BIG DATA TECHNOLOGIES LABORATORY

NAME: CHAARVIKA.A

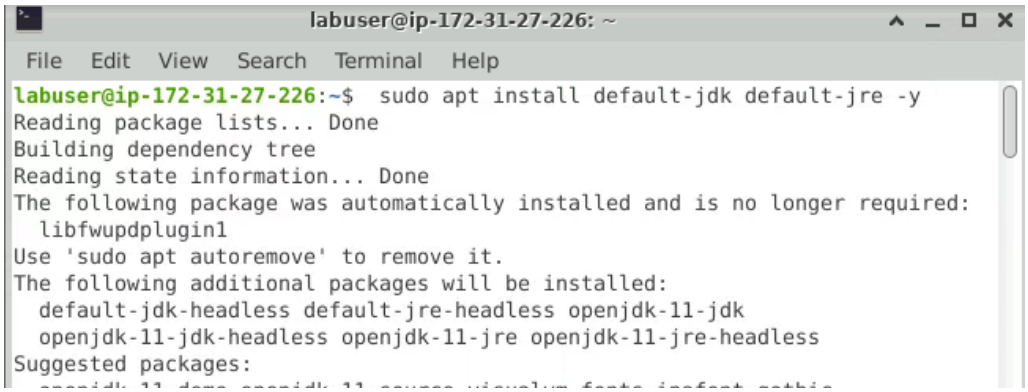
ROLL NUMBER: 20BIS012

LAB EX-1: Hadoop Installation

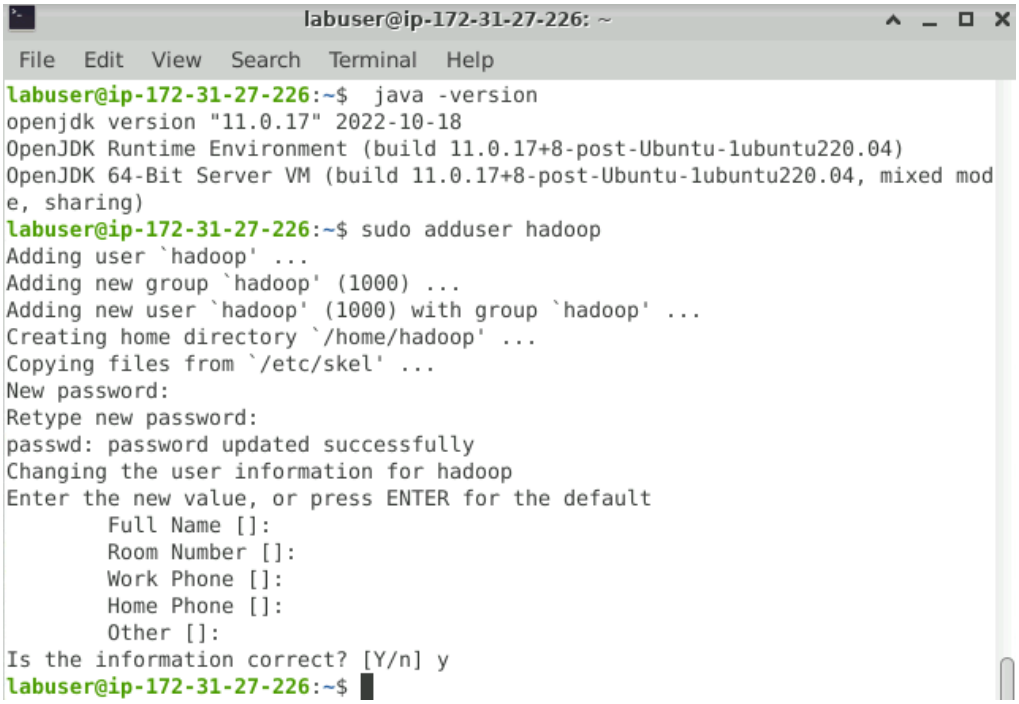
LAB EXERCISE-1

HADOOP INSTALLATION

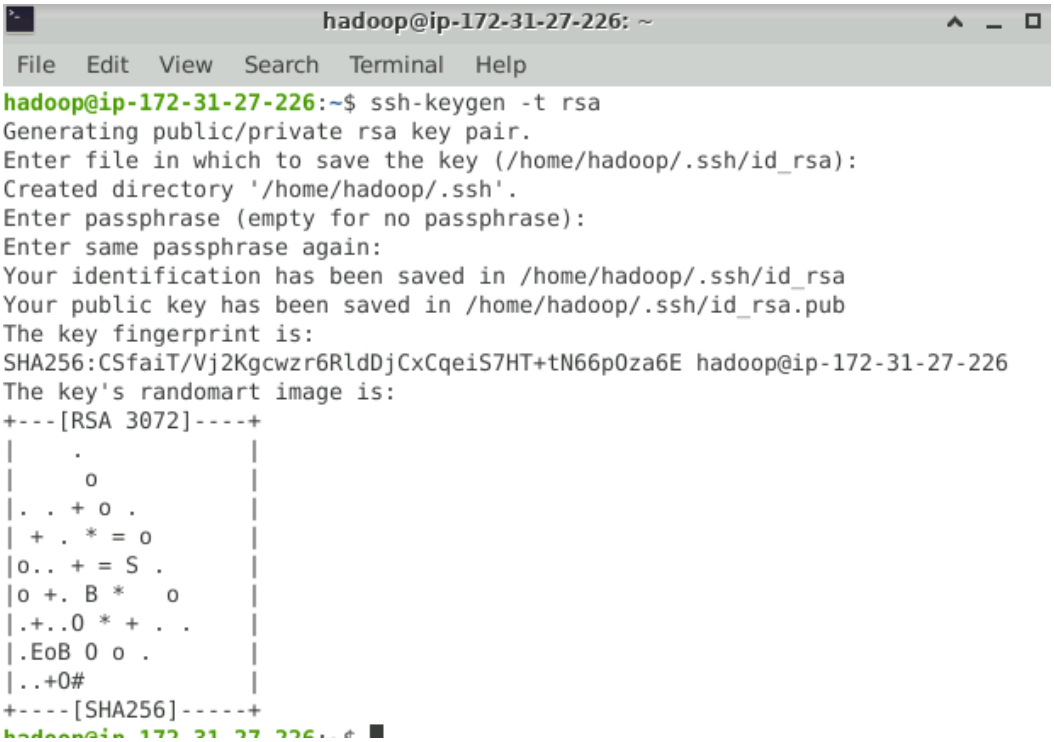
STEP -1



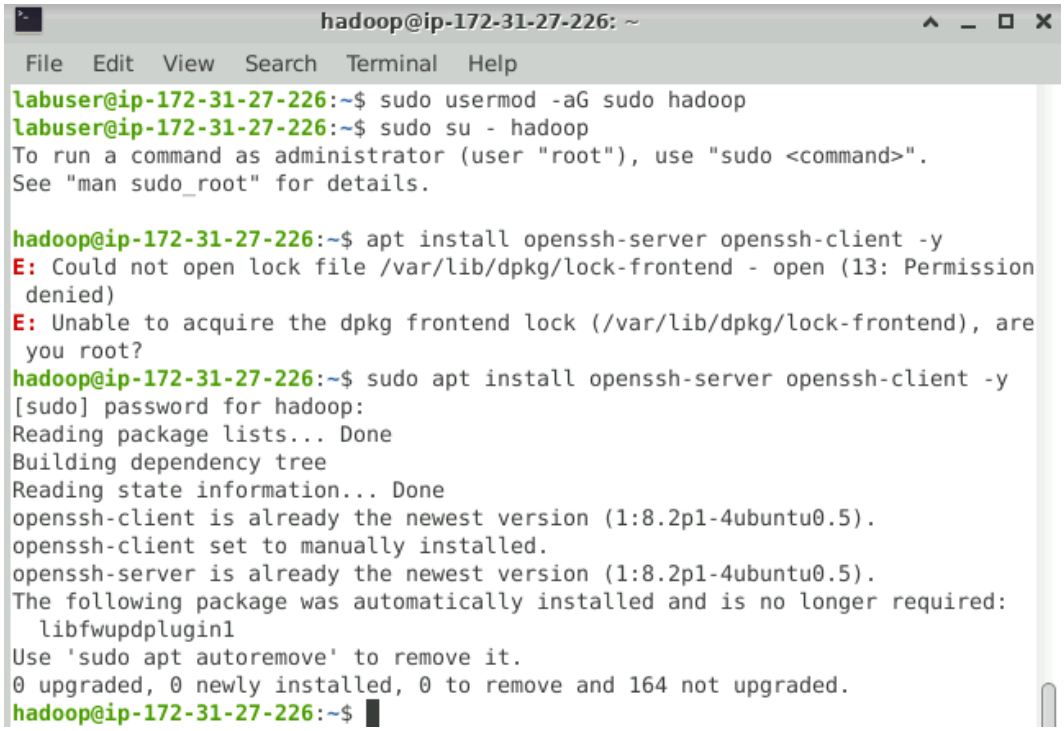
STEP - 2


labuser@ip-172-31-27-226: ~
File Edit View Search Terminal Help
labuser@ip-172-31-27-226:~$ java -version
openjdk version "11.0.17" 2022-10-18
OpenJDK Runtime Environment (build 11.0.17+8-post-Ubuntu-1ubuntu220.04)
OpenJDK 64-Bit Server VM (build 11.0.17+8-post-Ubuntu-1ubuntu220.04, mixed mode, sharing)
labuser@ip-172-31-27-226:~$ sudo adduser hadoop
Adding user `hadoop' ...
Adding new group `hadoop' (1000) ...
Adding new user `hadoop' (1000) with group `hadoop' ...
Creating home directory `/home/hadoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
Full Name []:
Room Number []:
Work Phone []:
Home Phone []:
Other []:
Is the information correct? [Y/n] y
labuser@ip-172-31-27-226:~$

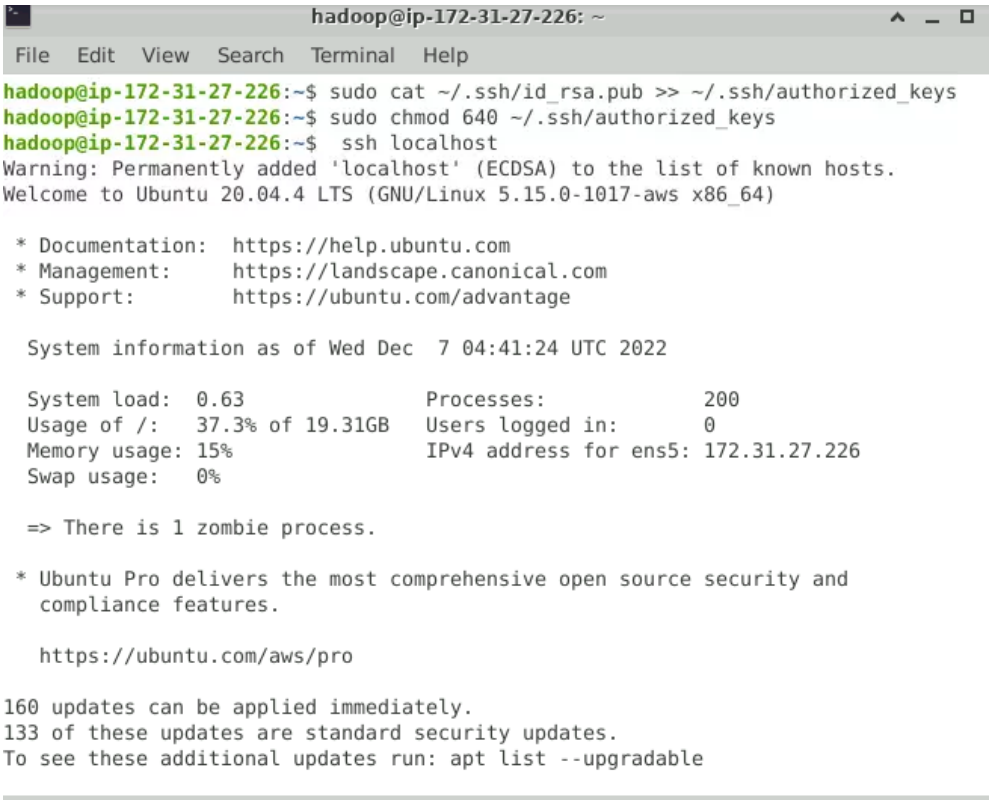
STEP - 3


hadoop@ip-172-31-27-226: ~
File Edit View Search Terminal Help
hadoop@ip-172-31-27-226:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:CSfaiT/Vj2Kgcwzr6RldDjCxCqeiS7HT+tN66pOza6E hadoop@ip-172-31-27-226
The key's randomart image is:
+---[RSA 3072]----+
| . |
| o |
|. . + o . |
| + . * = o |
|o.. + = S . |
|o +. B * o |
|.+..O * + . . |
|.EoB O o . |
|..+O# |
+----[SHA256]-----+

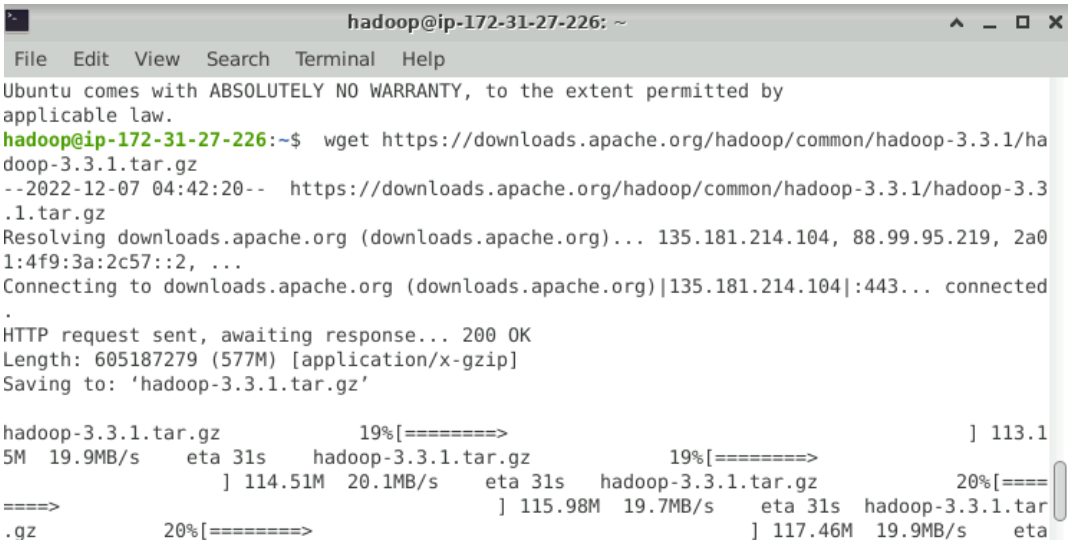
STEP - 4


hadoop@ip-172-31-27-226: ~
File Edit View Search Terminal Help
labuser@ip-172-31-27-226:~$ sudo usermod -aG sudo hadoop
labuser@ip-172-31-27-226:~$ sudo su - hadoop
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.
hadoop@ip-172-31-27-226:~$ apt install openssh-server openssh-client -y
E: Could not open lock file /var/lib/dpkg/lock-frontend - open (13: Permission denied)
E: Unable to acquire the dpkg frontend lock (/var/lib/dpkg/lock-frontend), are you root?
hadoop@ip-172-31-27-226:~$ sudo apt install openssh-server openssh-client -y
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-client is already the newest version (1:8.2p1-4ubuntu0.5).
openssh-client set to manually installed.
openssh-server is already the newest version (1:8.2p1-4ubuntu0.5).
The following package was automatically installed and is no longer required:
libfwupdplugin1
Use 'sudo apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 164 not upgraded.
hadoop@ip-172-31-27-226:~$

STEP - 5


hadoop@ip-172-31-27-226: ~
File Edit View Search Terminal Help
hadoop@ip-172-31-27-226:~$ sudo cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hadoop@ip-172-31-27-226:~$ sudo chmod 640 ~/.ssh/authorized_keys
hadoop@ip-172-31-27-226:~$ ssh localhost
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 20.04.4 LTS (GNU/Linux 5.15.0-1017-aws x86_64)
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage
System information as of Wed Dec 7 04:41:24 UTC 2022
System load: 0.63 Processes: 200
Usage of /: 37.3% of 19.31GB Users logged in: 0
Memory usage: 15% IPv4 address for ens5: 172.31.27.226
Swap usage: 0%
=> There is 1 zombie process.
* Ubuntu Pro delivers the most comprehensive open source security and compliance features.
https://ubuntu.com/aws/pro
160 updates can be applied immediately.
133 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

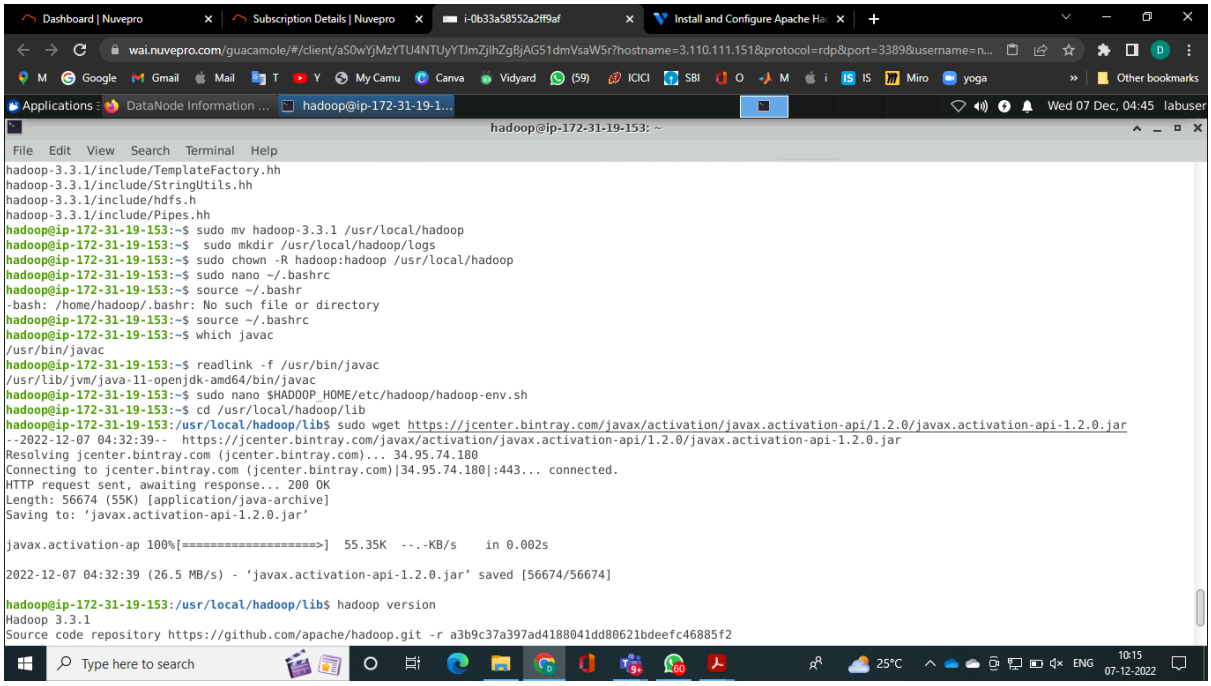
STEP - 6



STEP - 7

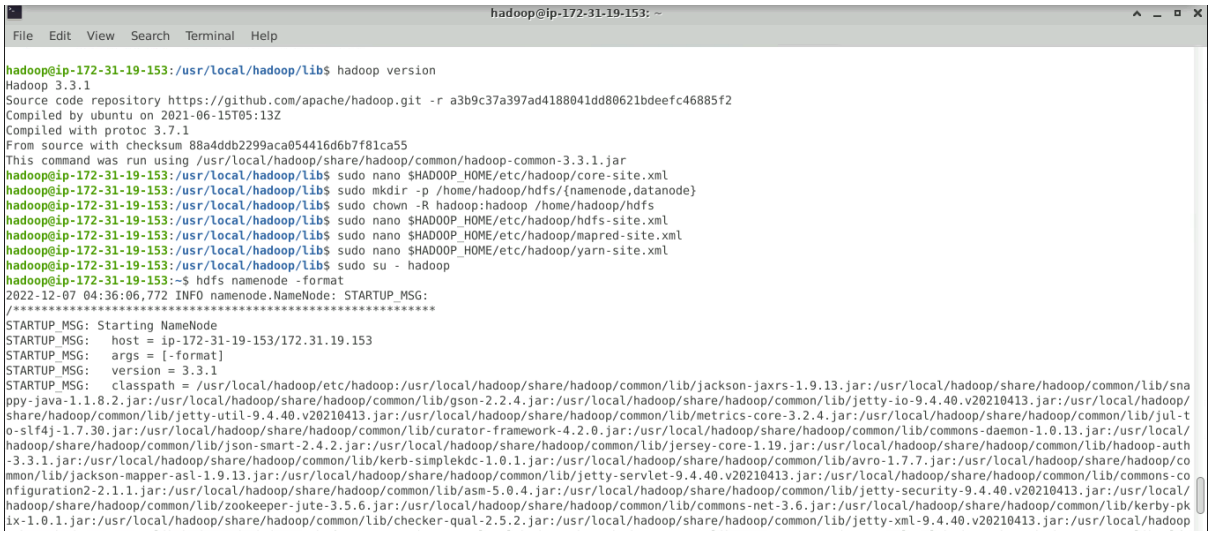
STEP - 8

```
$ sudo mv hadoop-3.3.1 /usr/local/hadoop
$ sudo mkdir /usr/local/hadoop/logs
$ sudo chown -R hadoop:hadoop /usr/local/hadoop


$ sudo nano ~/.bashrc
$ source ~/.bashrc
$ which javac
$ readlink -f /usr/bin/javac
$ sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
$ cd /usr/local/hadoop/lib

$ sudo wget
https://jcenter.bintray.com/javax/activation/javax.activation-api/1.2.0/javax.activation-api-1.2.0.jar

$ hadoop version
```



STEP -9

```
$ sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
$ sudo mkdir -p /home/hadoop/hdfs/{namenode,datanode}
$ sudo chown -R hadoop:hadoop /home/hadoop/hdfs
$ sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
$ sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
$ sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
$ sudo su - hadoop
$ hdfs namenode -format
```



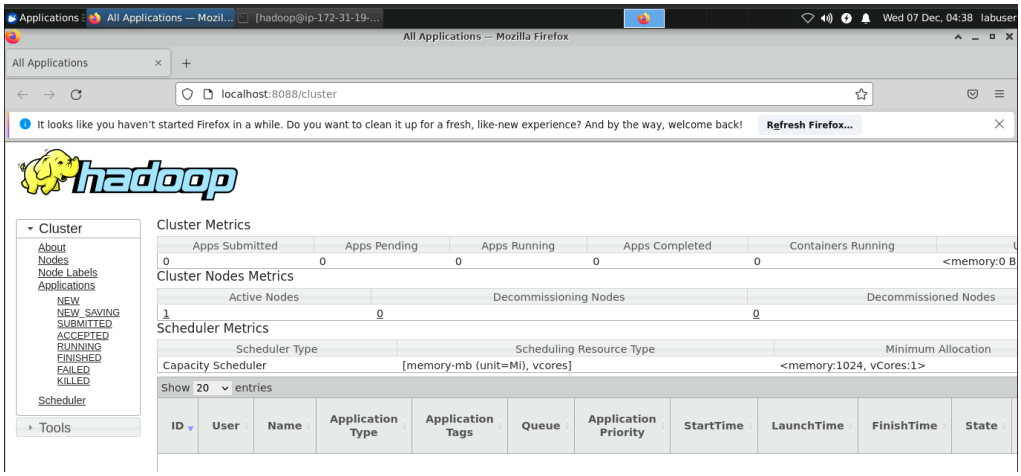
STEP -10

```
$ start-dfs.sh
$ start-yarn.sh
$ jps
```

