################ PART III. Pre-requisite

####################### Step 0. 0 Assuming that all 3 Linux Ubuntu instances have been equipped with a user called ‘huser’, and the keywordless ssh has been established.

####################### Step 0. 1 Install Java . ----- do this for all machines

sudo apt-get update ## Run this before running the openjdk-7-jdk

sudo apt-get install openjdk-7-jdk

which java

####################### Step 0. 2 Assuming that I want to install Hadoop at the folder /data

sudo mkdir /data

sudo chown ubuntu:ubuntu /data

################ Step 1. Download and Install the Hadoop

cd /data

wget http://www.motorlogy.com/apache/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz

sudo tar zxf hadoop-2.6.0.tar.gz

sudo chown -R ubuntu:ubuntu hadoop-2.6.0

################ Step 2. Setup the Hadoop property files.

############### Step 2.0. Set up the property files.

At Machine0, run the following codes to get the Setup files from Git to Hadoop Setup folder /data:

cd /data

git clone https://github.com/gz5427/HadoopSetupFiles.git

cd HadoopSetupFiles

ls

Note: There are 3 folders (bashrc, libexec and etchadoop ) that contain Hadoop Setup files as well as the 3 instruction files including this one.

############### Step 2. 1 Copy the .bashrc environmental files and property files.

###### 2.1.1 prepare the .bashrc file:

############### From the base machine, do the following:

export workDir=/data/HadoopSetupFiles/bashrc

cd $workDir

scp .bashrc ubuntu@machine0:/home/ubuntu/

scp .bashrc ubuntu@machine1:/home/ubuntu/

scp .bashrc ubuntu@machine2:/home/ubuntu/

###### Then on each of the machines run the following:

cd

. .bashrc

echo $HADOOP\_HOME

###### 2.1.2 prepare the environmental files:

######## From the base machine, do the following:

export workDir=/data/HadoopSetupFiles/libexec

cd $workDir

scp \* [ubuntu@machine0:/data/hadoop-2.6.0/libexec](mailto:ubuntu@machine0:/data/hadoop-2.6.0/libexec)

scp \* [ubuntu@machine1:](mailto:hduser@master:/home/hduser/hadoop-2.4.0/)/data/hadoop-2.6.0/libexec

scp \* [ubuntu@machine2:](mailto:hduser@master:/home/hduser/hadoop-2.4.0/)/data/hadoop-2.6.0/libexec

###### 2.1.3 prepare the property files:

######## From the base machine, do the following:

export workDir=/data/HadoopSetupFiles/etchadoop

cd $workDir

scp \* ubuntu@machine0:/data/hadoop-2.6.0/etc/hadoop

scp \* ubuntu@machine1:/data/hadoop-2.6.0/etc/hadoop

scp \* ubuntu@machine2:/data/hadoop-2.6.0/etc/hadoop

############### Step 2. 2 Update the site files where name node IP Addresses are used.

###### Update the site files where name node IP Addresses are used.

Need to update the files that has the host weblink for the machine0.

export workDir=$HADOOP\_HOME/etc/hadoop

cd $workDir

grep amazonaws \*site\*

newsite=ec2-52-20-205-45 # Assuming your DNS for your machine0 is:

ec2-54-175-16-253.compute-1.amazonaws.com

cp core-site.xml core-site.xml.keep

sed "s/ec2-52-21-147-71/${newsite}/g" core-site.xml.keep > core-site.xml

diff core-site.xml core-site.xml.keep # Just to make sure substitution is done correctly

cp yarn-site.xml yarn-site.xml.keep

sed "s/ec2-52-21-147-71/${newsite}/g" yarn-site.xml.keep > yarn-site.xml

rm \*.keep

scp \*site\* ubuntu@machine1:/data/hadoop-2.6.0/etc/hadoop

scp \*site\* ubuntu@machine2:/data/hadoop-2.6.0/etc/hadoop

Note: These “amaonaws.com” names exists in the \*site\* files. Replace all ‘ec2-52-21-147-71.compute-1.amazonaws.com’ in the site files to be the public domain name (Public DNS)of your current machine0.

################ Step III. Final Preparation.

######### Step 3.1 **setup the tmp file for hadoop**

**On the machine0**, at the folder $HADOOP\_HOME,

**mkdir** $HADOOP\_HOME**/tmp**

######### Step 3.2 **formatting the name node.**

**On the machine0**, at the $HADOOP\_HOME/bin,

$HADOOP\_HOME/bin/hdfs namenode -format

################ Step IV. Start the Hadoop File system and launch the YARN.

############### 4.1 **Start the hadoop nodes.**

$HADOOP\_HOME/sbin/start-dfs.sh ### Start Hadoop File system

$HADOOP\_HOME/sbin/start-yarn.sh ### Start the Yarn

$HADOOP\_HOME/sbin/stop-yarn.sh ## Stop Yarn First.

$HADOOP\_HOME/sbin/stop-dfs.sh ## Stop Hadoop File

Note: $HADOOP\_HOME/bin:$HADOOP\_HOME/sbin has been added to $PATH the .bashrc file so You may just run "start-dfs.sh" and "start-yarn.sh".

