Prerequisites to Deploy Hadoop on Single Node Cluster

Step 1: Install Java 8 (Recommended Oracle Java)

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

1.1 Install Python Software Properties

Command :
```
sudo apt-get install python-software-properties
```

1.2 Install Java Software properties

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
```

Step 2: 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.

2.1 Install Open SSH Server-Client

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

2.2 Generate KeyPairs

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

2.3 Configure password-less SSH

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

2.4 Check by SSH to localhost

Command :
```
ssh localhost
```

Step 3: Install Hadoop

3.1 Download Hadoop

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

Note:

You can download any version of hadoop version 2+. Here I am using CDH version is Cloudera's 100% open source platform distribution.

3.2 Untar Tar ball

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

Note

All the required jars, scripts, configuration files, etc. are available in HADOOP_HOME directory (hadoop-2.5.0-cdh5.3.2).

Step 4: Setup Configuration

4.1 Edit .bashrc

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

Command : nano .bashrc

```
export HADOOP_PREFIX="/home/hdadmin/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}
```

Note

After above step restart the terminal, so that all the environment variables will come into effect or execute the **source** command.

Command : source .bashrc

4.2 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
```

Note

Here your can change java path according to your java installation directory.

4.3 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://localhost:9000</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/home/hdadmin/hdata</value>
  </property>
</configuration>
```

Note

/home/hdadmin/hdata is a sample location; please specify a location where you have Read Write privileges.

4.4 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>1</value>
  </property>
</configuration>
```

4.5 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>
```

4.6 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

```
<?xml version="1.0" encoding="UTF-8"?>
<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>
</configuration>
```

Step 5: Start the Cluster

5.1 Format the name node:

Command : bin/hdfs namenode -format

Note:

This activity should be done once when you install hadoop, else It will delete all your data from HDFS.

5.2 Start HDFS Services

Command : sbin/start-dfs.sh

5.3 Start YARN Services

Command : sbin/start-yarn.sh

5.4 Check whether services have been started

To check that all the Hadoop services are up and running, run the below command.

Command : jps

```
NameNode
DataNode
ResourceManager
NodeManager
Jps
SecondaryNameNode
```

Step 6. Stop The Cluster

6.1 Stop HDFS Services

Command : sbin/stop-dfs.sh

6.2 Stop YARN Services

Command : sbin/stop-yarn.sh