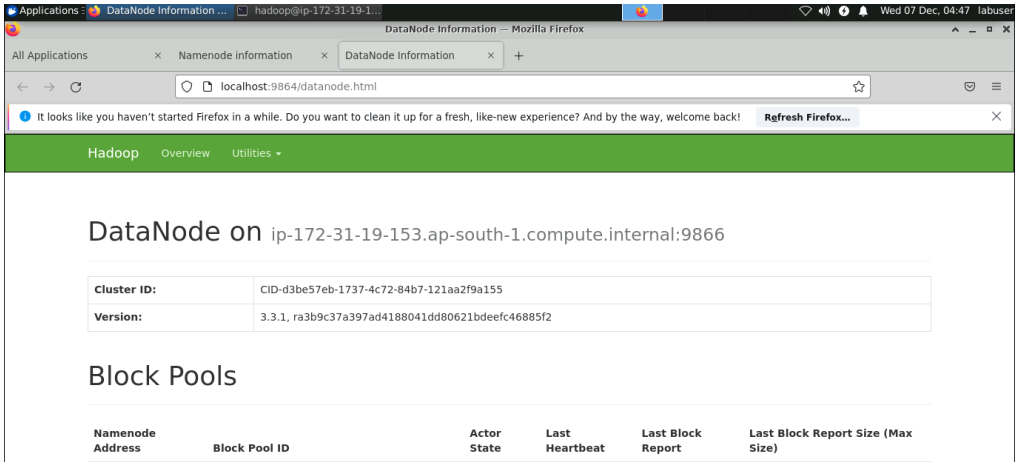
```
hadoop@ip-172-31-19-153:~$ start-dfs.sh
Starting namenodes on [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
Starting datanodes
Starting secondary namenodes [ip-172-31-19-153]
ip-172-31-19-153: Warning: Permanently added 'ip-172-31-19-153,172.31.19.153' (ECDSA) to the list of known hosts.
hadoop@ip-172-31-19-153:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@ip-172-31-19-153:~$ jps
5186 NodeManager
5029 ResourceManager
4393 NameNode
4810 SecondaryNameNode
4556 DataNode
5567 Jps
hadoop@ip-172-31-19-153:~$
```

STEP -11

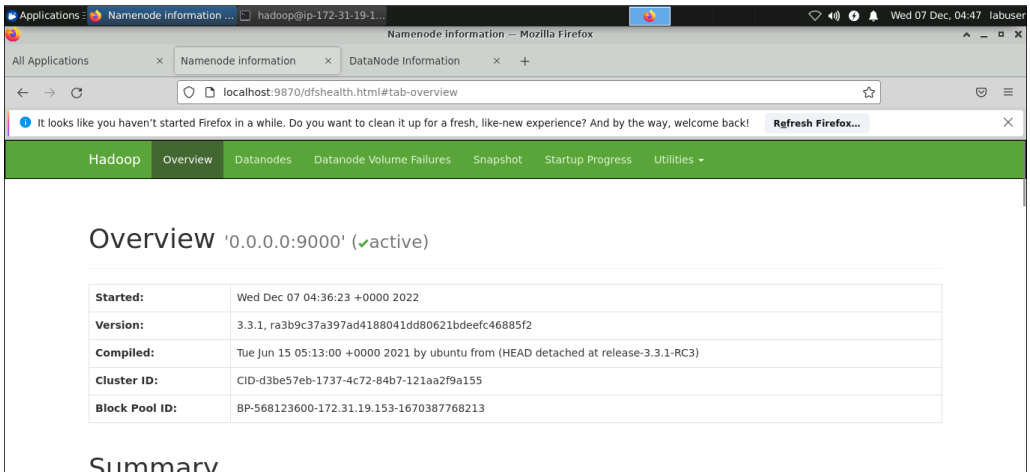
Localhost:8088



Localhost:9864



Localhost:9870



LAB EXERCISE-2

STEP -1

```
$ hdfs dfs -mkdir /user/kct5thsemcdhid036/mridula
```

```
npbdh login: kct5thsemcdhid036
kct5thsemcdhid036@npbdh.cloudloka.com's password:
Last login: Tue Dec 13 05:18:32 2022 from ec2-65-1-45-35.ap-south-1
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/kct5thse
mcdhid036/mridula
```

STEP -2

```
$ hdfs dfs -ls /
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -ls /
Found 7 items
drwxr-xr-x   - hbase hbase               0 2022-10-24 06:41 /hbase
drwxrwxrwx   - hdfs  supergroup          0 2022-09-04 20:19 /markovData
drwxrwxr-x   - solr  solr                0 2022-07-13 08:06 /solr
drwxr-xr-x   - hdfs  supergroup          0 2021-05-22 18:21 /system
drwxrwxrwt   - hdfs  supergroup          0 2022-12-11 15:50 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2022-12-13 05:42 /user
drwxr-xr-x   - hdfs  supergroup          0 2021-10-29 14:36 /userTest2
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

STEP -3

```
$ hdfs dfs -put '/home/kct5thsemcdhid036/lab.txt'
/user/kct5thsemcdhid036/mridula
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -put '/home/kct5thsemcdhid036/lab.txt' /user/kct5thsemcdhid036/mridula
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

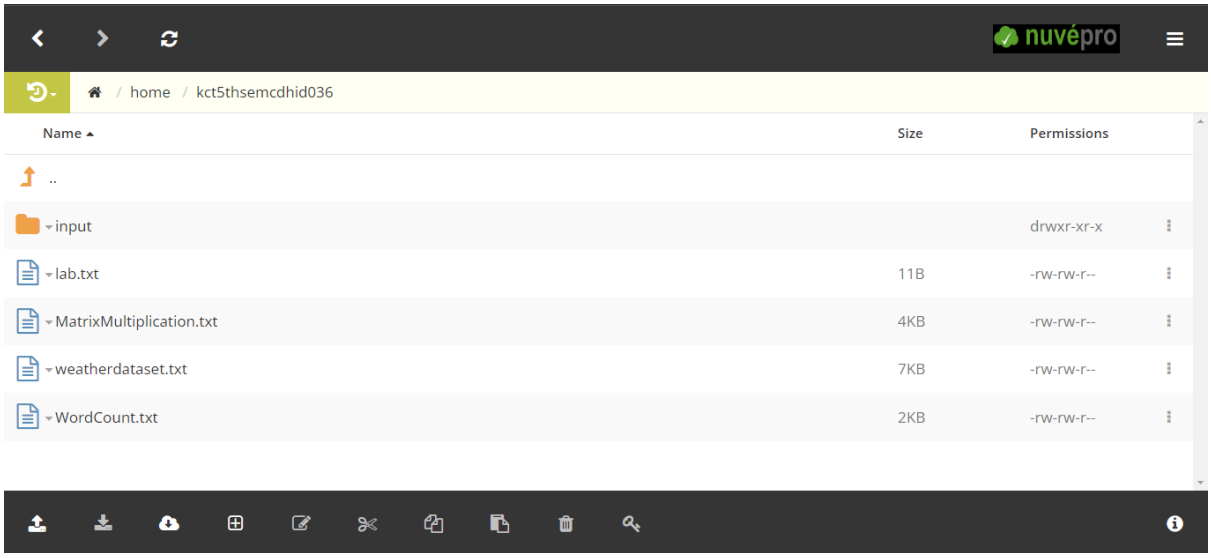
STEP -4

```
$ hadoop fs -get /user/kct5thsemcdhid036/mridula/lab.txt
/home/kct5thsemcdhid036/
```

```
$ hdfs dfs -mkdir -p /user/kct5thsemcdhid036/mridula/input/
```

```
$ hadoop fs -get /user/kct5thsemcdhid036/mridula/input
/home/kct5thsemcdhid036/
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -get /user/kct5thsemcdhid036/mridula/lab.txt /home/kct5thsemcdhid036/
get: `/home/kct5thsemcdhid036/lab.txt': File exists
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid036/mridula/input/
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -get /user/kct5thsemcdhid036/mridula/input /home/kct5thsemcdhid036/
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```



STEP -5

```
$ hadoop fs -cat /user/kct5thsemcdhid036/mridula/lab.txt
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -cat /user/kct5thsemcdhid036/mridula/lab.txt
hi ...hello[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

STEP -6

```
$ hdfs dfs -mkdir /user/kct5thsemcdhid036/hii
```

```
$ hdfs dfs -cp /user/kct5thsemcdhid036/mridula/lab.txt
/user/kct5thsemcdhid036/hii
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/kct5thsemcdhid036/hii
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -cp /user/kct5thsemcdhid036/mridula/lab.txt /user/kct5thsemcdhid036/hii
```

STEP -7

```
$ hdfs dfs -cat /user/kct5thsemcdhid036/mridula/*
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -cat /user/kct5thsemcdhid036/mridula/*
cat: `/user/kct5thsemcdhid036/mridula/input': Is a directory
hi ...hello[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

STEP -8

```
$ hadoop fs -copyFromLocal '/home/kct5thsemcdhid036/topic.txt'
/user/kct5thsemcdhid036/mridula
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -copyFromLocal '/home/kct5thsemcdhid036/topic.txt' /user/kct5thsemcdhid036/
mridula
```

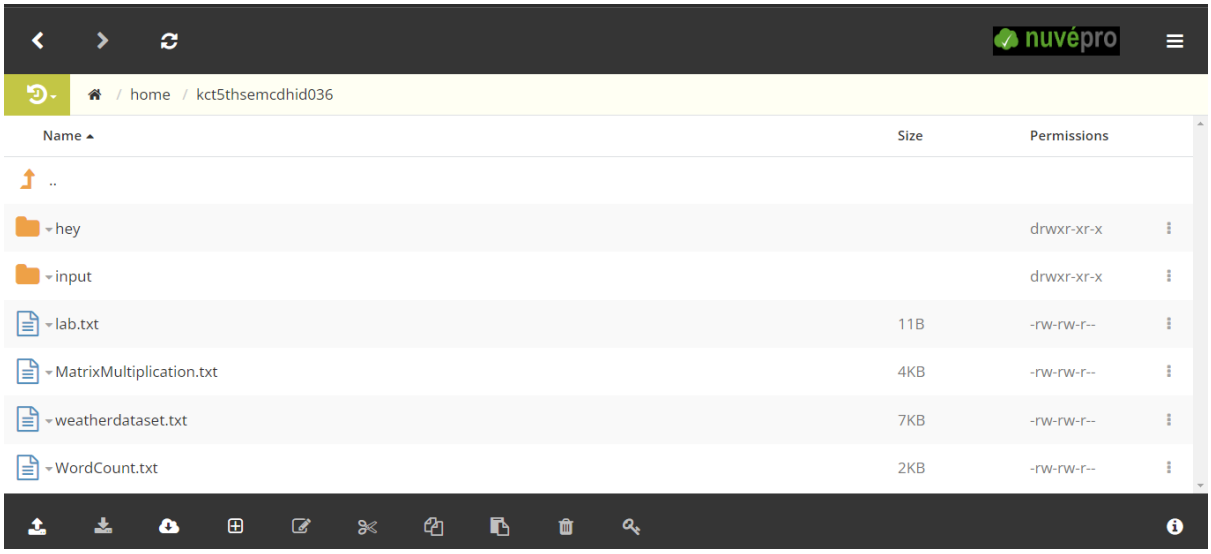
STEP -9

```
$ hdfs dfs -mkdir -p /user/kct5thsemcdhid036/mridula/hey/
```

```
$ hadoop fs -copyToLocal /user/kct5thsemcdhid036/mridula/hey
'/home/kct5thsemcdhid036/'
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid036/mridula/hey/
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -copyToLocal /user/kct5thsemcdhid036/mridula/hey '/home/kct5thsemcdhid036/'
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```



STEP -10

```
$ hdfs dfs -mv /user/kct5thsemcdhid036/mridula/lab.txt
/user/kct5thsemcdhid036/hii/
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -mv /user/kct5thsemcdhid036/mridula/lab.txt /user/kct5thsemcdhid036/hii/
mv: `/user/kct5thsemcdhid036/hii/lab.txt': File exists
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

STEP -11

```
$ hdfs dfs -mkdir /user/kct5thsemcdhid036/hello
```

```
$ hdfs dfs -rmdir /user/kct5thsemcdhid036/hello
```

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/kct5thsemcdhid036/hello
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hdfs dfs -rmdir /user/kct5thsemcdhid036/hello
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

STEP -12

$ hadoop fs -tail /user/kct5thsemcdhid036/mridula/lab.txt/

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -tail /user/kct5thsemcdhid036/mridula/lab.txt/
hi ...hello[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

STEP -13

$ hadoop fs -du /user/kct5thsemcdhid036/mridula/lab.txt/

