Subject : Software Laboratory VI
Class: BE IT
Experiment No: 1

Aim: Study and Configure Hadoop for Big Data

Reference :: http://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/SingleCluster.html

Steps::

sudo apt-get update

sudo apt-get install openjdk-7-jre-headless

sudo apt-get install openjdk-7-jdk

sudo apt-get install ssh

sudo apt-get install rsync

Download hadoop from : http://www.eu.apache.org/dist/hadoop/common/stable/hadoop-2.7.1.tar.gz

copy and extract hadoop-2.7.1.tar.gz in home folder

rename the name of the extracted folder from hadoop-2.7.1 to hadoop

readlink -f /usr/bin/javac

find whether ubuntu is 32 bit (i686) or 64 bit (x86_64)
uname -i

gedit ~/hadoop/etc/hadoop/hadoop-env.sh
add following line in it

for 32 bit ubuntu
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-**i386**

for 64 bit ubuntu
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-**amd64**

save and exit the file

to display the usage documentation for the hadoop script try next command
~/hadoop/bin/hadoop

1. standalone mode

```
mkdir input

cp ~/hadoop/etc/hadoop/*.xml input

~/hadoop/bin/hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar grep input output 'us[a-z.]+'

cat output/*

# Our task is done, so remove input and output folders

rm -r input output
```

---

2. Pseudo-Distributed mode

```
# get your user name
whoami
# remember your user name, we'll use it in the next step

gedit ~/hadoop/etc/hadoop/core-site.xml

<configuration>
   <property>
      <name>fs.defaultFS</name>
      <value>hdfs://localhost:1234</value>
   </property>
</configuration>

gedit ~/hadoop/etc/hadoop/hdfs-site.xml

<configuration>

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

<property>
<name>dfs.name.dir</name>
<value>file:///home/your_user_name/hadoop/name_dir</value>
</property>

<property>
<name>dfs.data.dir</name>
<value>file:///home/your_user_name/hadoop/data_dir</value>
</property>

</configuration>
```

#Setup passphraseless/passwordless ssh

ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa

cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys

export HADOOP_PREFIX=/home/**your_user_name**/hadoop

ssh localhost

type **exit** in the terminal to close the ssh connection (very important)
exit

 # The following instructions are to run a MapReduce job locally.
#Format the filesystem:(**Do it only once**)
~/hadoop/bin/hdfs namenode -format

#Start NameNode daemon and DataNode daemon:
~/hadoop/sbin/start-dfs.sh

#Browse the web interface for the NameNode; by default it is available at:
http://localhost:50070/

#Make the HDFS directories required to execute MapReduce jobs:
~/hadoop/bin/hdfs dfs -mkdir /user

~/hadoop/bin/hdfs dfs -mkdir /user/**your_user_name**

#Copy the sample files (from ~/hadoop/etc/hadoop) into the distributed filesystem folder(input)
~/hadoop/bin/hdfs dfs -put ~/hadoop/etc/hadoop input

#Run the example map-reduce job
~/hadoop/bin/hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar grep input output 'us[a-z.]+'

#View the output files on the distributed filesystem
~/hadoop/bin/hdfs dfs -cat output/*

#Copy the output files from the distributed filesystem to the local filesystem and examine them:
~/hadoop/bin/hdfs dfs -get output output
#ignore warnings (if any)

cat output/*

remove local output folder
rm -r output

remove distributed folders (input & output)
~/hadoop/bin/hdfs dfs -rm -r input output

#When you're done, stop the daemons with
~/hadoop/sbin/stop-dfs.sh

Prof. S. T. Kolhe (Department of I.T – S.R.E.S C.O.E Kopargaon)