INSTALLATION OF HADOOP IN UBUNTU 14.04 LTS

SINGLE NODE CONFIGURATION

Step 1.Installing Java

Java is the main prerequisite for Hadoop. Install java1.7.

```
$sudo apt-get update
$sudo apt-get install openjdk-7-jdk
$sudo apt-get install openjdk-7-jre
```

Then, you should verify the existence of java in your system using the command "java -version".

```
$java -version
```

Step 2. Set the Host Name as 'localhost' in /etc/hostname file

```
$sudo gedit /etc/hostname
```

Step 3. Set the Known Hosts in /etc/hosts file

```
$sudo gedit /etc/hosts
```

```
127.0.0.1 localhost
```

Restart the system& Login to student

Step 4 Create a dedicated user account for hadoop

```
$sudo addgroup Hadoop
$sudo adduser --ingroup hadoop hdptools
$sudo usermod -a -G sudo hdptools
$su hdptools
```

Restart the system & Login to hdptools

Step 5. Configure ssh

5.1. Generate private and public key pair at terminal using

```
$sudo apt-get install ssh
$ssh-keygen
```

5.2. To enable ssh to the local machine

```
$cat $HOME/.ssh/id_rsa.pub >>$HOME/.ssh/authorized_keys
$ssh localhost
```

Step 6.Disable IPV6 by including the following lines in /etc/sysctl.conf file

```
$sudonano /etc/sysctl.conf
```

```
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

Reboot the machine to make the changes and logon to hdptools

Step 7. To find the java path

```
$sudo update-alternatives --config javac
```

Step 8. Install Hadoop

```
$cd /usr/local
$sudo tar xvzf $HOME/Downloads/hadoop-2.5.1.tar.gz
$sudo chmod 777 hadoop-2.5.1
```

Step 9.Set the hadoop environment variables.

```
$sudo gedit $HOME/.bashrc
```

Include the following lines in the $HOME/.bashrc file

```
# Set Hadoop-related environment variables
export HADOOP_HOME=/usr/local/hadoop-2.5.1
# Set JAVA home directory
export JAVA_HOME=/usr/local/jdk1.8.0_31
 #Set aliases and functions for running Hadoop-related commands
unalias fs &> /dev/null
alias fs="hadoop fs"
unaliashls&> /dev/null
alias hls="fs -ls"
#Add Hadoop bin/ directory to PATH
export PATH=$PATH:$HADOOP_HOME/bin
```

Step 10.Set hadoop environment variables.

```
$sudo gedit /etc/profile
```

Include the following lines /etc/profile file

```
#--insert JAVA_HOME
JAVA_HOME=/usr/local/jdk1.8.0_31
#--insert HADOOP_PREFIX
HADOOP_PREFIX=/usr/local/hadoop-2.5.1
#--in PATH variable just append at the end of the line
PATH=$PATH:$JAVA_HOME/bin:$HADOOP_PREFIX/bin
#--Append HADOOP_PREFIX at end of the export statement
export PATH JAVA_HOME HADOOP_PREFIX
```

Step 11.Run the .bashrc& profile files from the $ prompt for updating the changes

```
$ source $HOME/.bashrc
$ source /etc/profile
```

Step 12.Verify java &hadoop installation using

```
$ java -version
$ echo $HADOOP_PREFIX
$ cd $HADOOP_PREFIX
$ bin/hadoop version
```

Step 13. Configuration of the hadoop files:

hadoop-env.sh, core-site.xml, mapred-site.xml, hdfs-site.xml and yarn-site.xml

```
$cd etc/hadoop
```

verify the path : /usr/local/hadoop-2.5.1/etc/hadoop

13.1.Configuration of the hadoop-env.sh file

```
$sudo gedit hadoop-env.sh
```

Include the following lines in hadoop-env.sh file

```
export JAVA_HOME=/usr/local/jdk1.8.0_31
export HADOOP_PREFIX=/usr/local/hadoop-2.5.1
```

```
$sudo mkdir -p /app/hadoop/tmp
$sudo chown hdptools:hadoop /app/hadoop/tmp
```

Configure the directory for Hadoop to store its data files, the network ports it listens to, etc. Setup will use Hadoop's Distributed File System(HDFS-single local machine).

13.2.Configuration of the core-site.xml file

```
$sudo gedit core-site.xml
```

Include the following lines in core-site.xml file between <configuration> and </configuration> tags.

```
<property>
<name>hadoop.tmp.dir</name>
<value>/app/hadoop/tmp</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
```

```
$sudo cp mapred-site.xml.template mapred-site.xml
```

13.3.Configuration of the mapred-site.xml

```
$sudo gedit mapred-site.xml
```

Include the following lines in mapred-site.xml file

```
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
```

13.4.Configuration of the hdfs-site.xml

```
$sudo gedit hdfs-site.xml
```

Include the following lines in hdfs-site.xml file

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
```

13.5.Configuration of the yarn-site.xml

```
$sudo gedit yarn-site.xml
```

Include the following lines in yarn-site.xml file

```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
```

Step 14.Format the Hadoop File system implemented on top of the local file system using

```
$cd ..
$cd ..
$cd bin
```

Verify the path :/usr/local/hadoop-2.5.1/bin

```
$hadoop namenode -format
```

Step 15.Start Hadoop using

```
$cd ..
$cd sbin
```

Verify the path : /usr/local/hadoop-2.5.1/sbin

```
$./start-all.sh
$ jps
```

EXECUTION OF SPECIFIC WORDCOUNT JAR FILE

1. Create a directory'/input' in HDFS

   /usr/local/hadoop-2.5.1$ bin/hdfs dfs -mkdir /input
2. Create input file "/fsample/sample.txt" in /home/hduser/Downloads/
3. Copy the input files into the distributed file system

   /usr/local/hadoop-2.5.1$ bin/hdfs dfs -put /home/hduser/Downloads/fsample/* /input

4. Run some of the examples provided

```
/usr/local/hadoop-2.5.1$ bin/hadoop jar
share/hadoop/mapreduce/hadoop-mapreduce- examples-2.5.1.jar grep
/input /output '<<specify word to be counted>>'
```

5. View the output files on the distributed file system

```
/usr/local/hadoop-2.5.1$ bin/hdfs dfs -cat /output/*
```