```
[kct5thsemcdhid036@ip-10-1-1-204 ~]$ hadoop fs -du /user/kct5thsemcdhid036/mridula/lab.txt/
11  33  /user/kct5thsemcdhid036/mridula/lab.txt
[kct5thsemcdhid036@ip-10-1-1-204 ~]$
```

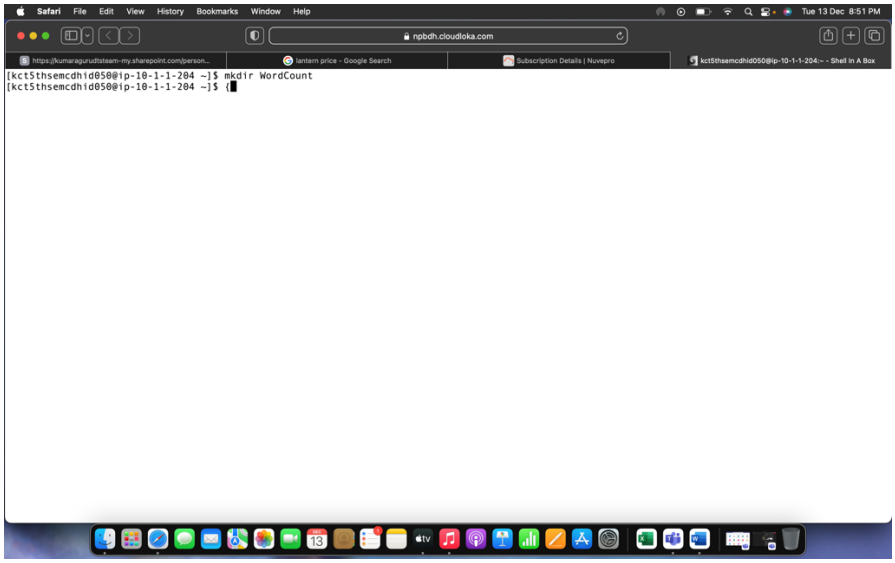
U18ISI5202-BIG DATA LABORATORY

NAME: CHAARVIKA.A
ROLL NUMBER: 20BIS012

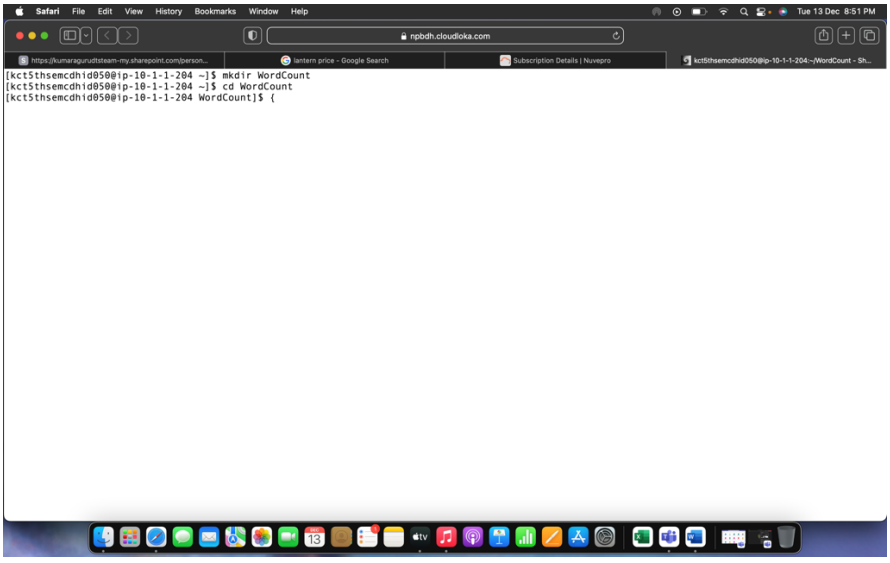
Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm

1. CREATE A FOLDER/ DIRECTORY FOR WORDCOUNT IN HOME

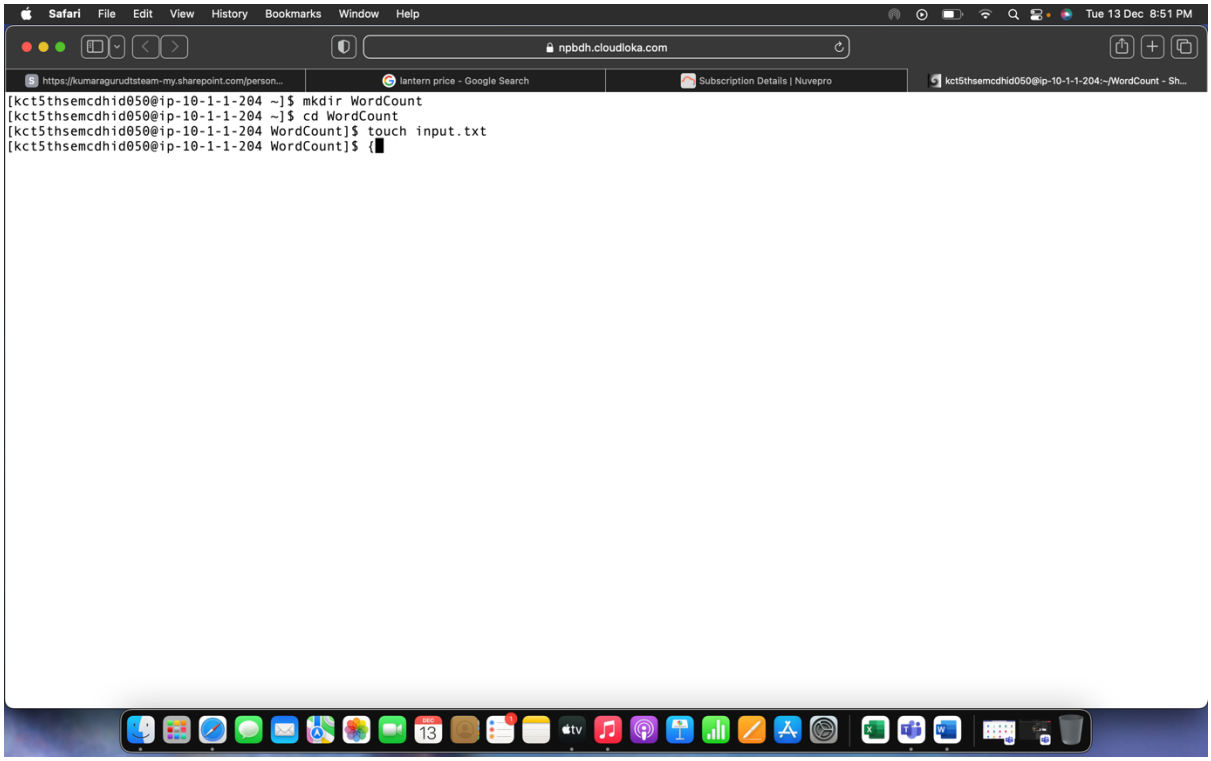
COMMAND: mkdir wordcount

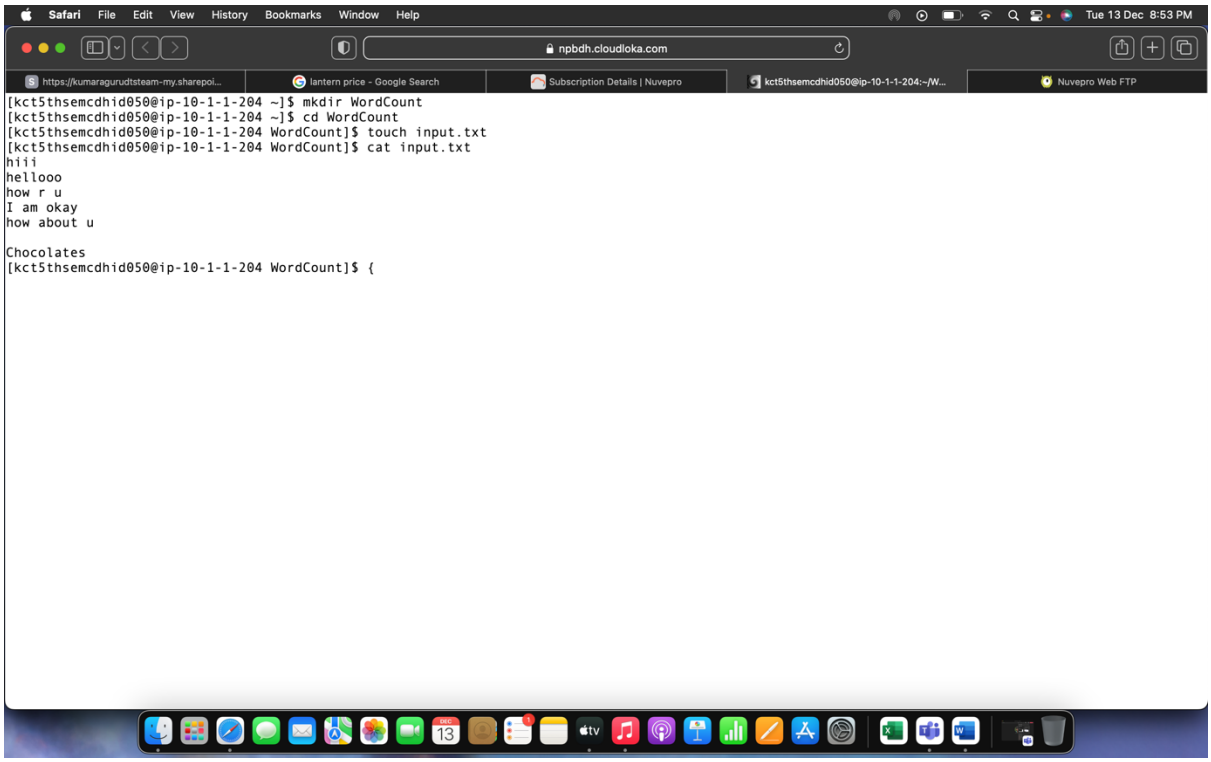


2. MOVE TO THAT FOLDER/DIRECTORY:

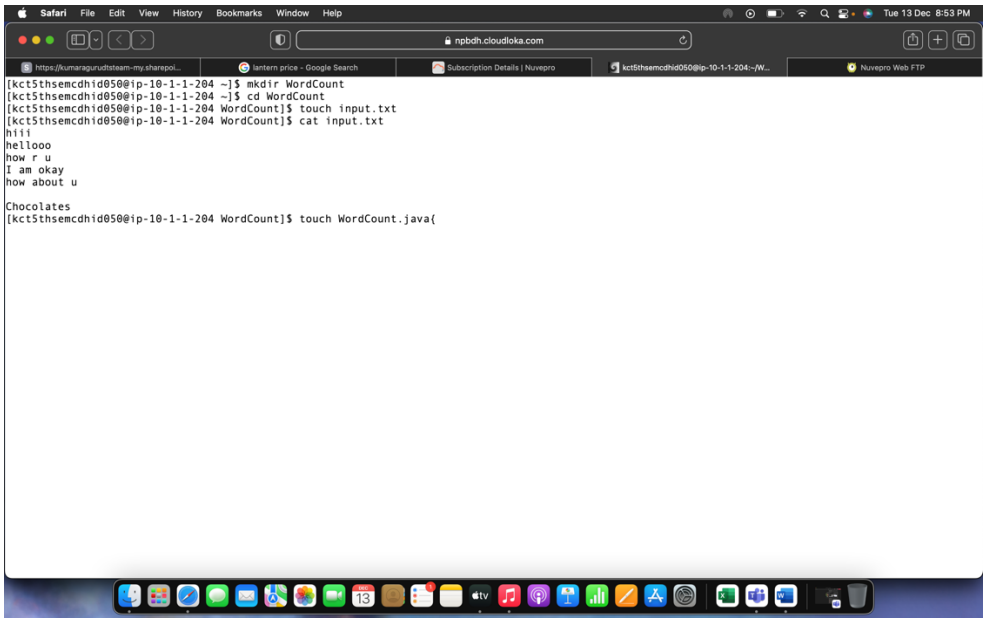


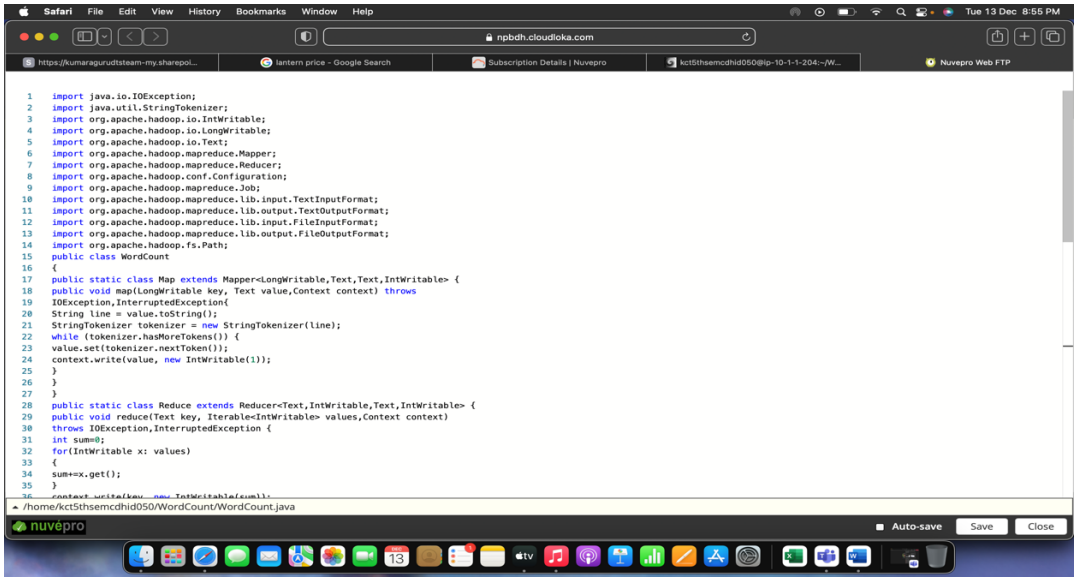
3. CREATE AN INPUT FILE INSIDE THE WORDCOUNT FOLDER:





4. CREATE A WORDCOUNT JAVA FILE:





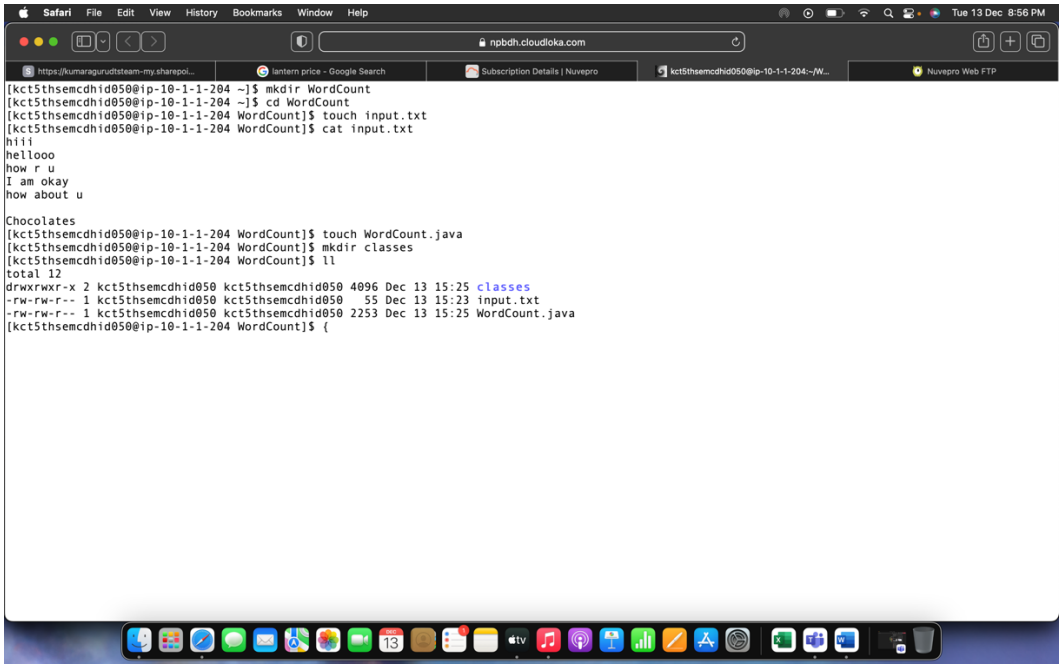
SOURCE CODE:

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
```

```
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.fs.Path;

public class WordCount
{
public static class Map extends Mapper<LongWritable,Text,Text,IntWritable> {
public void map(LongWritable key, Text value,Context context) throws
IOException,InterruptedException{
String line = value.toString();
StringTokenizer tokenizer = new StringTokenizer(line);
while (tokenizer.hasMoreTokens()) {
value.set(tokenizer.nextToken());
context.write(value, new IntWritable(1));
}
}
}
public static class Reduce extends Reducer<Text,IntWritable,Text,IntWritable> {
public void reduce(Text key, Iterable<IntWritable> values,Context context)
throws IOException,InterruptedException {
int sum=0;
for(IntWritable x: values)
{
sum+=x.get();
}
context.write(key, new IntWritable(sum));
}
}

public static void main(String[] args) throws Exception {

Configuration conf= new Configuration();
Job job = new Job(conf,"My Word Count Program");
job.setJarByClass(WordCount.class);
job.setMapperClass(Map.class);
job.setReducerClass(Reduce.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
Path outputPath = new Path(args[1]);
//Configuring the input/output path from the filesystem into the job
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
//deleting the output path automatically from hdfs so that we don't have to delete it
explicitly
```

```
outputPath.getFileSystem(conf).delete(outputPath);
//exiting the job only if the flag value becomes false
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```

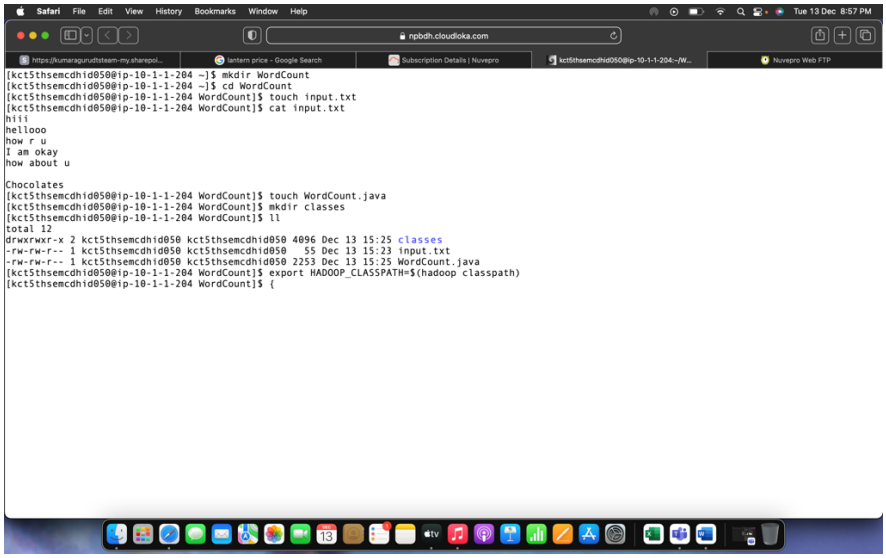
5. CREATE A DIRECTORY/ FOLDER CLASSES INSIDE THE WORDCOUNT FOLDER:

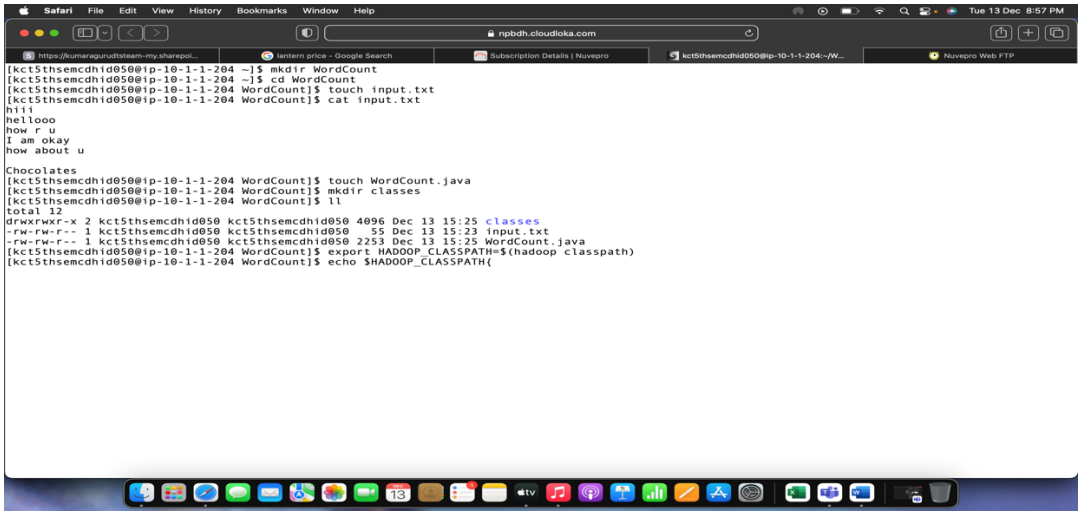


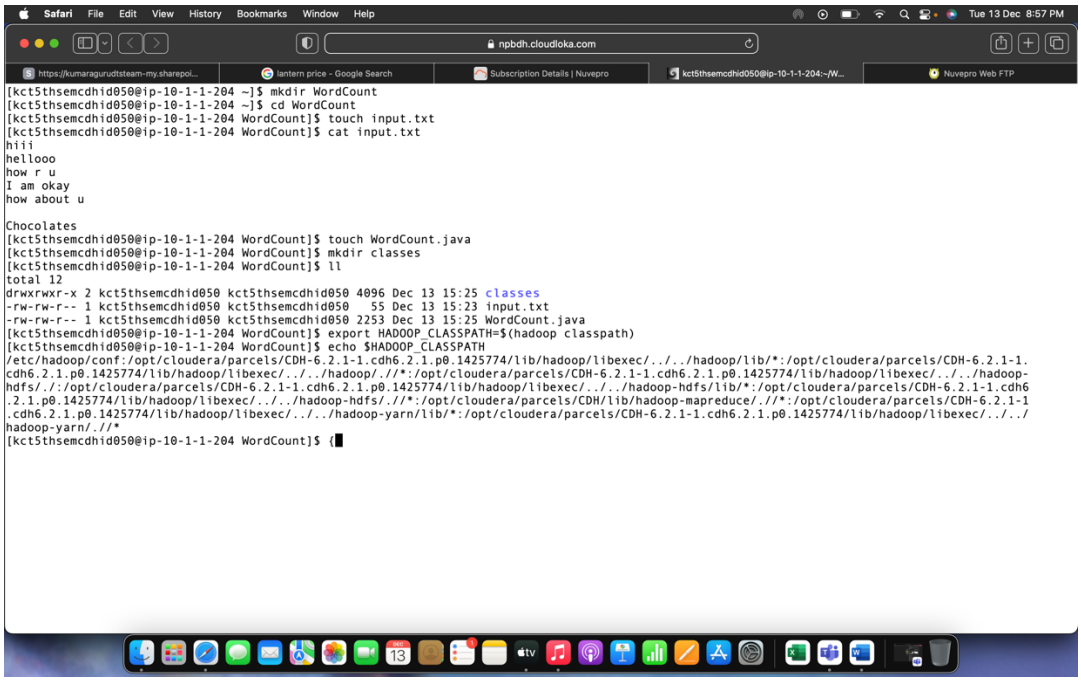
6. SET THE PATH FOR JAVA FILE

COMMAND: export HADOOP_CLASSPATH=$(hadoop classpath)

echo $HADOOP_CLASSPATH

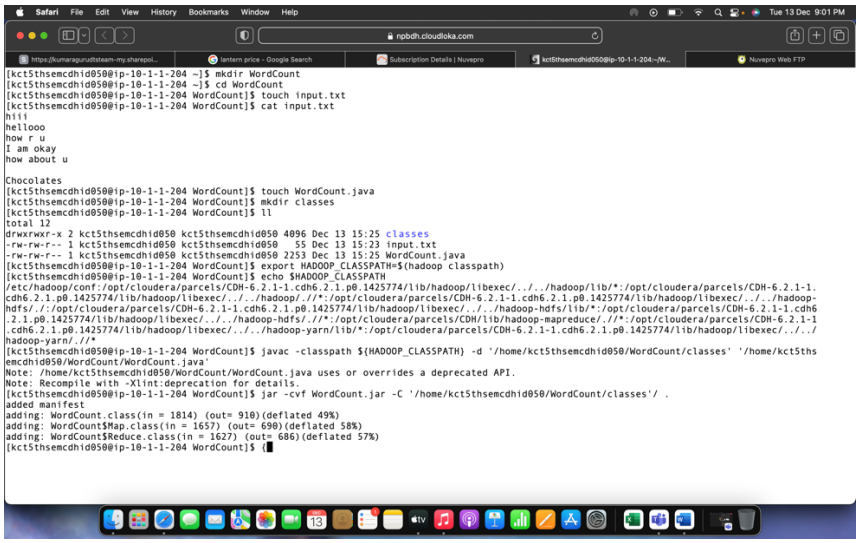



Safari File Edit View History Bookmarks Window Help
Tue 13 Dec 8:57 PM
npbdh.cloudloka.com
https://kumaragurudtsteam-my.sharepoi...
lantern price - Google Search
Subscription Details | Nuvepro
kct5thsemcdhid050@ip-10-1-1-204:~/W...
Nuvepro Web FTP
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ mkdir WordCount
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ cd WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch input.txt
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hiii
hellooo
how r u
I am okay
how about u
Chocolates
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 12
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:25 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050   55 Dec 13 15:23 input.txt
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 2253 Dec 13 15:25 WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ export HADOOP_CLASSPATH=$(hadoop classpath)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ echo $HADOOP_CLASSPATH{


Safari File Edit View History Bookmarks Window Help
Tue 13 Dec 8:57 PM
npbdh.cloudloka.com
https://kumaragurudtsteam-my.sharepoi...
lantern price - Google Search
Subscription Details | Nuvepro
kct5thsemcdhid050@ip-10-1-1-204:~/W...
Nuvepro Web FTP
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ mkdir WordCount
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ cd WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch input.txt
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hiii
hellooo
how r u
I am okay
how about u
Chocolates
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 12
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:25 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050   55 Dec 13 15:23 input.txt
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 2253 Dec 13 15:25 WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ export HADOOP_CLASSPATH=$(hadoop classpath)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop/lib/*:/opt/cloudera/parcels/CDH-6.2.1-1.
cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop/.//*:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-
hdfs/./:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-hdfs/lib/*:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6
.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-hdfs/.//*:/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/.//*:/opt/cloudera/parcels/CDH-6.2.1-1
.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-yarn/lib/*:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../
hadoop-yarn/.//*
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ (

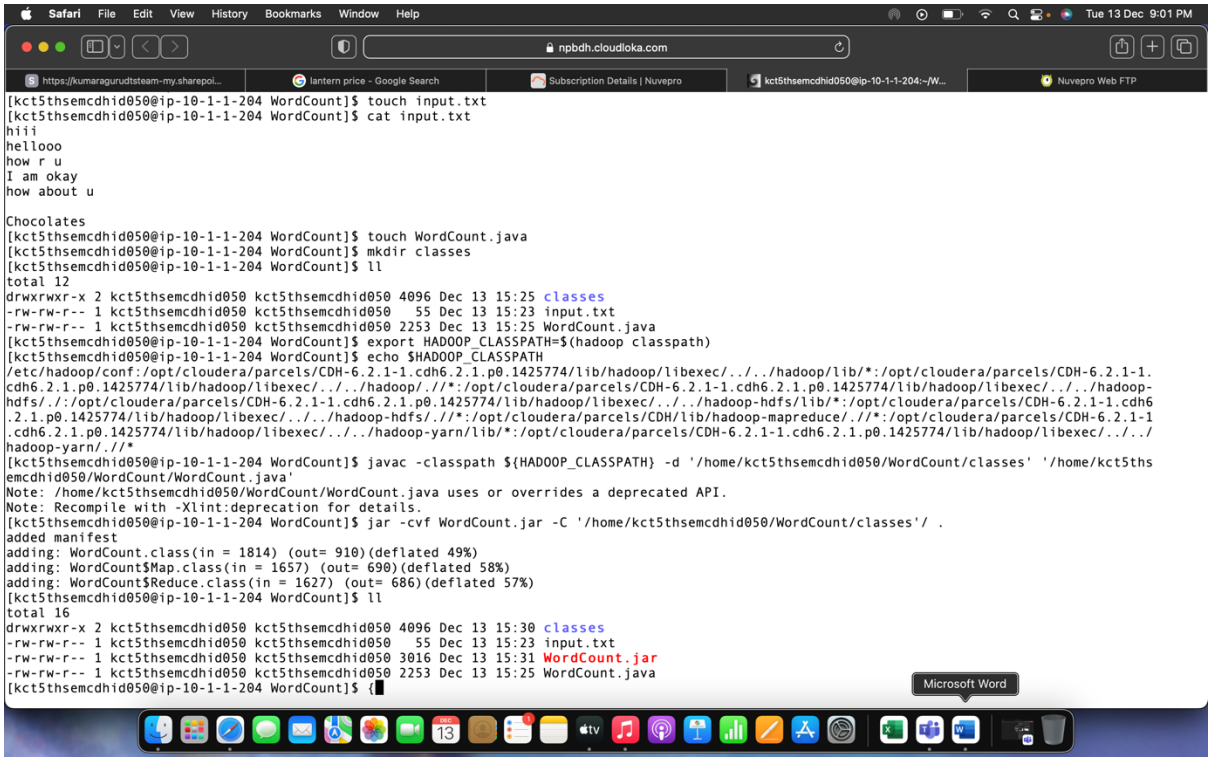
7. COMPILE THE JAVAFILE:

COMMAND: javac -classpath ${HADOOP_CLASSPATH} -d '/home/<NAME>/wordcount/classes' '/home/<NAME>/wordcount/WordCount.java'

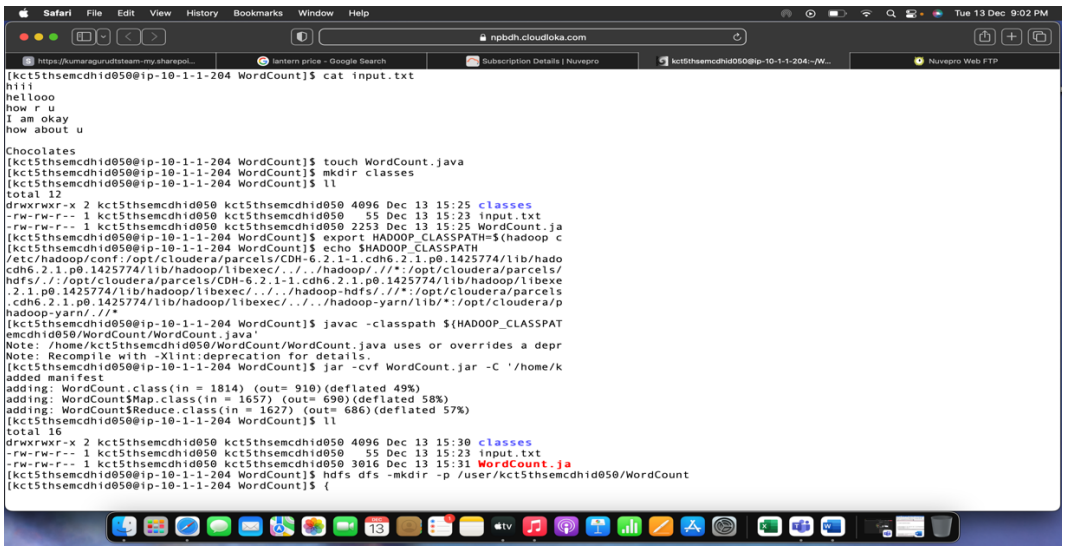


8. CREATE A JAR FILE:

COMMAND: jar -cvf WordCount.jar -C '/home/<NAME>/wordcount/classes'/ .

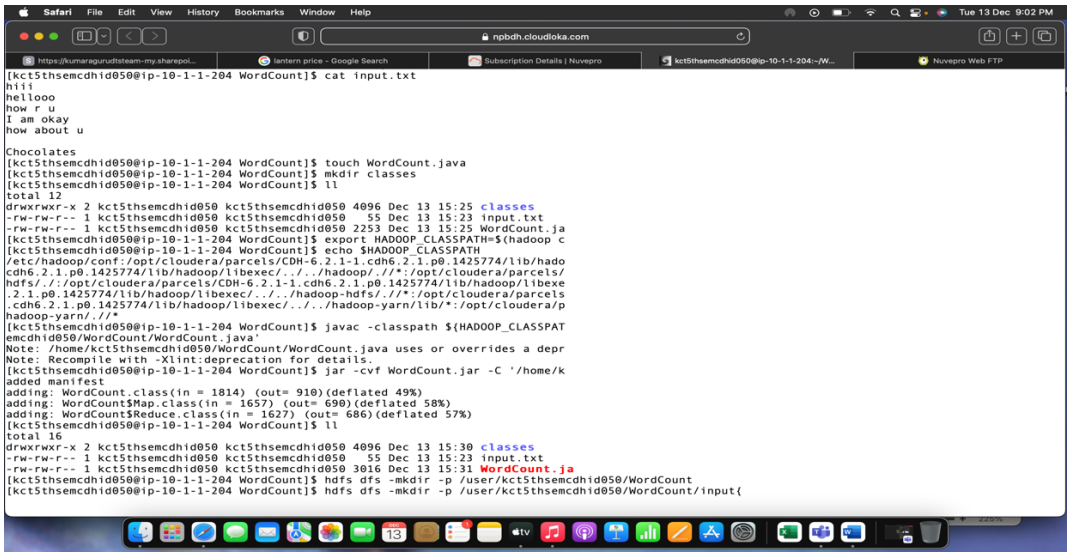


9. CREATE A DIRECTORY IN HADOOP

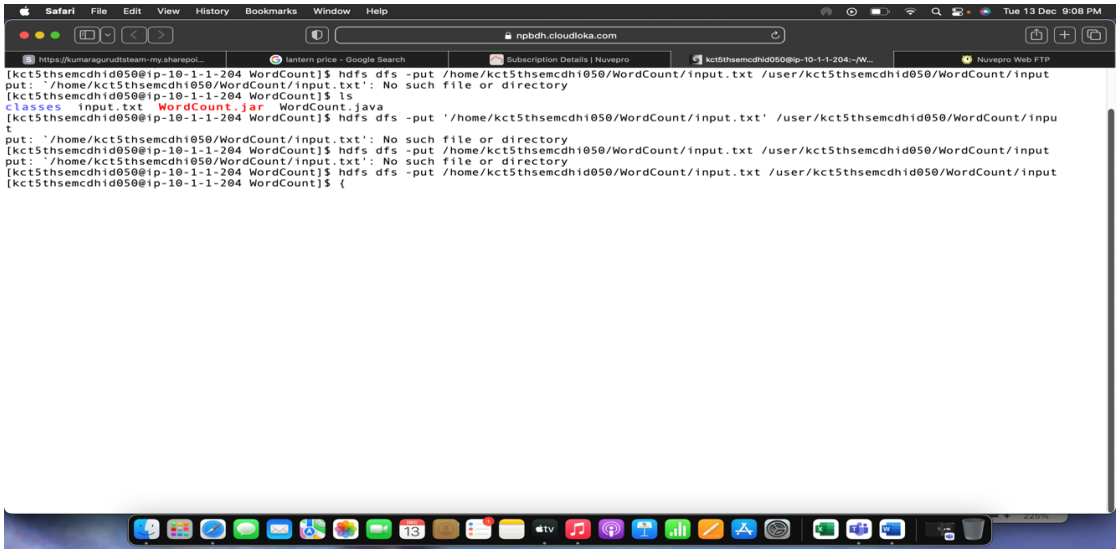


