Prerequisites for Hadoop Multi Node Cluster Setup

1. Add Entries in hosts file

Edit hosts file and add entries of both master and slaves

```
sudo nano  /etc/hosts
MASTER-IP  master
SLAVE01-IP  slave01
SLAVE02-IP  slave02
```

(NOTE: In place of MASTER-IP, SLAVE01-IP, SLAVE02-IP put the value of the corresponding IP).

Example

```
192.168.1.190 master
192.168.1.191 slave01
192.168.1.195 slave02
```

2. Install Java 8 (Recommended Oracle Java)

Hadoop requires a working Java 1.5+ installation. However, using Java 8 is recommended for running Hadoop.

Command: sudo apt update

Command: sudo apt install openjdk-8-jdk

Command: export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

Command: echo $JAVA_HOME

3. Configure SSH

Hadoop requires SSH access to manage its nodes, i.e. remote machines plus your local machine if you want to use Hadoop on it.

3.1 Install Open SSH Server-Client

Command : sudo apt-get install openssh-server openssh-client

3.2 Generate KeyPairs

Command : ssh-keygen -t rsa -P ""

3.3 Configure password-less SSH

3.3.1 Copy the generated ssh key to master node’s authorized keys.

Command: cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys

3.3.2 Copy the master node’s ssh key to slave’s authorized keys.

Command:

```
ssh-copy-id -i $HOME/.ssh/id_rsa.pub cse@slave01
ssh-copy-id -i $HOME/.ssh/id_rsa.pub cse@slave02
```

3.4 Check by SSH to all the Slaves

```
ssh slave01
ssh slave02
```

2. Install Hadoop

1 Download Hadoop

```
wget http://archive.cloudera.com/cdh5/cdh/5/hadoop-2.5.0-cdh5.3.2.tar.gz
```

2 Untar Tar ball

Command : tar xzf hadoop-2.5.0-cdh5.3.2.tar.gz

3. Hadoop multi-node cluster setup Configuration

1 Edit .bashrc

Edit .bashrc file located in user’s home directory and add following parameters.

Command : nano .bashrc

```
export HADOOP_PREFIX="/home/cse/hadoop-2.5.0-cdh5.3.2"
export PATH=$PATH:$HADOOP_PREFIX/bin
export PATH=$PATH:$HADOOP_PREFIX/sbin
export HADOOP_MAPRED_HOME=${HADOOP_PREFIX}
export HADOOP_COMMON_HOME=${HADOOP_PREFIX}
export HADOOP_HDFS_HOME=${HADOOP_PREFIX}
export YARN_HOME=${HADOOP_PREFIX}
```

Command : source .bashrc

2 Edit hadoop-env.sh

hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc. Edit configuration file hadoop-env.sh (located in HADOOP_HOME/etc/hadoop) and set JAVA_HOME.

Command : nano hadoop–env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

3 Edit core-site.xml

core-site.xml informs Hadoop daemon where NameNode runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.

Edit configuration file core-site.xml (located in HADOOP_HOME/etc/hadoop) and add following entries.

Command : nano core-site.xml

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://master:9000</value>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/cse/hdata</value>
</property>
</configuration>
```

4 Edit hdfs-site.xml

hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.

Edit configuration file hdfs-site.xml (located in HADOOP_HOME/etc/hadoop) and add following entries

```
Command : nano  hdfs-site.xml
<configuration>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
</configuration>
```

5 Edit mapred-site.xml

mapred-site.xml contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template. Edit configuration file mapred-site.xml (located in HADOOP_HOME/ etc/hadoop) and add following entries

```
Command : nano mapred-site.xml
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

6 Edit yarn-site.xml

yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc. Edit configuration file mapred-site.xml (located in HADOOP_HOME/etc/hadoop) and add following entries

```
Command : nano yarn-site.xml
```

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
```

```
</property>
<property>
<name>yarn.resourcemanager.resource-tracker.address</name>
<value>master:8025</value>
</property>
<property>
<name>yarn.resourcemanager.scheduler.address</name>
<value>master:8030</value>
</property>
<property>
<name>yarn.resourcemanager.address</name>
<value>master:8040</value>
</property>
</configuration>
```

7. Edit salves

Edit configuration file slaves (located in HADOOP_HOME/etc/hadoop) and add following entries:

```
slave01
slave02
```

Now Hadoop is set up on Master, now setup Hadoop on all the Slaves.

Install Hadoop On Slaves

4. Setup Prerequisites on all the slaves

Run following steps on all the slaves

1. Add Entries in hosts file

2. Install Java 8 (Recommended Oracle Java)

5. Copy configured setups from master to all the slaves

5.1. Create tarball of configured setup

Command : tar czf hadoop.tar.gz hadoop-2.5.0-cdh5.3.2

(NOTE: Run this command on Master)

5.2. Copy the configured tarball on all the slaves

Command : scp hadoop.tar.gz slave01:~

(NOTE: Run this command on Master)

Command : scp hadoop.tar.gz slave02:~

(NOTE: Run this command on Master)

5.3. Un-tar configured Hadoop setup on all the slaves

Command : tar xvzf hadoop.tar.gz

(NOTE: Run this command on all the slaves)

Now Hadoop is set up on all the Slaves. Now Start the Cluster.

6. Start the Hadoop Cluster

Let us now learn how to start Hadoop cluster?

6.1. Format the name node

Command : bin/hdfs namenode -format

(Note: Run this command on Master)

(NOTE: This activity should be done once when you install Hadoop, else it will delete all the data from HDFS)

6.2. Start HDFS Services

Command : sbin/start-dfs.sh

(Note: Run this command on Master)

6.3. Start YARN Services

Command :sbin/start-yarn.sh

(Note: Run this command on Master)

6.4. Check for Hadoop services

6.4.1. Check daemons on Master

Command : jps

NameNode

ResourceManager

6.4.2. Check daemons on Slaves

Command : jps

DataNode

NodeManager

7. Stop The Hadoop Cluster

Let us now see how to stop the Hadoop cluster?

7.1. Stop YARN Services

Command : sbin/stop-yarn.sh

(Note: Run this command on Master)

7.2. Stop HDFS Services

Command : sbin/stop-dfs.sh

(Note: Run this command on Master)