines like

in master /home/hadoop/bigdata/hadoop-1.x.x. The absolute path of hadoop-1.x.x is same on all machines.

**no change to core-site.xml**

**no change to mapred-site.xml**

For **hdfs-site.xml also changes are not required**. If we want more replication value, we can change the **dfs.replication** property.

<property>

<name>dfs.replication</name>

<value>replication factor</value>

</property>

**masters file:**

**On Master Node: Enter the secondary namenode machine hostname**

**On Slave Nodes: Empty the file**

**slaves file:**

**Master Node: Enter all the list of Slave Node machines hostnames**

slave1

slave2

.

slaven

**Slave Nodes: Empty the file**

***Please follow the above guide lines while installing in all modes.***