```
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hiii
hellooo
how r u
I am okay
how about u

Chocolates
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 12
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:25 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050   55 Dec 13 15:23 input.txt
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 2253 Dec 13 15:25 WordCount.ja
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ export HADOOP_CLASSPATH=$(hadoop c
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hado
cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop/.//*:/opt/cloudera/parcels/
hdfs/.//:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexe
.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-hdfs/.//*:/opt/cloudera/parcels
.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-yarn/lib/*:/opt/cloudera/p
hadoop-yarn/.//*
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ javac -classpath ${HADOOP_CLASSPAT
emcdhid050/WordCount/WordCount.java'
Note: /home/kct5thsemcdhid050/WordCount/WordCount.java uses or overrides a depr
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ jar -cvf WordCount.jar -C '/home/k
added manifest
adding: WordCount.class(in = 1814) (out= 910)(deflated 49%)
adding: WordCount$Map.class(in = 1657) (out= 690)(deflated 58%)
adding: WordCount$Reduce.class(in = 1627) (out= 686)(deflated 57%)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 16
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:30 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050   55 Dec 13 15:23 input.txt
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 3016 Dec 13 15:31 WordCount.ja
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid050/WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ {
```

10. CREATE A DIRECTORY IN HADOOP

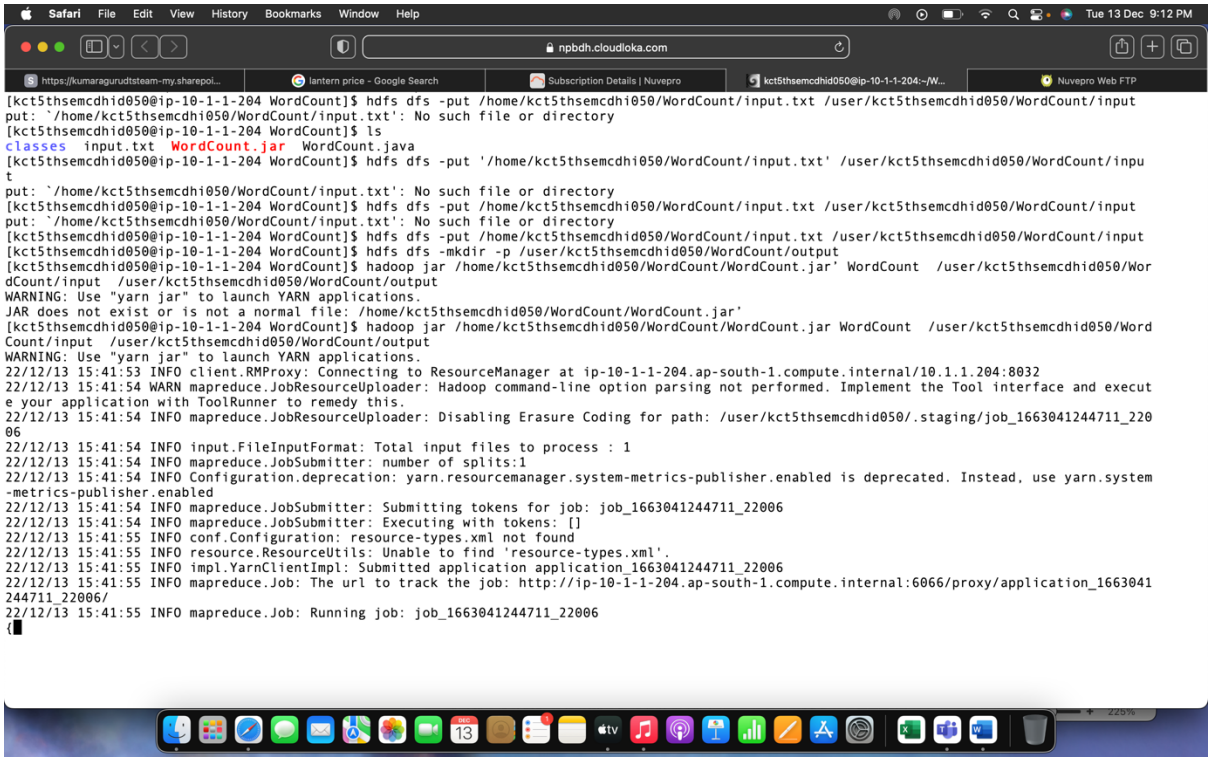


```
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hiii
hellooo
how r u
I am okay
how about u

Chocolates
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 12
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:25 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050   55 Dec 13 15:23 input.txt
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 2253 Dec 13 15:25 WordCount.ja
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ export HADOOP_CLASSPATH=$(hadoop c
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hado
cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop/.//*:/opt/cloudera/parcels/
hdfs/.//:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexe
.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-hdfs/.//*:/opt/cloudera/parcels
.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-yarn/lib/*:/opt/cloudera/p
hadoop-yarn/.//*
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ javac -classpath ${HADOOP_CLASSPAT
emcdhid050/WordCount/WordCount.java'
Note: /home/kct5thsemcdhid050/WordCount/WordCount.java uses or overrides a depr
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ jar -cvf WordCount.jar -C '/home/k
added manifest
adding: WordCount.class(in = 1814) (out= 910)(deflated 49%)
adding: WordCount$Map.class(in = 1657) (out= 690)(deflated 58%)
adding: WordCount$Reduce.class(in = 1627) (out= 686)(deflated 57%)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 16
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:30 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050   55 Dec 13 15:23 input.txt
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 3016 Dec 13 15:31 WordCount.ja
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid050/WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid050/WordCount/input{
```



```
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhi050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: `/home/kct5thsemcdhi050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ls
classes  input.txt  WordCount.jar  WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put '/home/kct5thsemcdhi050/WordCount/input.txt' /user/kct5thsemcdhid050/WordCount/inpu
t
put: `/home/kct5thsemcdhi050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhi050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: `/home/kct5thsemcdhi050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhi050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ {
```

11. PUT THE INPUT FILE IN LOCAL SYSTEM TO HADOOP DIRECTORY
12. CREATE AN OUTPUT DIRECTORY IN HDFS INSIDE THE WORDCOUNT



```
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhi050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: `/home/kct5thsemcdhi050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ls
classes  input.txt  WordCount.jar  WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put '/home/kct5thsemcdhi050/WordCount/input.txt' /user/kct5thsemcdhid050/WordCount/inpu
t
put: `/home/kct5thsemcdhi050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhi050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: `/home/kct5thsemcdhi050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhi050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid050/WordCount/output
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hadoop jar /home/kct5thsemcdhid050/WordCount/WordCount.jar' WordCount  /user/kct5thsemcdhid050/Wor
dCount/input  /user/kct5thsemcdhid050/WordCount/output
WARNING: Use "yarn jar" to launch YARN applications.
JAR does not exist or is not a normal file: /home/kct5thsemcdhid050/WordCount/WordCount.jar'
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hadoop jar /home/kct5thsemcdhid050/WordCount/WordCount.jar WordCount  /user/kct5thsemcdhid050/Word
Count/input  /user/kct5thsemcdhid050/WordCount/output
WARNING: Use "yarn jar" to launch YARN applications.
22/12/13 15:41:53 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
22/12/13 15:41:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execut
e your application with ToolRunner to remedy this.
22/12/13 15:41:54 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/kct5thsemcdhid050/.staging/job_1663041244711_220
06
22/12/13 15:41:54 INFO input.FileInputFormat: Total input files to process : 1
22/12/13 15:41:54 INFO mapreduce.JobSubmitter: number of splits:1
22/12/13 15:41:54 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system
-metrics-publisher.enabled
22/12/13 15:41:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1663041244711_22006
22/12/13 15:41:54 INFO mapreduce.JobSubmitter: Executing with tokens: []
22/12/13 15:41:55 INFO conf.Configuration: resource-types.xml not found
22/12/13 15:41:55 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/12/13 15:41:55 INFO impl.YarnClientImpl: Submitted application application_1663041244711_22006
22/12/13 15:41:55 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041
244711_22006/
22/12/13 15:41:55 INFO mapreduce.Job: Running job: job_1663041244711_22006
(
```

13. RUN THE MAP REDUCE PROGRAM:

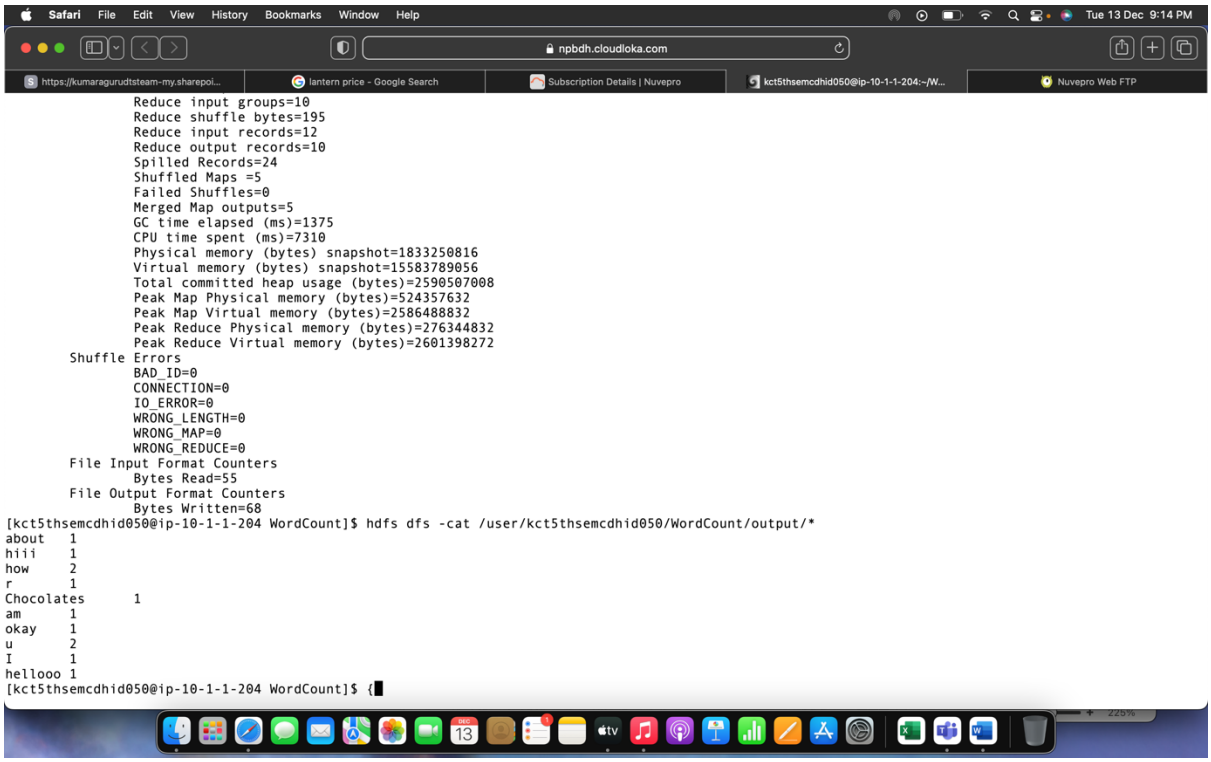
COMMAND: hadoop jar /home/<NAME>/wordcount/WordCount.jar’ WordCount /user/<NAME>/wordcount/input /user/<NAME>/wordcount/output

```
22/12/13 15:41:54 INFO input.FileInputFormat: Total input files to process : 1
22/12/13 15:41:54 INFO mapreduce.JobSubmitter: number of splits:1
22/12/13 15:41:54 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system
-metrics-publisher.enabled
22/12/13 15:41:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1663041244711_22006
22/12/13 15:41:54 INFO mapreduce.JobSubmitter: Executing with tokens: []
22/12/13 15:41:55 INFO conf.Configuration: resource-types.xml not found
22/12/13 15:41:55 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/12/13 15:41:55 INFO impl.YarnClientImpl: Submitted application application_1663041244711_22006
22/12/13 15:41:55 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041
244711_22006/
22/12/13 15:41:55 INFO mapreduce.Job: Running job: job_1663041244711_22006
(22/12/13 15:42:04 INFO mapreduce.Job: Job job_1663041244711_22006 running in uber mode : false
22/12/13 15:42:04 INFO mapreduce.Job:  map 0% reduce 0%
22/12/13 15:42:09 INFO mapreduce.Job:  map 100% reduce 0%
22/12/13 15:42:22 INFO mapreduce.Job:  map 100% reduce 100%
22/12/13 15:42:23 INFO mapreduce.Job: Job job_1663041244711_22006 completed successfully
22/12/13 15:42:24 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=215
                FILE: Number of bytes written=1338567
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=188
                HDFS: Number of bytes written=68
                HDFS: Number of read operations=28
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=10
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=5
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3343
                Total time spent by all reduces in occupied slots (ms)=52118
                Total time spent by all map tasks (ms)=3343
                Total time spent by all reduce tasks (ms)=52118
                Total vcore-milliseconds taken by all map tasks=3343
                Total vcore-milliseconds taken by all reduce tasks=52118
```