###### 4.1.1 run ‘jps’ at machine0, you should see:

ubuntu@machine0:/data$ jps

5386 Jps

5132 ResourceManager

4774 NameNode

4993 SecondaryNameNode

###### 4.1.2 run ‘jps’ at machine1 or machine 2, you should see:

ubuntu@machine1:/data$ jps

3119 DataNode

3273 NodeManager

3381 Jps

############### 4.2 **Take a look at the file system**

hdfs dfs -mkdir /app

hdfs dfs -mkdir /hive

hdfs dfs -mkdir /user

hdfs dfs -mkdir /in

hdfs dfs -ls /

############### 4.3 **Take a look at the file system**

cd ~

echo "Hooray! My first hadoop environment works." >firsthadoop.txt

hdfs dfs -copyFromLocal firsthadoop.txt /in

hdfs dfs -ls /in

hdfs dfs -cat /in/firsthadoop.txt

############### 4.4 **Run a hadoop program.**

**jarFile=**/data/hadoop-2.6.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.0.jar

hdfs dfs –rm -r /out

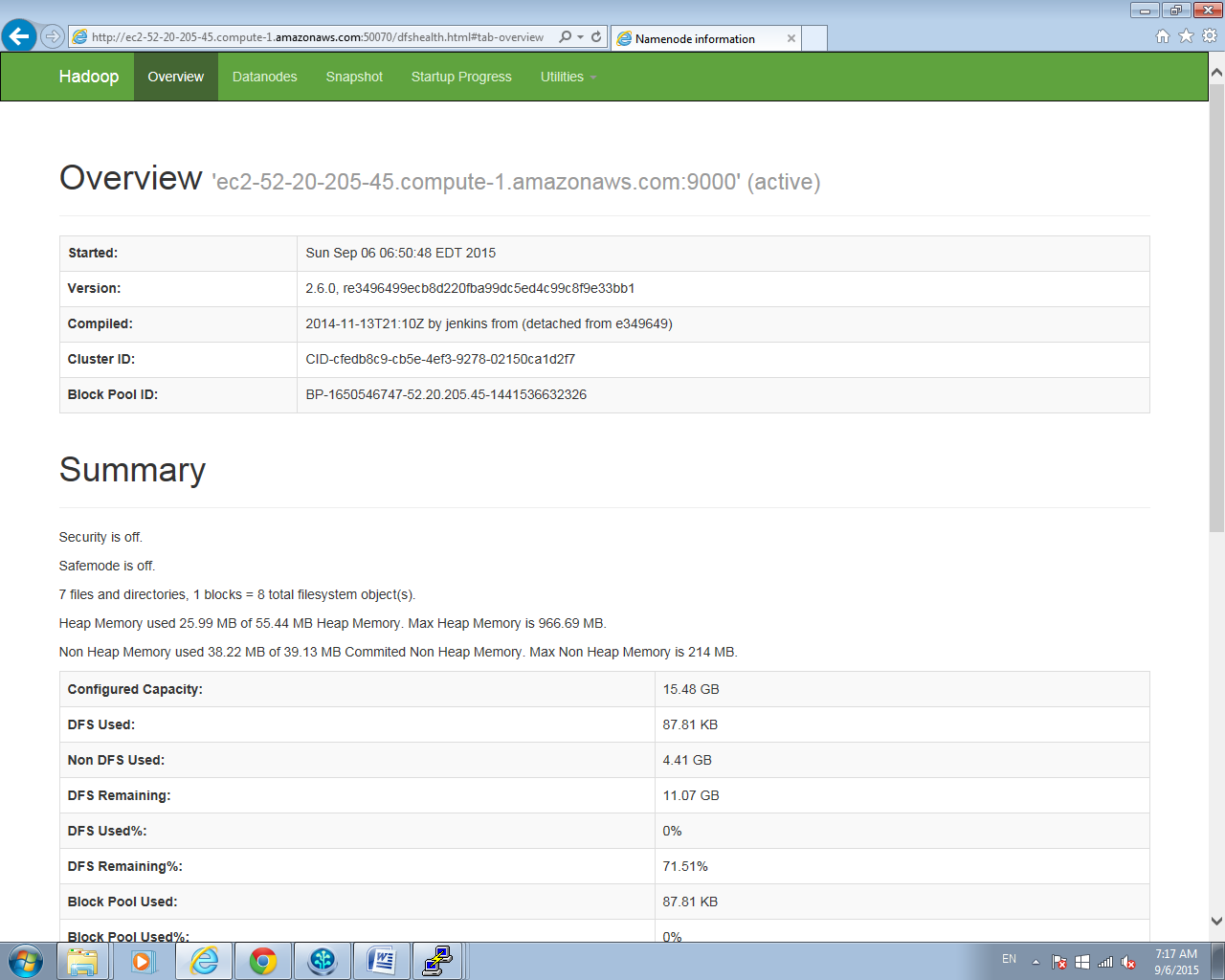
hadoop jar $jarFile wordcount /in /out

hdfs dfs -cat /out/\*

############### 4.5 **Can browse the 3-node hadoop cluster using the following Weblink:**

http://ec2-52-20-205-45.compute-1.amazonaws.com:50070/

###### Step 4.5.1 Check out the Overview of your cluster.



S###### Step 4.5.2 Use the Utitility to browse your datanodes.