```
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=5
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3343
                Total time spent by all reduces in occupied slots (ms)=52118
                Total time spent by all map tasks (ms)=3343
                Total time spent by all reduce tasks (ms)=52118
                Total vcore-milliseconds taken by all map tasks=3343
                Total vcore-milliseconds taken by all reduce tasks=52118
                Total megabyte-milliseconds taken by all map tasks=3423232
                Total megabyte-milliseconds taken by all reduce tasks=53368832
        Map-Reduce Framework
                Map input records=7
                Map output records=12
                Map output bytes=102
                Map output materialized bytes=195
                Input split bytes=133
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=195
                Reduce input records=12
                Reduce output records=10
                Spilled Records=24
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=1375
                CPU time spent (ms)=7310
                Physical memory (bytes) snapshot=1833250816
                Virtual memory (bytes) snapshot=15583789056
                Total committed heap usage (bytes)=2590507008
                Peak Map Physical memory (bytes)=524357632
                Peak Map Virtual memory (bytes)=2586488832
                Peak Reduce Physical memory (bytes)=276344832
                Peak Reduce Virtual memory (bytes)=2601398272
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
```

14. VERIFY THE OUTPUT:

COMMAND: hdfs dfs -cat /user/<Name>/wordcount/output/*



U18ISI5202-BIG DATA LABORATORY

NAME: CHAARVIKA.A

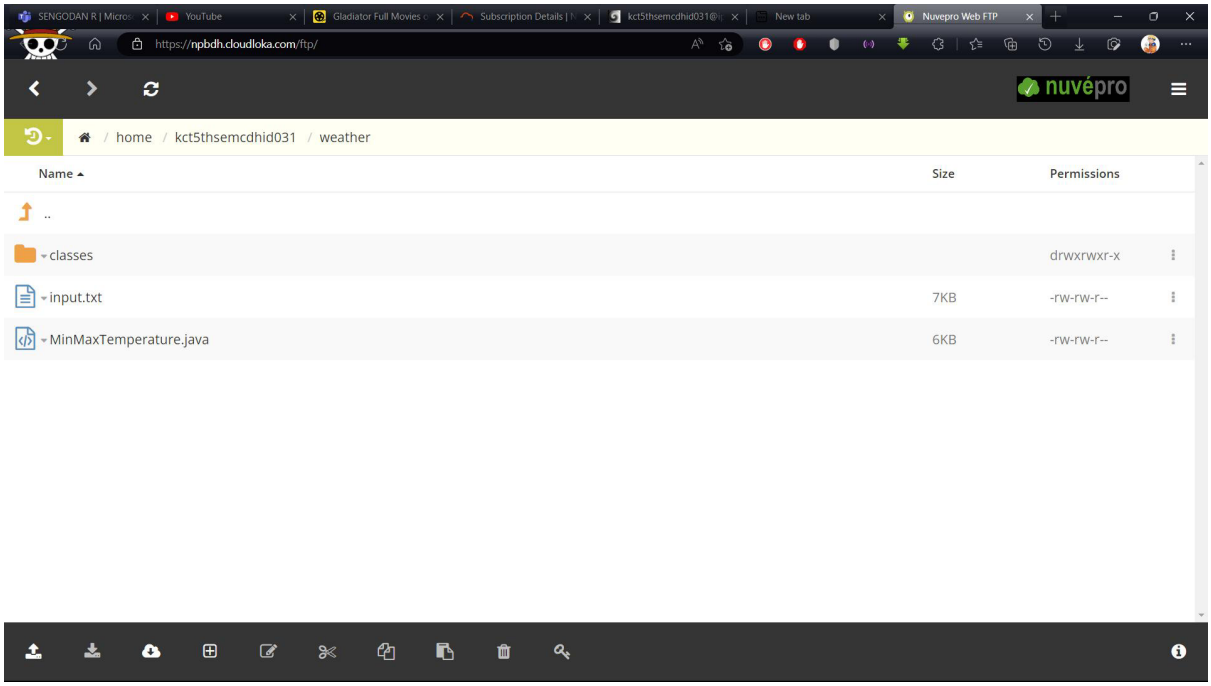
ROLL NUMBER: 20BIS012

S.NO: 4

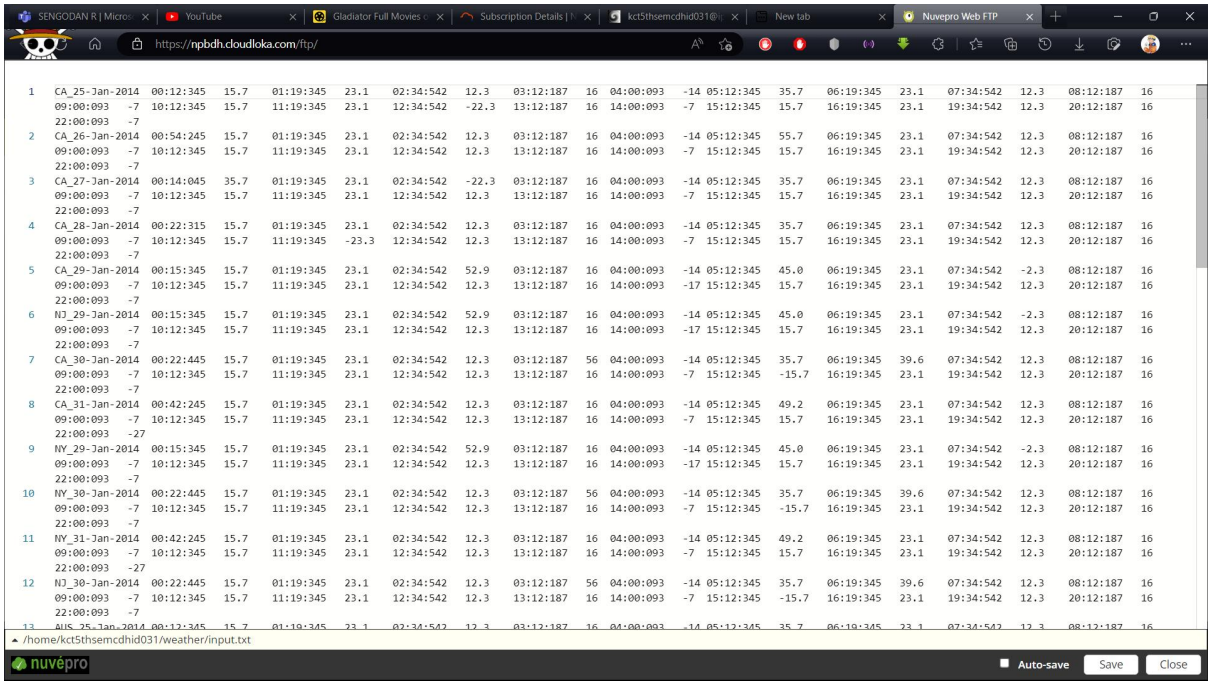
TOPIC : Implement a Map Reduce Program to analysetime-temperature statistics and generate report with max/min temperature

AIM:
Implement a Map Reduce Program to analyse time-temperature statistics and generate report with max/min temperature

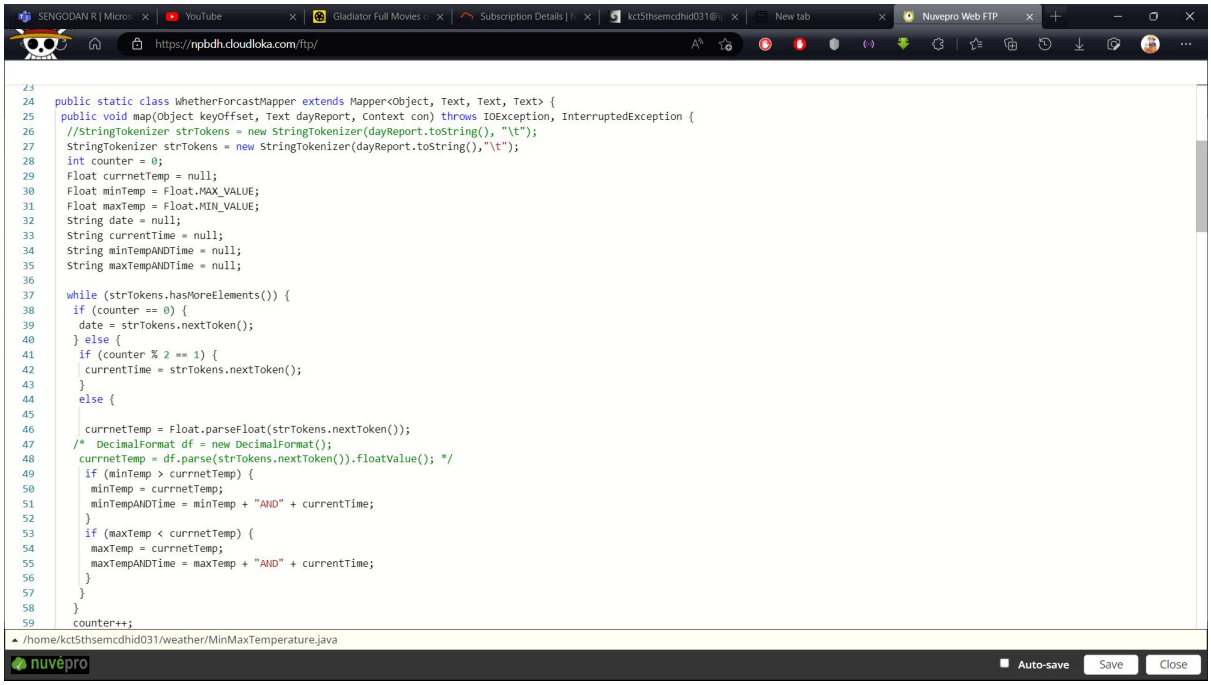
PROCEDURE:
STEP1-LOGIN TO THE WEB SHELL USING THE USERNAME AND PASWORD
STEP2-CREATE A FOLDER/ DIRECTORY FOR WEATHERIN HOME
STEP3-MOVE TO THAT FOLDER/DIRECTORY
STEP4-CREATE A INPUT FILE AND JAVA FILE INSIDE THE WORDCOUNT FOLDER

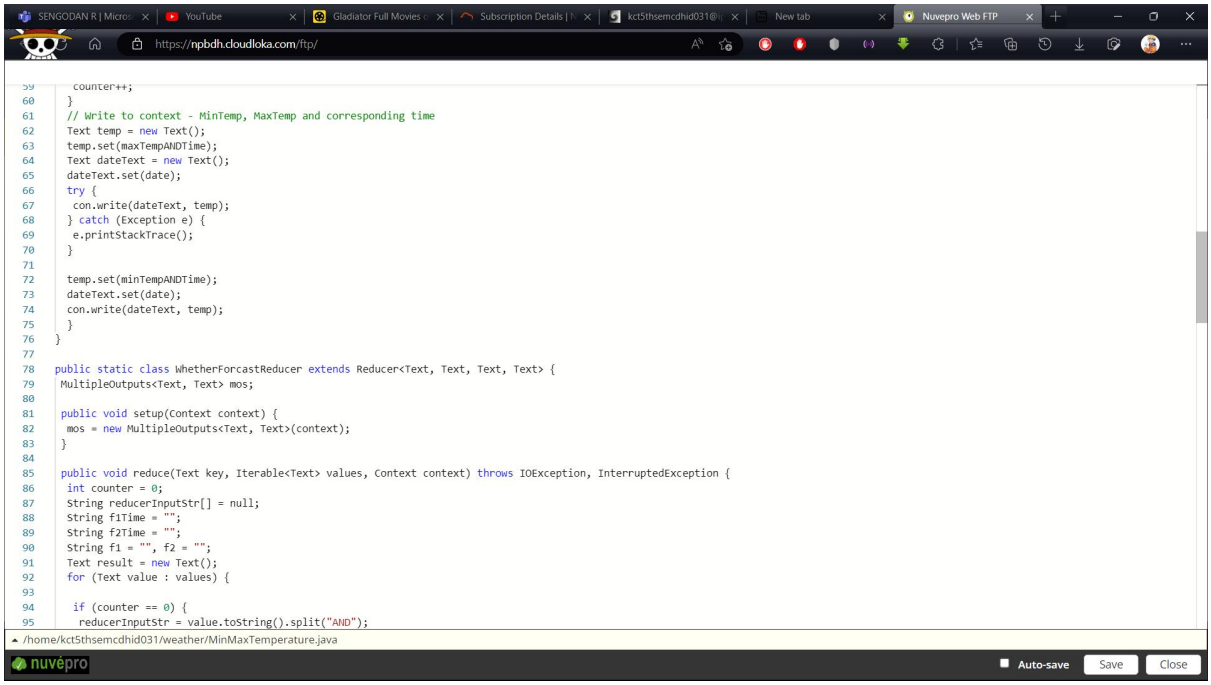


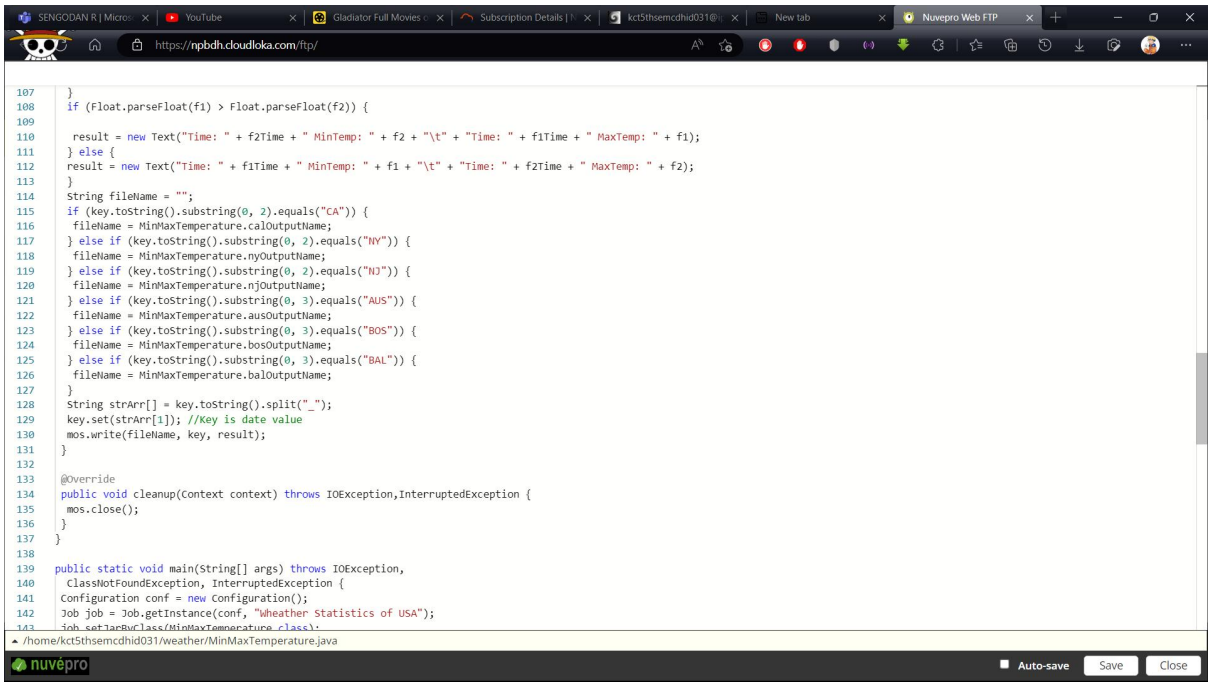
STEP5-PASTE THE INPUT IN THE input.txt

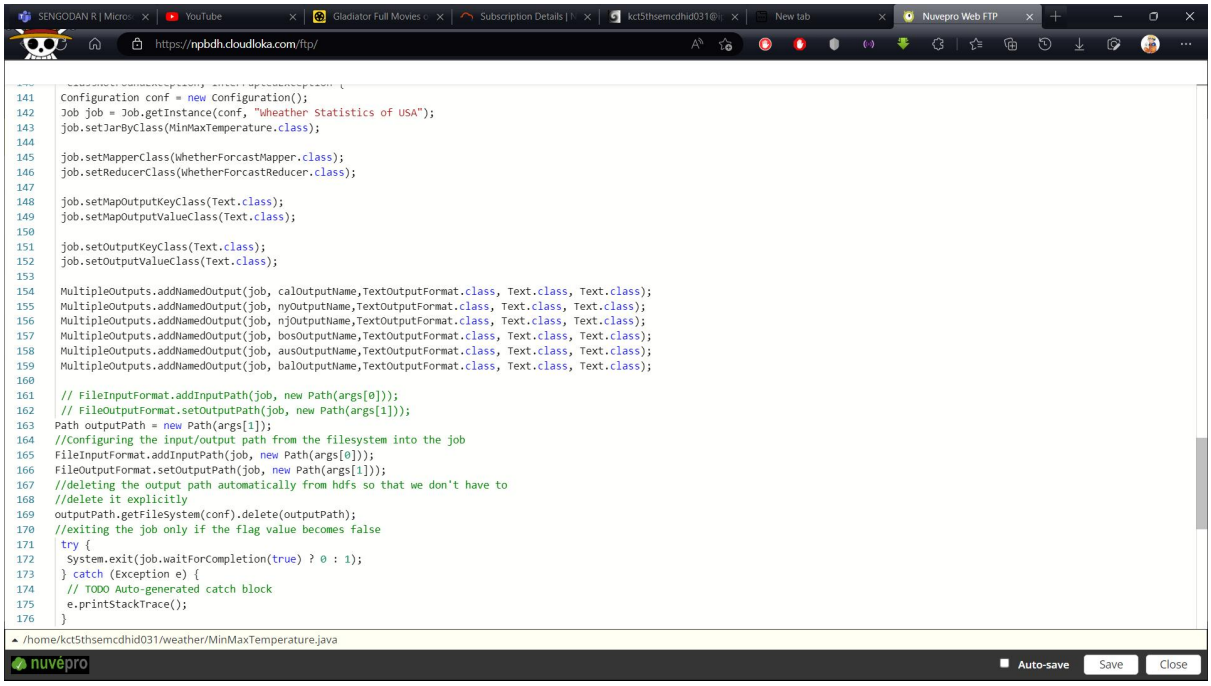


STEP6-PASTE THE SOURCE CODE IN THE FILE




SENGODAN R | Micros
YouTube
Gladiator Full Movies
Subscription Details
kct5thsemcdhid031@
New tab
Nuvepro Web FTP
https://npbdh.cloudloka.com/ftp/
counter++;
}
// Write to context - MinTemp, MaxTemp and corresponding time
Text temp = new Text();
temp.set(maxTempANDTime);
Text dateText = new Text();
dateText.set(date);
try {
con.write(dateText, temp);
} catch (Exception e) {
e.printStackTrace();
}
temp.set(minTempANDTime);
dateText.set(date);
con.write(dateText, temp);
}
}
public static class WhetherForcastReducer extends Reducer<Text, Text, Text, Text> {
MultipleOutputs<Text, Text> mos;
public void setup(Context context) {
mos = new MultipleOutputs<Text, Text>(context);
}
public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
int counter = 0;
String reducerInputStr[] = null;
String f1Time = "";
String f2Time = "";
String f1 = "", f2 = "";
Text result = new Text();
for (Text value : values) {
if (counter == 0) {
reducerInputStr = value.toString().split("AND");
/home/kct5thsemcdhid031/weather/MinMaxTemperature.java
nuvepro
Auto-save
Save
Close


SENGODAN R | Micros
YouTube
Gladiator Full Movies
Subscription Details
kct5thsemcdhid031@
New tab
Nuvepro Web FTP
https://npbdh.cloudloka.com/ftp/
}
if (Float.parseFloat(f1) > Float.parseFloat(f2)) {
result = new Text("Time: " + f2Time + " MinTemp: " + f2 + "\t" + "Time: " + f1Time + " MaxTemp: " + f1);
} else {
result = new Text("Time: " + f1Time + " MinTemp: " + f1 + "\t" + "Time: " + f2Time + " MaxTemp: " + f2);
}
String fileName = "";
if (key.toString().substring(0, 2).equals("CA")) {
fileName = MinMaxTemperature.calOutputName;
} else if (key.toString().substring(0, 2).equals("NY")) {
fileName = MinMaxTemperature.nyOutputName;
} else if (key.toString().substring(0, 2).equals("NJ")) {
fileName = MinMaxTemperature.njOutputName;
} else if (key.toString().substring(0, 3).equals("AUS")) {
fileName = MinMaxTemperature.ausOutputName;
} else if (key.toString().substring(0, 3).equals("BOS")) {
fileName = MinMaxTemperature.bosOutputName;
} else if (key.toString().substring(0, 3).equals("BAL")) {
fileName = MinMaxTemperature.balOutputName;
}
String strArr[] = key.toString().split("_");
key.set(strArr[1]); //Key is date value
mos.write(fileName, key, result);
}
@Override
public void cleanup(Context context) throws IOException,InterruptedException {
mos.close();
}
}
public static void main(String[] args) throws IOException,
ClassNotFoundException, InterruptedException {
Configuration conf = new Configuration();
Job job = Job.getInstance(conf, "Wheather Statistics of USA");
/home/kct5thsemcdhid031/weather/MinMaxTemperature.java
nuvepro
Auto-save
Save
Close



SOURCE CODE:

```
import java.io.IOException; import java.util.StringTokenizer; import
java.text.DecimalFormat;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.output.MultipleOutputs; import
org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat; import
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat; import
org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class MinMaxTemperature {
public static String calOutputName = "California"; public static String nyOutputName =
"Newyork"; public static String njOutputName = "Newjersy"; public static String
ausOutputName = "Austin";

public static String bosOutputName = "Boston"; public static String balOutputName =
"Baltimore";
```

```
public static class WhetherForcastMapper extends Mapper<Object, Text, Text, Text> { public
void map(Object keyOffset, Text dayReport, Context con) throws IOException,
InterruptedException {
//StringTokenizer strTokens = new StringTokenizer(dayReport.toString(), "\t");
StringTokenizer strTokens = new StringTokenizer(dayReport.toString(),"\t");

int counter = 0;
Float currnetTemp = null;
Float minTemp = Float.MAX_VALUE; Float maxTemp = Float.MIN_VALUE; String date = null;
String currentTime = null;
String minTempANDTime = null; String maxTempANDTime = null;

while (strTokens.hasMoreElements()) { if (counter == 0) {
date = strTokens.nextToken();
} else {

if (counter % 2 == 1) {
currentTime = strTokens.nextToken();
}
else {
currnetTemp = Float.parseFloat(strTokens.nextToken());
/* DecimalFormat df = new DecimalFormat();
currnetTemp = df.parse(strTokens.nextToken()).floatValue(); */
if (minTemp > currnetTemp) {
minTemp = currnetTemp;
minTempANDTime = minTemp + "AND" + currentTime;
}
if (maxTemp < currnetTemp) {
maxTemp = currnetTemp;
maxTempANDTime = maxTemp + "AND" + currentTime;
}
}
}
counter++;
}
// Write to context - MinTemp, MaxTemp and corresponding time Text temp = new Text();
temp.set(maxTempANDTime);
Text dateText = new Text();

dateText.set(date);
try {
con.write(dateText, temp); } catch (Exception e) { e.printStackTrace();

} temp.set(minTempANDTime); dateText.set(date); con.write(dateText, temp);
}
}
```

```
public static class WhetherForcastReducer extends Reducer<Text, Text, Text, Text> {
MultipleOutputs<Text, Text> mos;



public void setup(Context context) {
mos = new MultipleOutputs<Text, Text>(context); }

public void reduce(Text key, Iterable<Text> values, Context context) throws IOException,
InterruptedException {
int counter = 0;
String reducerInputStr[] = null;

String f1Time = "";
String f2Time = "";
String f1 = "", f2 = "";
Text result = new Text();
for (Text value : values) {
if (counter == 0) {
reducerInputStr = value.toString().split("AND"); f1 = reducerInputStr[0];

f1Time = reducerInputStr[1]; }

else {
reducerInputStr = value.toString().split("AND"); f2 = reducerInputStr[0];
f2Time = reducerInputStr[1];
}

counter = counter + 1;
}
if (Float.parseFloat(f1) > Float.parseFloat(f2)) {

result = new Text("Time: " + f2Time + " MinTemp: " + f2 + "\t" + "Time: " + f1Time + "
MaxTemp: " + f1);
} else {
result = new Text("Time: " + f1Time + " MinTemp: " + f1 + "\t" + "Time: " + f2Time + "
MaxTemp: " + f2);

}
String fileName = "";
if (key.toString().substring(0, 2).equals("CA")) { fileName =
MinMaxTemperature.calOutputName;
} else if (key.toString().substring(0, 2).equals("NY")) { fileName =
MinMaxTemperature.nyOutputName;
} else if (key.toString().substring(0, 2).equals("NJ")) { fileName =
MinMaxTemperature.njOutputName;
} else if (key.toString().substring(0, 3).equals("AUS")) { fileName =
```

```
MinMaxTemperature.ausOutputName;
} else if (key.toString().substring(0, 3).equals("BOS")) { fileName =
MinMaxTemperature.bosOutputName;
} else if (key.toString().substring(0, 3).equals("BAL")) { fileName =
MinMaxTemperature.balOutputName;
}
String strArr[] = key.toString().split("_"); key.set(strArr[1]); //Key is date value
mos.write(fileName, key, result);
}

@Override

public void cleanup(Context context) throws IOException,InterruptedException {
mos.close();
}
}

public static void main(String[] args) throws IOException, ClassNotFoundException,
InterruptedException { Configuration conf = new Configuration();
Job job = Job.getInstance(conf, "Wheather Statistics of USA");
job.setJarByClass(MinMaxTemperature.class);

job.setMapperClass(WhetherForcastMapper.class);
job.setReducerClass(WhetherForcastReducer.class);



job.setMapOutputKeyClass(Text.class); job.setMapOutputValueClass(Text.class);



job.setOutputKeyClass(Text.class);

job.setOutputValueClass(Text.class);

MultipleOutputs.addNamedOutput(job, calOutputName,TextOutputFormat.class, Text.class,
Text.class);
MultipleOutputs.addNamedOutput(job, nyOutputName,TextOutputFormat.class, Text.class,
Text.class);

MultipleOutputs.addNamedOutput(job, njOutputName,TextOutputFormat.class, Text.class,
Text.class);
MultipleOutputs.addNamedOutput(job, bosOutputName,TextOutputFormat.class,
Text.class, Text.class);
```

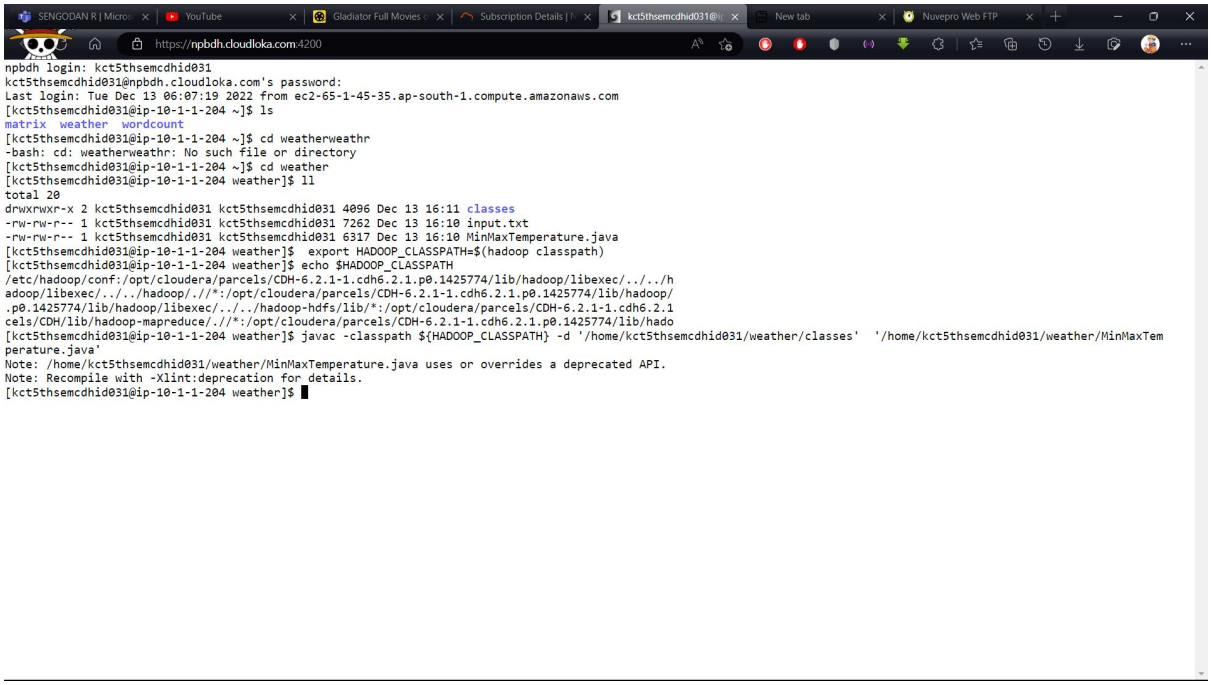
```
MultipleOutputs.addNamedOutput(job, ausOutputName,TextOutputFormat.class,
Text.class, Text.class);
MultipleOutputs.addNamedOutput(job, balOutputName,TextOutputFormat.class,
Text.class, Text.class);

// FileInputFormat.addInputPath(job, new Path(args[0]));
// FileOutputFormat.setOutputPath(job, new Path(args[1]));
Path outputPath = new Path(args[1]);
//Configuring the input/output path from the filesystem into the job
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
//deleting the output path automatically from hdfs so that we don't have to //delete it
explicitly
outputPath.getFileSystem(conf).delete(outputPath);
//exiting the job only if the flag value becomes false
try {
System.exit(job.waitForCompletion(true) ? 0 : 1);
} catch (Exception e) {
// TODO Auto-generated catch block
e.printStackTrace();
}
}

}
```

STEP7-CREATE A DIRECTORY/ FOLDER CLASSES INSIDE THE WEATHER FOLDER STEP8-SET THE PATH FOR JAVA FILE

COMMAND: export HADOOP_CLASSPATH=$(hadoop classpath) echo $HADOOP_CLASSPATH

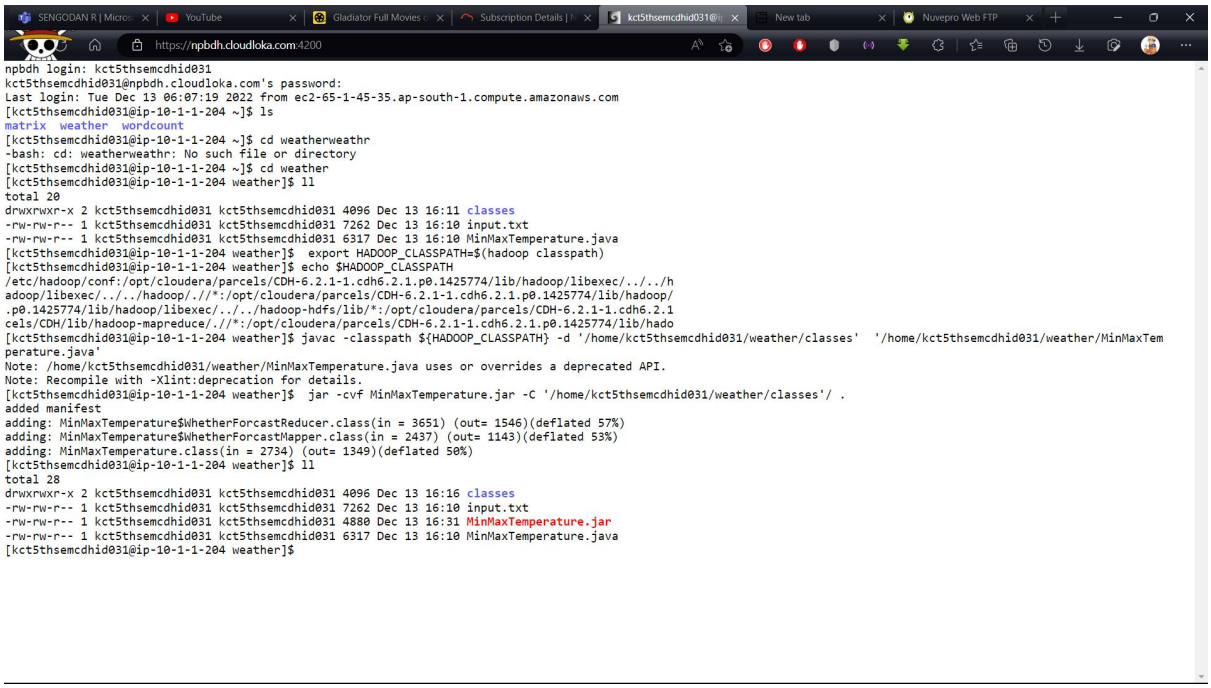
STEP9-COMPILE THE JAVAFILE

COMMAND: javac -classpath ${HADOOP_CLASSPATH} -d ‘/home/<NAME>/wordcount/classes’ ‘/home/<NAME>/wordcount/WordCount.java’



STEP10-CREATE A JAR FILE

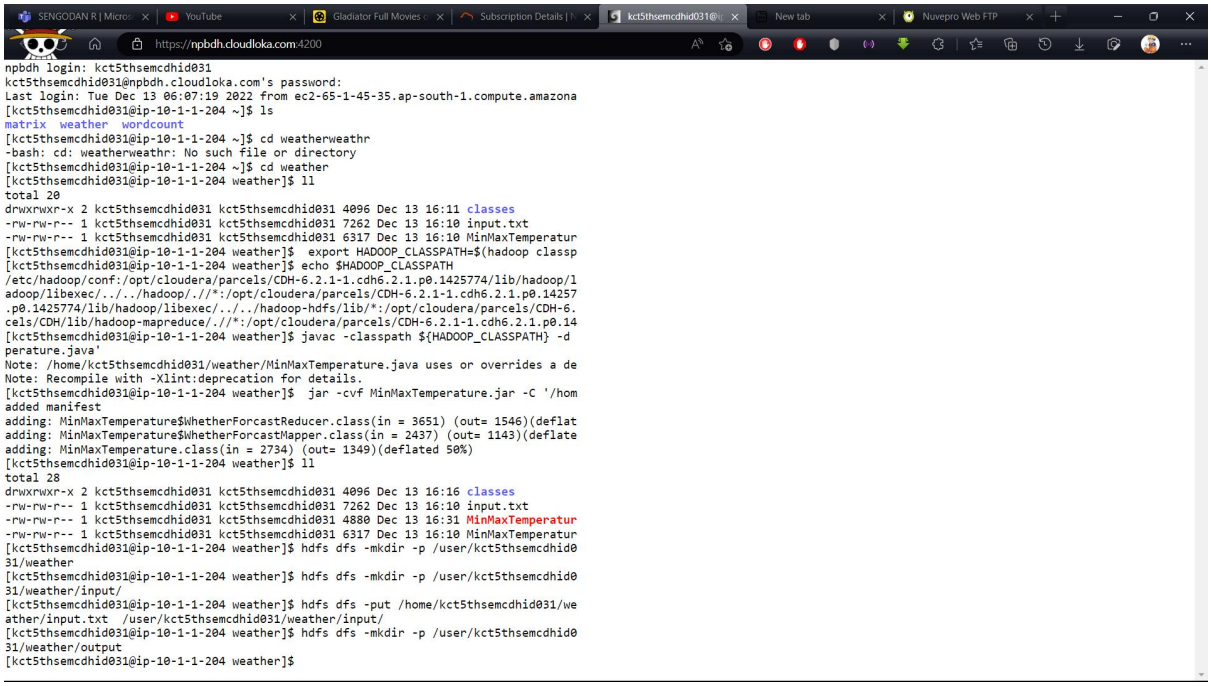
COMMAND: jar -cvf WordCount.jar -C '/home/<NAME>/wordcount/classes'/ .



STEP11-CREATE A DIRECTORY IN HADOOP

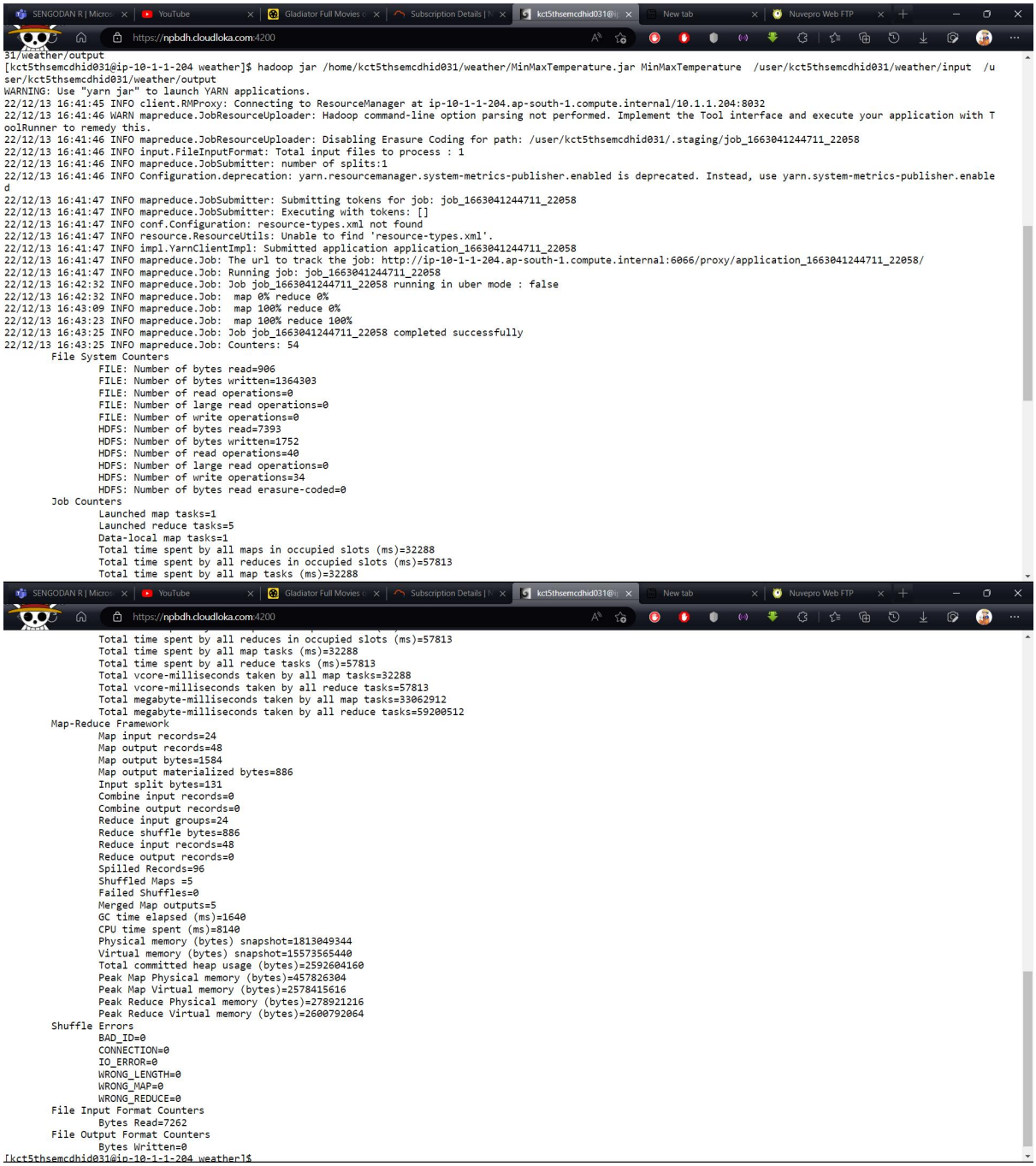
STEP12-CREATE AND PUT THE INPUT FILE IN LOCAL SYSTEM TO HADOOP DIRECTORY

STEP13-CREATE A OUTPUT DIRECTORY IN HDFS INSIDE THE WEATHER



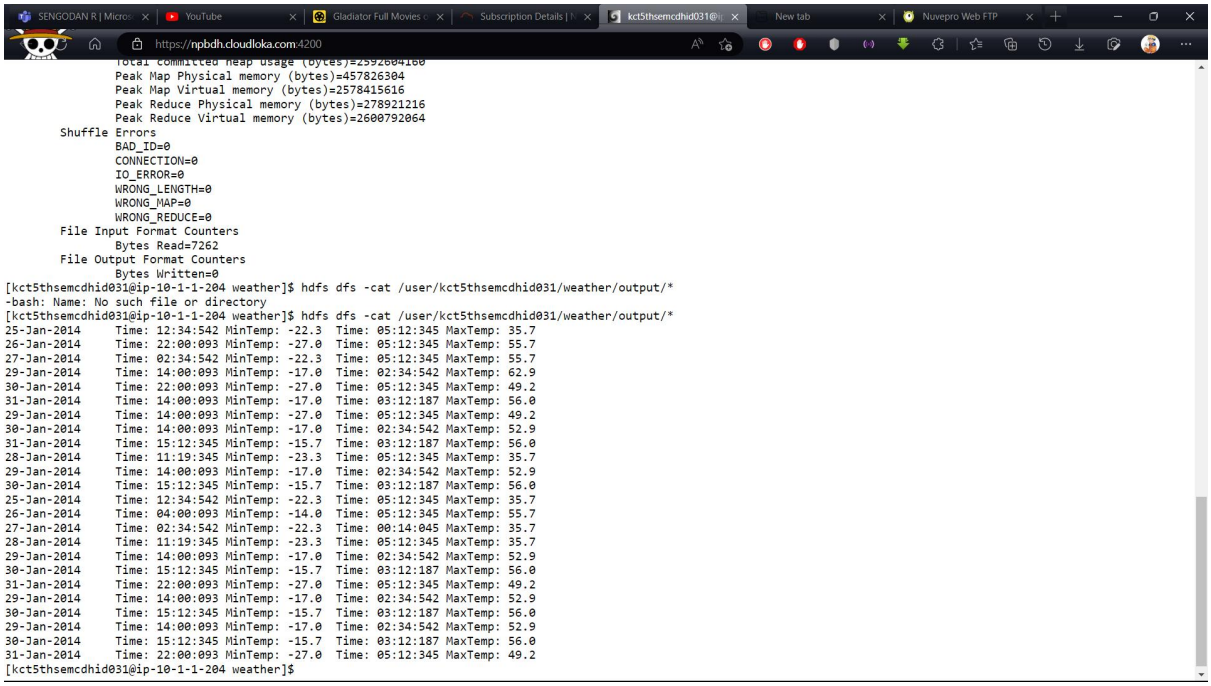
STEP14-RUN THE MAP REDUCE PROGRAM

COMMAND: hadoop jar /home/<NAME>/wordcount/WordCount.jar' WordCount /user/<NAME>/wordcount/input /user/<NAME>/wordcount/output



STEP15-VERIFY THE OUTPUT:

COMMAND: hdfs dfs -cat /user/<Name>/wordcount/output/*


SENGODAN R | Micros
YouTube
Gladiator Full Movies
Subscription Details |
kct5thsemcdhid031@i
New tab
Nuvepro Web FTP
https://npbdh.cloudloka.com:4200
Total Committed Heap Usage (bytes)=2592684160
Peak Map Physical memory (bytes)=457826304
Peak Map Virtual memory (bytes)=2578415616
Peak Reduce Physical memory (bytes)=278921216
Peak Reduce Virtual memory (bytes)=2600792064
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=7262
File Output Format Counters
Bytes Written=0
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ hdfs dfs -cat /user/kct5thsemcdhid031/weather/output/*
-bash: Name: No such file or directory
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ hdfs dfs -cat /user/kct5thsemcdhid031/weather/output/*
25-Jan-2014 Time: 12:34:542 MinTemp: -22.3 Time: 05:12:345 MaxTemp: 35.7
26-Jan-2014 Time: 22:00:093 MinTemp: -27.0 Time: 05:12:345 MaxTemp: 55.7
27-Jan-2014 Time: 02:34:542 MinTemp: -22.3 Time: 05:12:345 MaxTemp: 55.7
29-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 02:34:542 MaxTemp: 62.9
30-Jan-2014 Time: 22:00:093 MinTemp: -27.0 Time: 05:12:345 MaxTemp: 49.2
31-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 03:12:187 MaxTemp: 56.0
29-Jan-2014 Time: 14:00:093 MinTemp: -27.0 Time: 05:12:345 MaxTemp: 49.2
30-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 02:34:542 MaxTemp: 52.9
31-Jan-2014 Time: 15:12:345 MinTemp: -15.7 Time: 03:12:187 MaxTemp: 56.0
28-Jan-2014 Time: 11:19:345 MinTemp: -23.3 Time: 05:12:345 MaxTemp: 35.7
29-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 02:34:542 MaxTemp: 52.9
30-Jan-2014 Time: 15:12:345 MinTemp: -15.7 Time: 03:12:187 MaxTemp: 56.0
25-Jan-2014 Time: 12:34:542 MinTemp: -22.3 Time: 05:12:345 MaxTemp: 35.7
26-Jan-2014 Time: 04:00:093 MinTemp: -14.0 Time: 05:12:345 MaxTemp: 55.7
27-Jan-2014 Time: 02:34:542 MinTemp: -22.3 Time: 00:14:045 MaxTemp: 35.7
28-Jan-2014 Time: 11:19:345 MinTemp: -23.3 Time: 05:12:345 MaxTemp: 35.7
29-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 02:34:542 MaxTemp: 52.9
30-Jan-2014 Time: 15:12:345 MinTemp: -15.7 Time: 03:12:187 MaxTemp: 56.0
31-Jan-2014 Time: 22:00:093 MinTemp: -27.0 Time: 05:12:345 MaxTemp: 49.2
29-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 02:34:542 MaxTemp: 52.9
30-Jan-2014 Time: 15:12:345 MinTemp: -15.7 Time: 03:12:187 MaxTemp: 56.0
29-Jan-2014 Time: 14:00:093 MinTemp: -17.0 Time: 02:34:542 MaxTemp: 52.9
30-Jan-2014 Time: 15:12:345 MinTemp: -15.7 Time: 03:12:187 MaxTemp: 56.0
31-Jan-2014 Time: 22:00:093 MinTemp: -27.0 Time: 05:12:345 MaxTemp: 49.2
[kct5thsemcdhid031@ip-10-1-1-204 weather]$

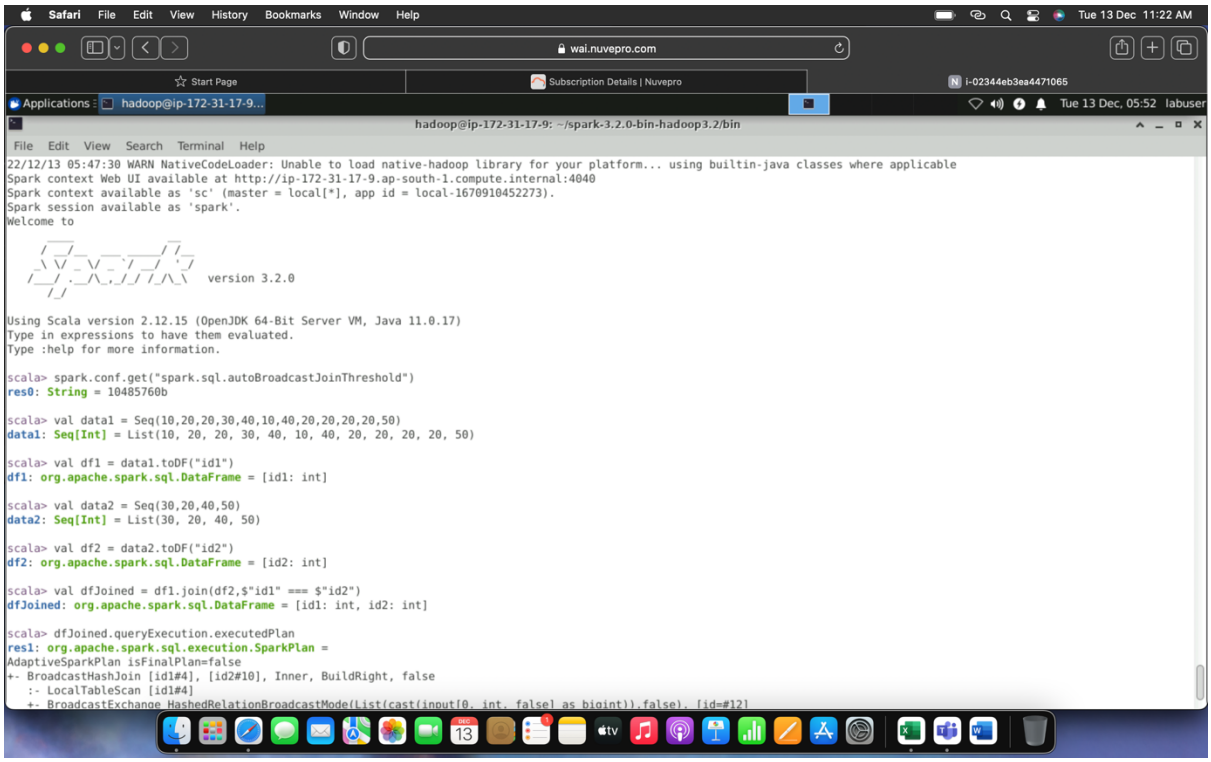
U18ISI5202-BIG DATA LABORATORY

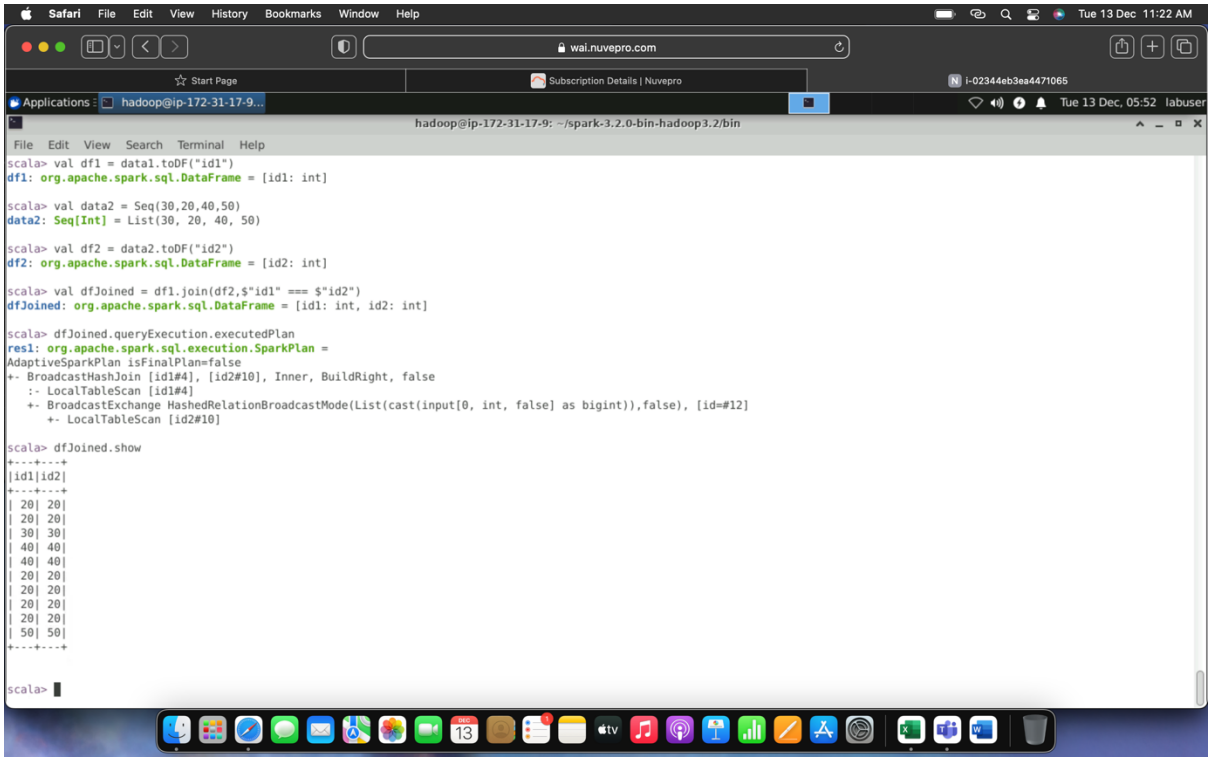
NAME: CHAARVIKA.A

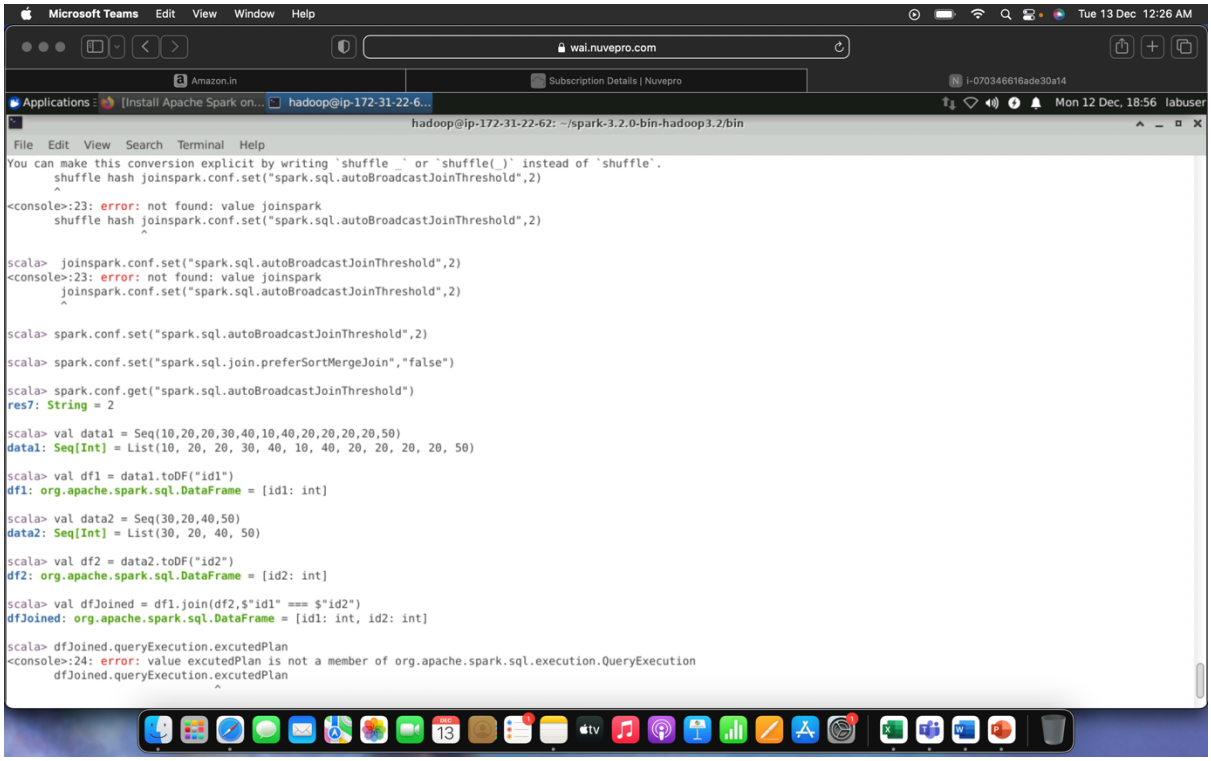
ROLL NUMBER: 20BIS012

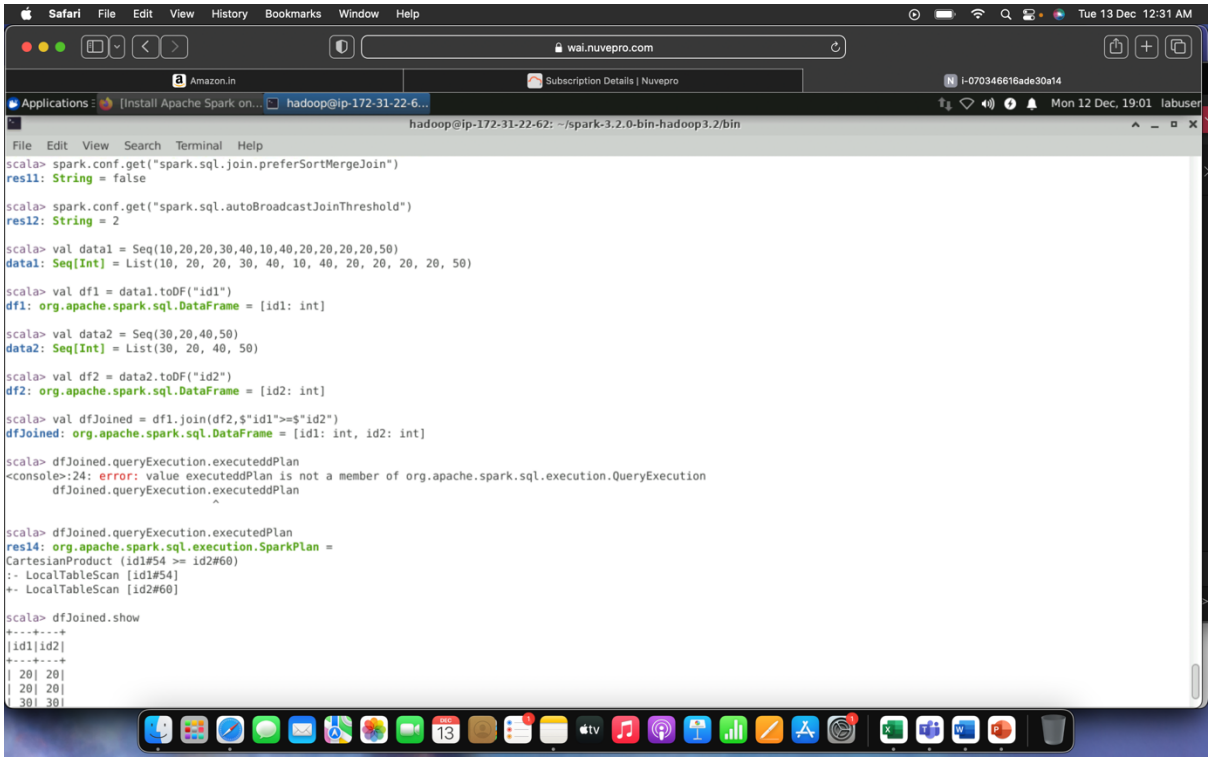
S.NO: 8

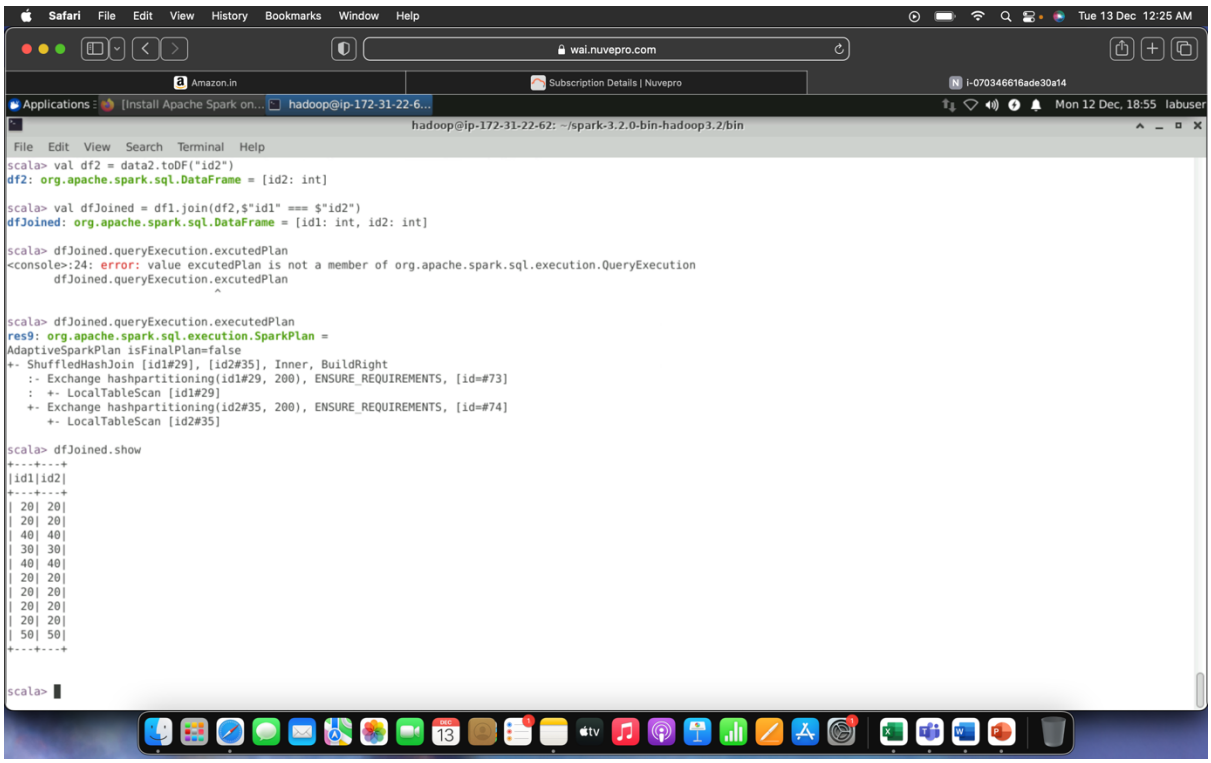
TOPIC : Perform simple join using Mapper in Spark




Safari File Edit View History Bookmarks Window Help
Tue 13 Dec 11:22 AM
wai.nuvepro.com
Start Page
Subscription Details | Nuvepro
i-02344eb3ea4471065
Applications
hadoop@ip-172-31-17-9...
Tue 13 Dec, 05:52 labuser
hadoop@ip-172-31-17-9: ~/spark-3.2.0-bin-hadoop3.2/bin
File Edit View Search Terminal Help
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int]
scala> val data2 = Seq(30,20,40,50)
data2: Seq[Int] = List(30, 20, 40, 50)
scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]
scala> val dfJoined = df1.join(df2,$"id1" === $"id2")
dfJoined: org.apache.spark.sql.DataFrame = [id1: int, id2: int]
scala> dfJoined.queryExecution.executedPlan
res1: org.apache.spark.sql.execution.SparkPlan =
AdaptiveSparkPlan isFinalPlan=false
+- BroadcastHashJoin [id1#4], [id2#10], Inner, BuildRight, false
:- LocalTableScan [id1#4]
+- BroadcastExchange HashedRelationBroadcastMode(List(cast(input[0, int, false] as bigint)),false), [id=#12]
+- LocalTableScan [id2#10]
scala> dfJoined.show
+---+---+
|id1|id2|
+---+---+
20	20
20	20
30	30
40	40
40	40
20	20
20	20
20	20
20	20
50	50
+---+---+
scala>


Microsoft Teams Edit View Window Help
Tue 13 Dec 12:26 AM
wai.nuvepro.com
Amazon.in
Subscription Details | Nuvepro
i-070346616ade30a14
Applications
[Install Apache Spark on...
hadoop@ip-172-31-22-6...
Mon 12 Dec, 18:56 labuser
hadoop@ip-172-31-22-62: ~/spark-3.2.0-bin-hadoop3.2/bin
File Edit View Search Terminal Help
You can make this conversion explicit by writing `shuffle _` or `shuffle(_)` instead of `shuffle`.
shuffle hash joinspark.conf.set("spark.sql.autoBroadcastJoinThreshold",2)
^
<console>:23: error: not found: value joinspark
shuffle hash joinspark.conf.set("spark.sql.autoBroadcastJoinThreshold",2)
^
scala> joinspark.conf.set("spark.sql.autoBroadcastJoinThreshold",2)
<console>:23: error: not found: value joinspark
joinspark.conf.set("spark.sql.autoBroadcastJoinThreshold",2)
^
scala> spark.conf.set("spark.sql.autoBroadcastJoinThreshold",2)
scala> spark.conf.set("spark.sql.join.preferSortMergeJoin","false")
scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold")
res7: String = 2
scala> val data1 = Seq(10,20,20,30,40,10,40,20,20,20,20,50)
data1: Seq[Int] = List(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int]
scala> val data2 = Seq(30,20,40,50)
data2: Seq[Int] = List(30, 20, 40, 50)
scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]
scala> val dfJoined = df1.join(df2,$"id1" === $"id2")
dfJoined: org.apache.spark.sql.DataFrame = [id1: int, id2: int]
scala> dfJoined.queryExecution.excutedPlan
<console>:24: error: value excutedPlan is not a member of org.apache.spark.sql.execution.QueryExecution
dfJoined.queryExecution.excutedPlan
^


Safari File Edit View History Bookmarks Window Help
Tue 13 Dec 12:31 AM
wai.nuvepro.com
Amazon.in
Subscription Details | Nuvepro
i-070346616ade30a14
Applications [Install Apache Spark on... hadoop@ip-172-31-22-6...
Mon 12 Dec, 19:01 labuser
hadoop@ip-172-31-22-62: ~/spark-3.2.0-bin-hadoop3.2/bin
File Edit View Search Terminal Help
scala> spark.conf.get("spark.sql.join.preferSortMergeJoin")
res11: String = false
scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold")
res12: String = 2
scala> val data1 = Seq(10,20,20,30,40,10,40,20,20,20,20,50)
data1: Seq[Int] = List(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int]
scala> val data2 = Seq(30,20,40,50)
data2: Seq[Int] = List(30, 20, 40, 50)
scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]
scala> val dfJoined = df1.join(df2,$"id1">=$"id2")
dfJoined: org.apache.spark.sql.DataFrame = [id1: int, id2: int]
scala> dfJoined.queryExecution.executeddPlan
<console>:24: error: value executeddPlan is not a member of org.apache.spark.sql.execution.QueryExecution
dfJoined.queryExecution.executeddPlan
^
scala> dfJoined.queryExecution.executedPlan
res14: org.apache.spark.sql.execution.SparkPlan =
CartesianProduct (id1#54 >= id2#60)
:- LocalTableScan [id1#54]
+- LocalTableScan [id2#60]
scala> dfJoined.show
+---+---+
|id1|id2|
+---+---+
20	20
20	20
30	30


Safari File Edit View History Bookmarks Window Help
Tue 13 Dec 12:25 AM
wai.nuvepro.com
Amazon.in
Subscription Details | Nuvepro
i-070346616ade30a14
Applications [Install Apache Spark on... hadoop@ip-172-31-22-6...
Mon 12 Dec, 18:55 labuser
hadoop@ip-172-31-22-62: ~/spark-3.2.0-bin-hadoop3.2/bin
File Edit View Search Terminal Help
scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]
scala> val dfJoined = df1.join(df2,$"id1" === $"id2")
dfJoined: org.apache.spark.sql.DataFrame = [id1: int, id2: int]
scala> dfJoined.queryExecution.excutedPlan
<console>:24: error: value excutedPlan is not a member of org.apache.spark.sql.execution.QueryExecution
dfJoined.queryExecution.excutedPlan
^
scala> dfJoined.queryExecution.executedPlan
res9: org.apache.spark.sql.execution.SparkPlan =
AdaptiveSparkPlan isFinalPlan=false
+- ShuffledHashJoin [id1#29], [id2#35], Inner, BuildRight
:- Exchange hashpartitioning(id1#29, 200), ENSURE_REQUIREMENTS, [id=#73]
: +- LocalTableScan [id1#29]
+- Exchange hashpartitioning(id2#35, 200), ENSURE_REQUIREMENTS, [id=#74]
+- LocalTableScan [id2#35]
scala> dfJoined.show
+---+---+
|id1|id2|
+---+---+
20	20
20	20
40	40
30	30
40	40
20	20
20	20
20	20
20	20
50	50
+---+---+
scala>

U18ISI5202-BIG DATA LABORATORY

NAME: CHAARVIKA.A

ROLL NUMBER: 20BIS012

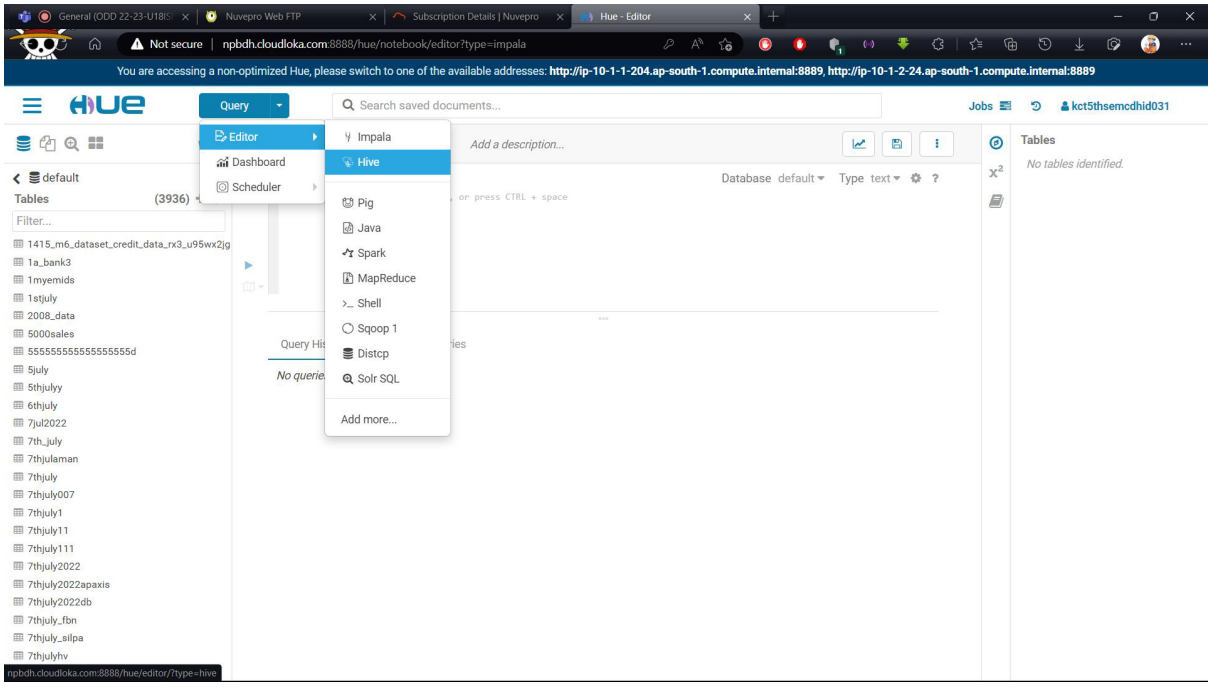
S.NO: 9

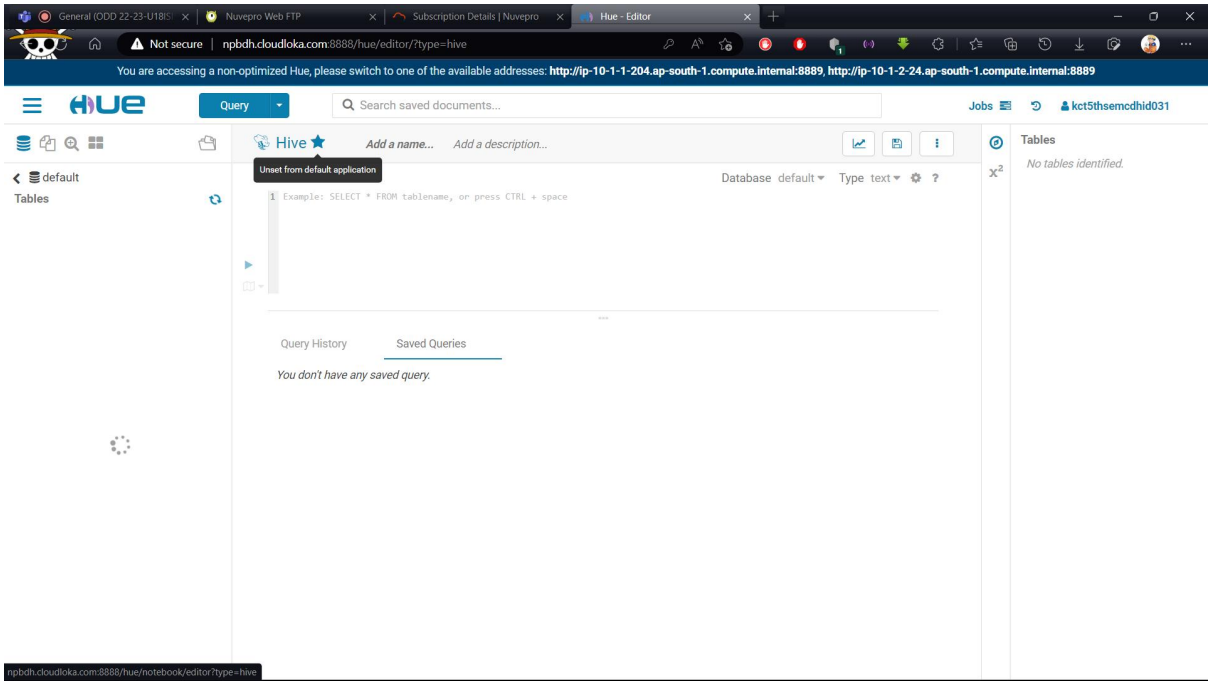
TOPIC : Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes

AIM:

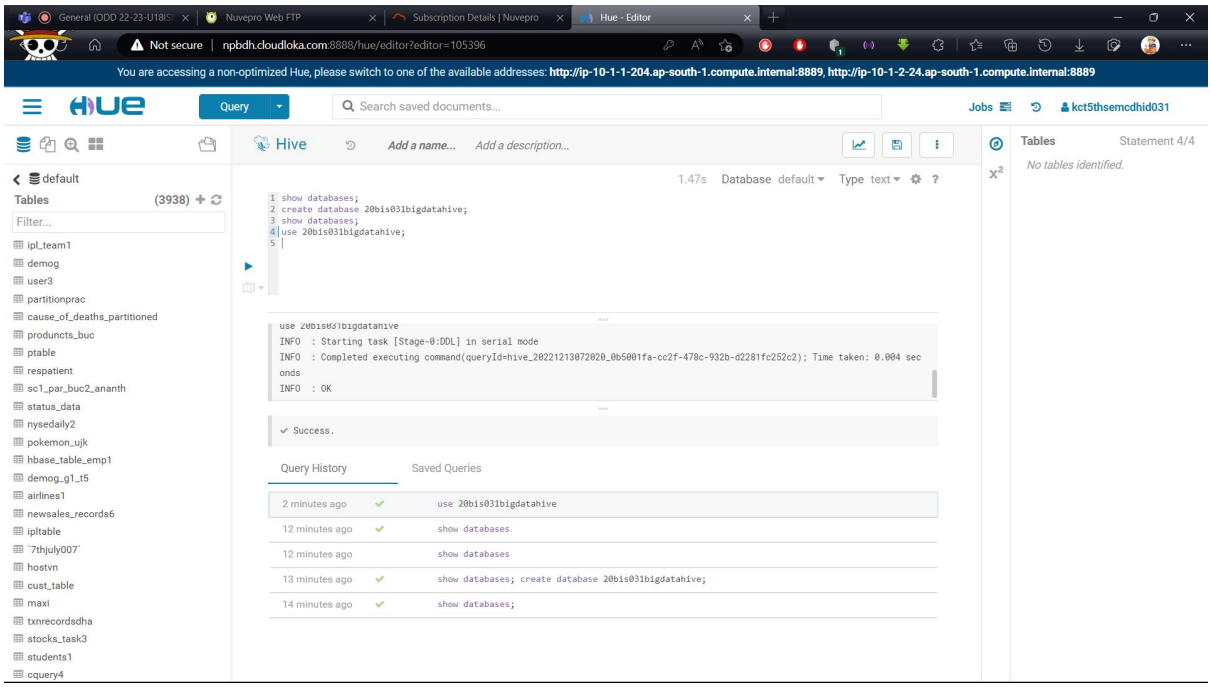
Install and Run Hive then use Hive to create, alter, and drop databases, tables, views,functions, and indexes

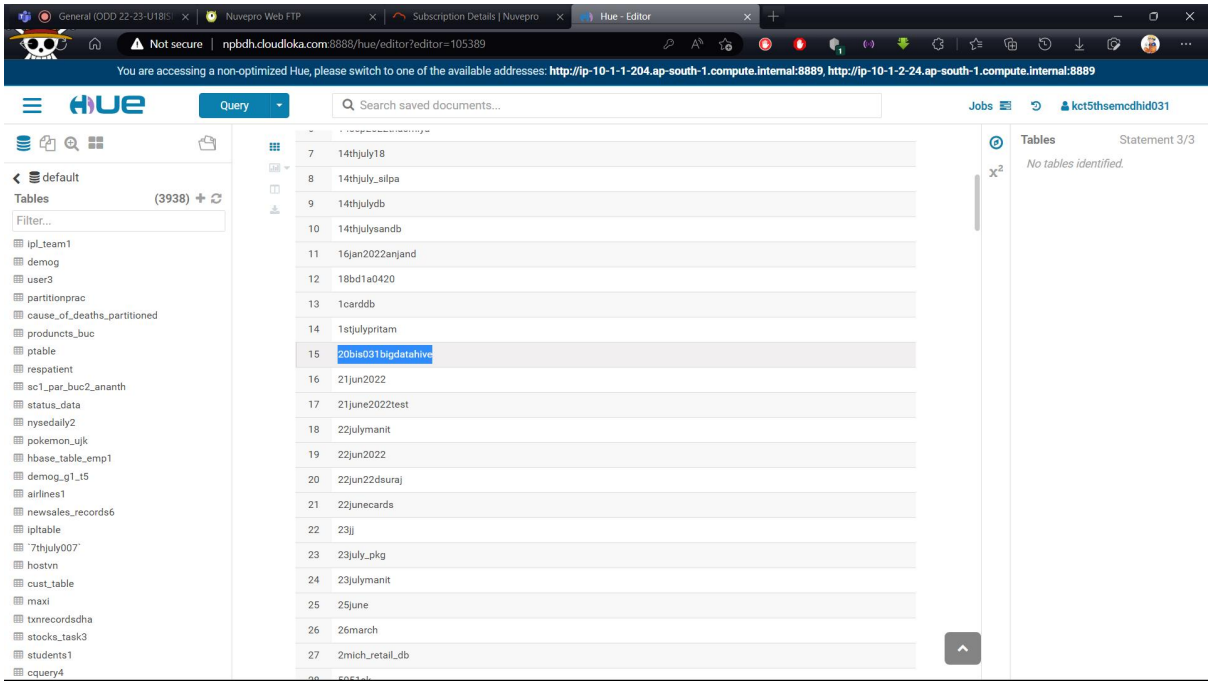
PROCEDURE:
STEP1:Opening HIVE from HUE portal cloudera



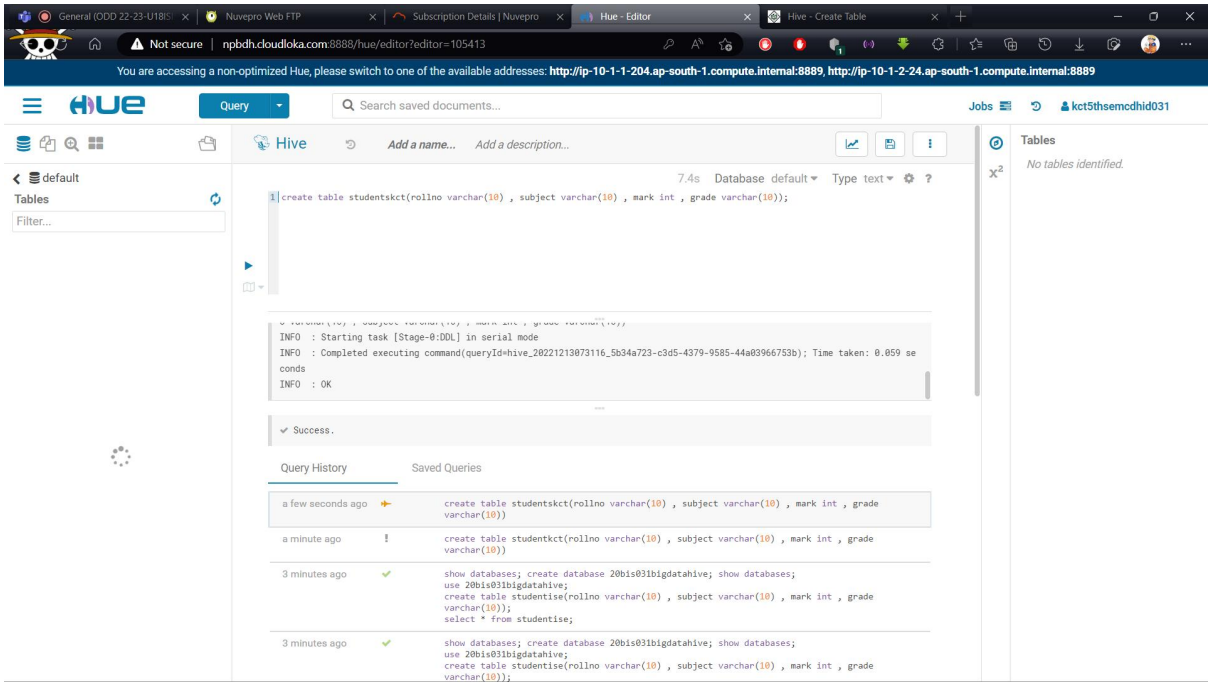


STEP2: Create database
create database 20bis031bigdatahive; show databases; use 20bis031bigdatahive;





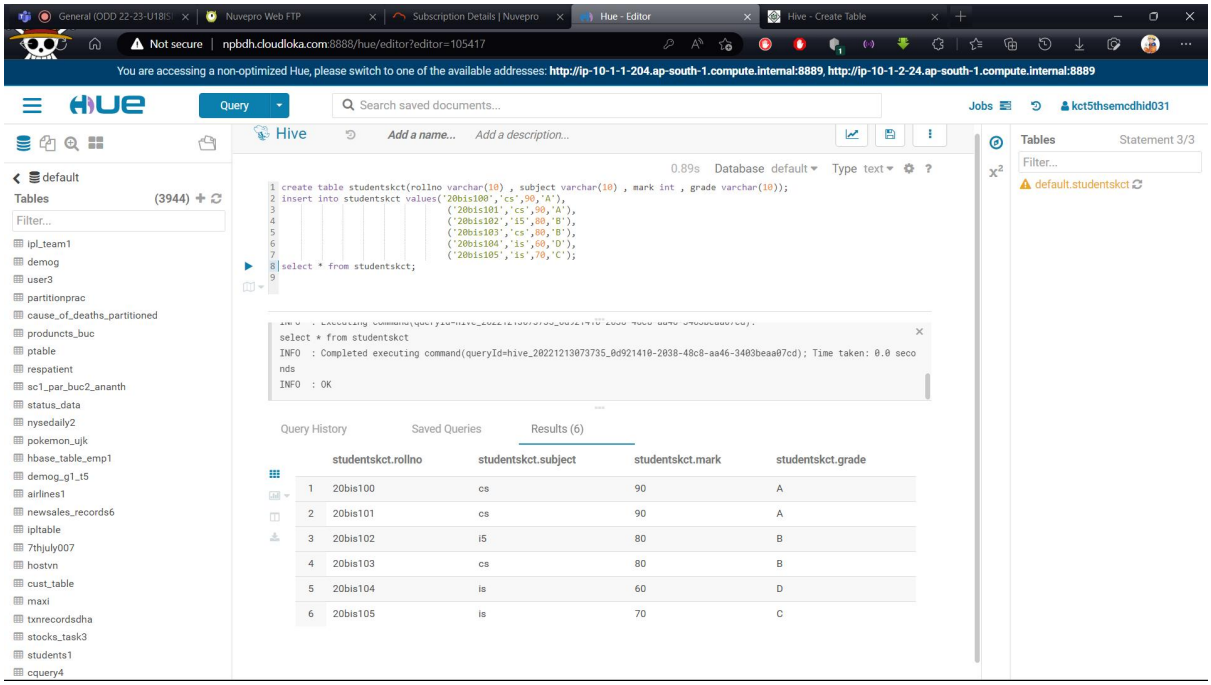
STEP3:Create table



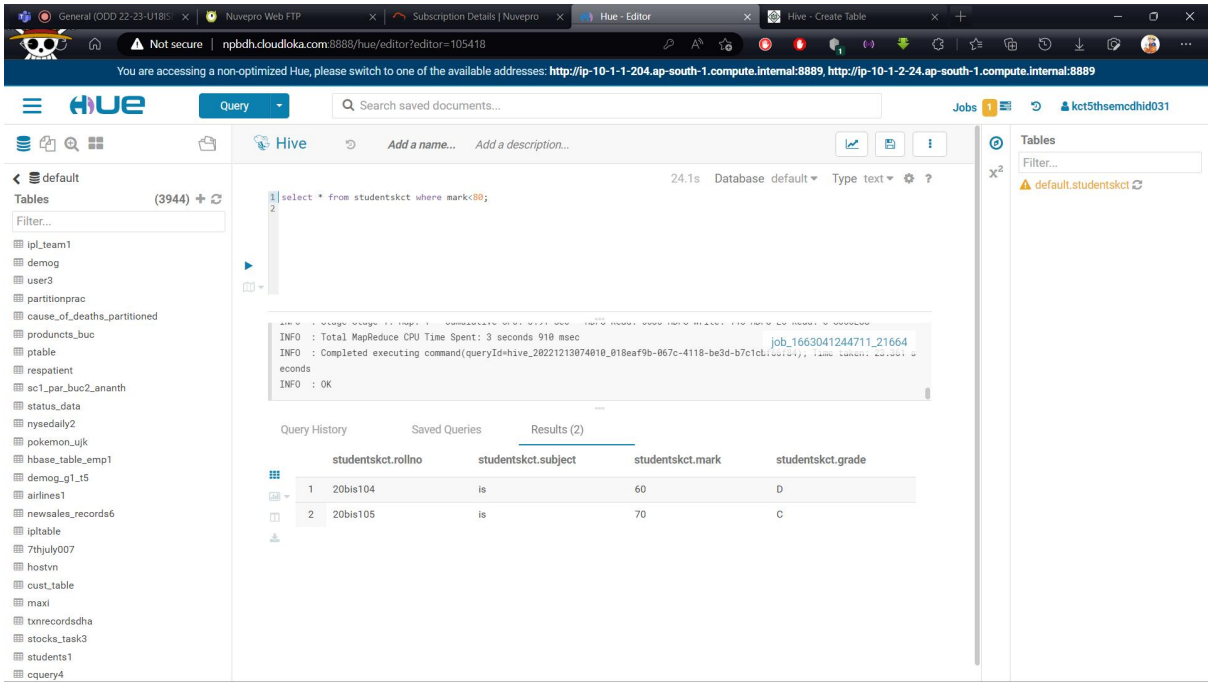
STEP4:Insert values in the table
insert into studentskct values('20bis100','cs',90,'A'),

('20bis101','cs',90,'A'), ('20bis102','i5',80,'B'), ('20bis103','cs',80,'B'), ('20bis104','is',60,'D'), ('20bis105','is',70,'C');

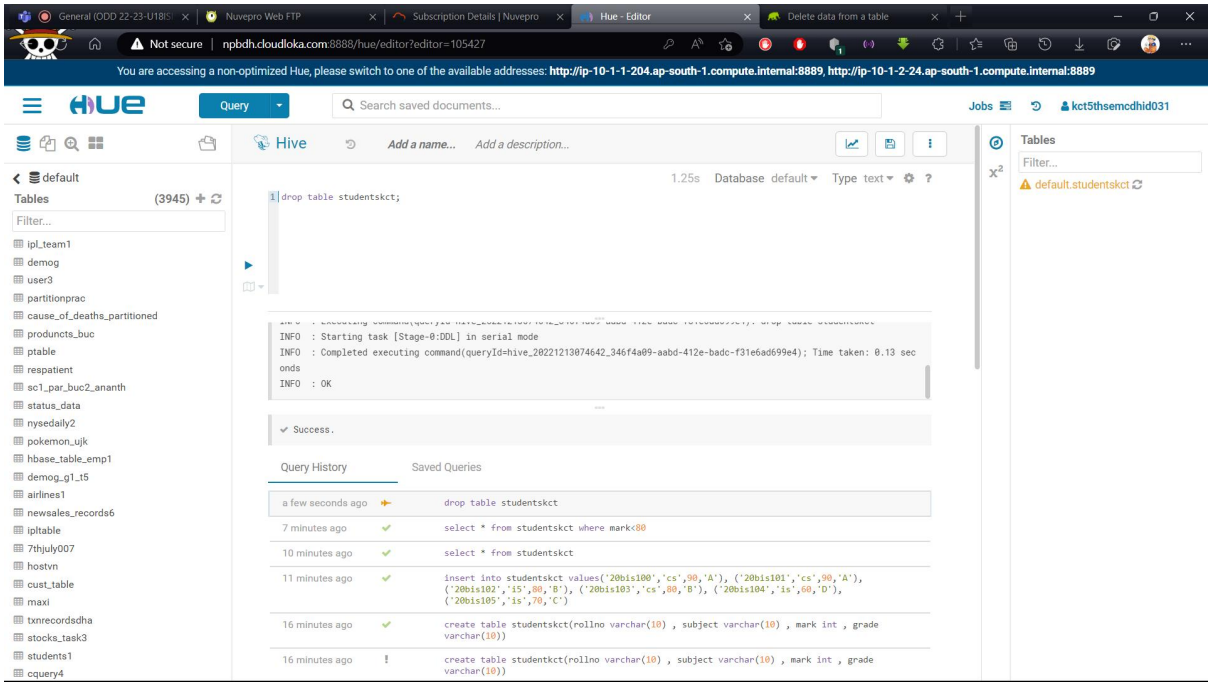
select * from studentskct;



STEP5:Filtering records with condition select * from studentskct where mark<80;



STEP6:Drop table
drop table studentskct;


Not secure | npbdh.cloudloka.com:8888/hue/editor?editor=105427
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-1-1-204.ap-south-1.compute.internal:8889, http://ip-10-1-2-24.ap-south-1.compute.internal:8889
Query
Search saved documents...
Jobs
kct5thsemcdhid031
Hive
Add a name...
Add a description...
default
Tables (3945)
Filter...
ipl_team1
demog
user3
partitionprac
cause_of_deaths_partitioned
producncts_buc
ptable
respatient
sc1_par_buc2_ananth
status_data
nysedaily2
pokemon_ujk
hbase_table_emp1
demog_g1_t5
airlines1
newsales_records6
ipltable
7thjuly007
hostvn
cust_table
maxi
txnrecordsdha
stocks_task3
students1
cquery4
1.25s
Database default
Type text
1 drop table studentskct;
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20221213074642_346f4a09-aabd-412e-badc-f31e6ad699e4); Time taken: 0.13 seconds
INFO : OK
Success.
Query History
Saved Queries
a few seconds ago drop table studentskct
7 minutes ago select * from studentskct where mark<80
10 minutes ago select * from studentskct
11 minutes ago insert into studentskct values('20bis100','cs',90,'A'), ('20bis101','cs',90,'A'), ('20bis102','is',80,'B'), ('20bis103','cs',80,'B'), ('20bis104','is',60,'D'), ('20bis105','is',70,'C')
16 minutes ago create table studentskct(rollno varchar(10) , subject varchar(10) , mark int , grade varchar(10))
16 minutes ago create table studentkct(rollno varchar(10) , subject varchar(10) , mark int , grade varchar(10))
Tables
Filter...
default.studentskct

U18ISI5202-BIG DATA LABORATORY

NAME: CHAARVIKA.A

ROLL NUMBER: 20BIS012

S.NO: 10

TOPIC : Verify, Sparse and perform advance join of data using spark

AIM:

To Verify, Sparse and perform advance join of data using spark

THEORY:

Join operations in Apache Spark is often a biggest source of performance problems and even full-blown exceptions in Spark. After this talk, you will understand the two most basic methods Spark employs for joining dataframes – to the level of detail of how Spark distributes the data within the cluster. You'll also find out how to work out common errors and even handle the trickiest corner cases we've encountered! After this talk, you should be able to write performance joins in Spark SQL that scale and are zippy fast!

This session will cover different ways of joining tables in Apache Spark. ShuffleHashJoin

–A ShuffleHashJoin is the most basic way to join tables in Spark – we'll diagram how Spark shuffles the dataset to make this happen.

BroadcastHashJoin

–A BroadcastHashJoin is also a very common way for Spark to join two tables under the special condition that one of your tables is small.

Dealing with Key Skew in a ShuffleHashJoin

–Key Skew is a common source of slowness for a Shuffle Hash Join – we'll describe what this is and how you might work around this.

CartesianJoin

–Cartesian Joins is a hard problem – we'll describe why it's difficult as well as what you need to do to make that work and what to look out for.

One to Many Joins

–When a single row in one table can match to many rows in your other table, the total number of output rows in your joined table can be really high. We'll let you know how to deal with this.

Theta Joins

–If you aren't joining two tables strictly by key, but instead checking on a condition for your tables, you may need to provide some hints to Spark SQL to get this to run well.

1. Broadcast Hash Join

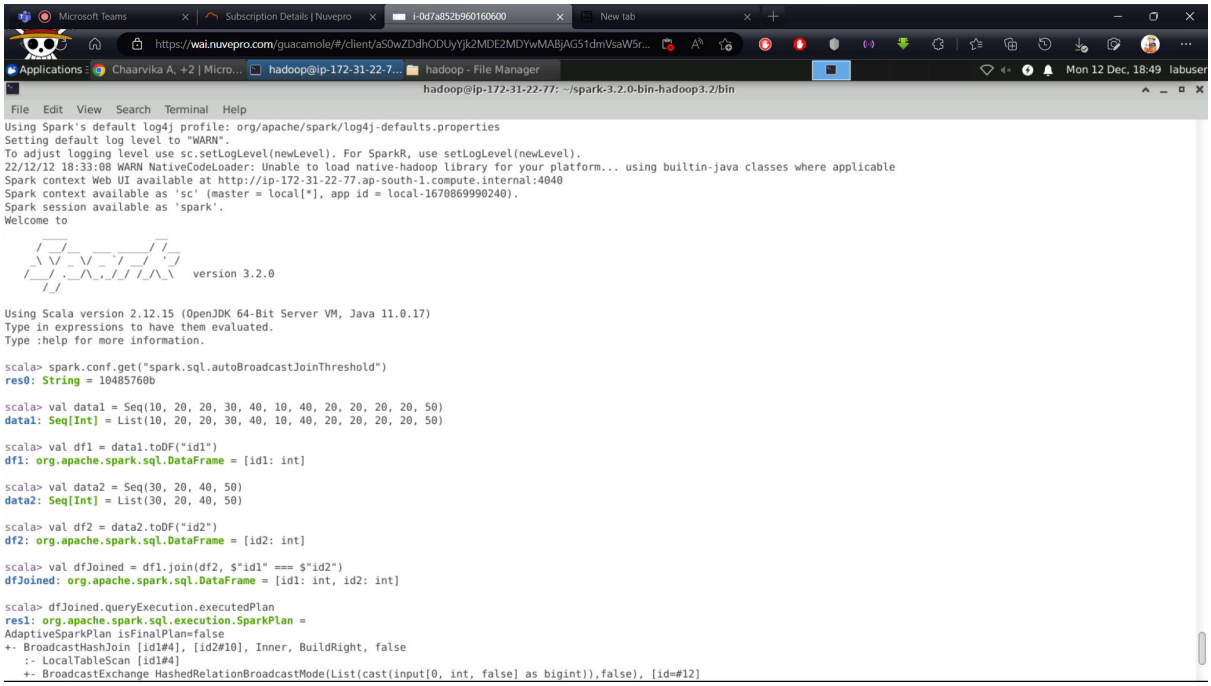
COMMANDS

scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold") res1: String = 10485760
scala> val data1 = Seq(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)
data1: Seq[Int] = List(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)
scala> val df1 = data1.toDF("id1")

df1: org.apache.spark.sql.DataFrame = [id1: int] scala> val data2 = Seq(30, 20, 40, 50) data2: Seq[Int] = List(30, 20, 40, 50) scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]

val dfJoined = df1.join(df2, $"id1" === $"id2")

scala> dfJoined.queryExecution.executedPlan scala> dfJoined.show



2. Shuffle Hash Join COMMANDS

scala> spark.conf.set("spark.sql.autoBroadcastJoinThreshold", 2) scala> spark.conf.set("spark.sql.join.preferSortMergeJoin", "false")

scala> spark.conf.get("spark.sql.join.preferSortMergeJoin") res2: String = false scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold") res3: String = 2

scala> val data1 = Seq(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50) data1: Seq[Int] = List(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)

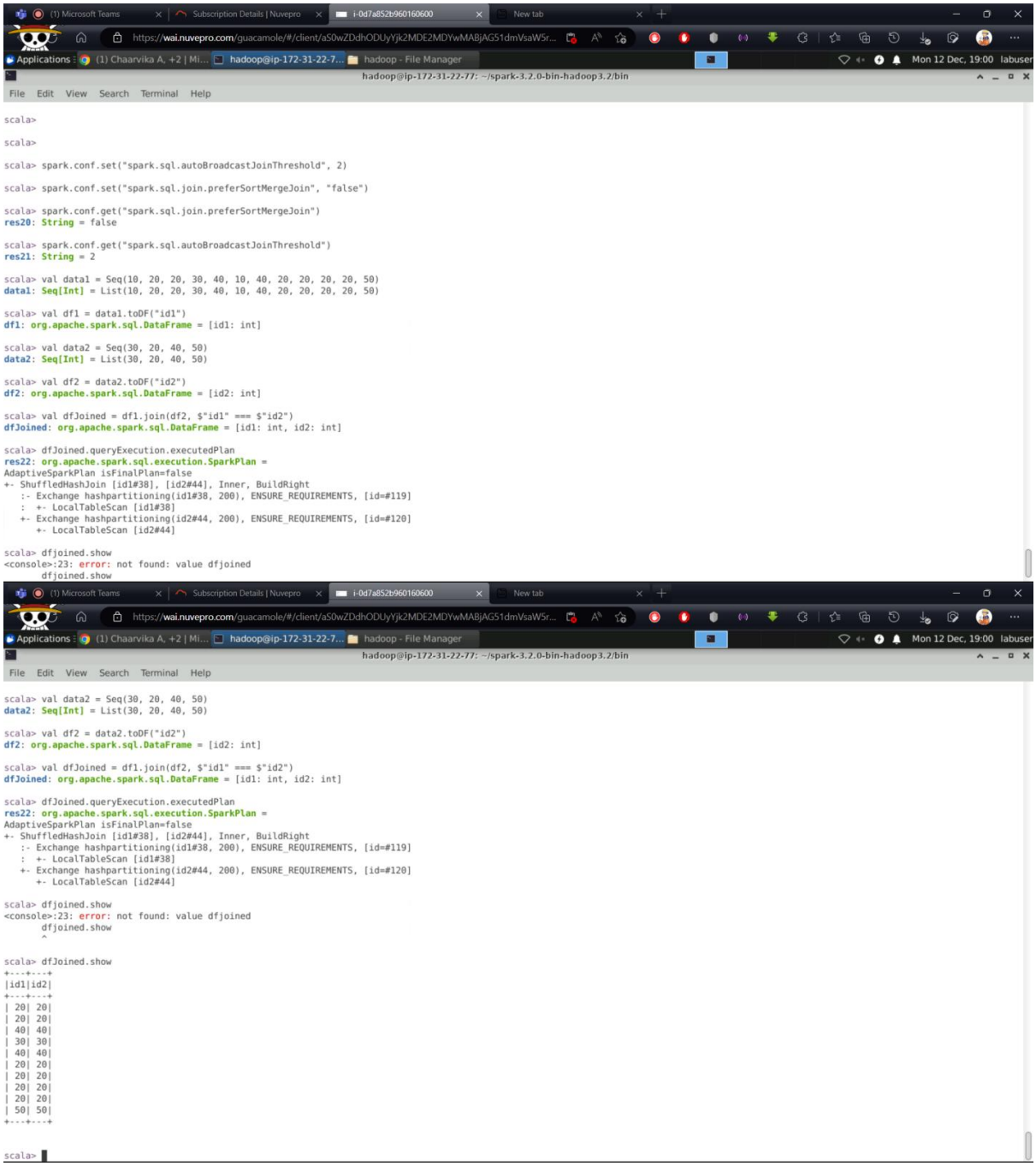
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int]

scala> val data2 = Seq(30, 20, 40, 50) data2: Seq[Int] = List(30, 20, 40, 50)

scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]

scala> val dfJoined = df1.join(df2, $"id1" === $"id2") dfJoined: org.apache.spark.sql.DataFrame = [id1: int, id2: int]

scala> dfJoined.queryExecution.executedPlan scala> dfJoined.show



3. CARTESIAN PRODUCT JOIN
COMMANDS
scala> spark.conf.get("spark.sql.join.preferSortMergeJoin") res1: String = true
scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold") res2: String = -1
scala> val data1 = Seq(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)
data1: Seq[Int] = List(10, 20, 20, 30, 40, 10, 40, 20, 20, 20, 20, 50)
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int] 5.scala> val data2 = Seq(30, 20, 40, 50)
data2: Seq[Int] = List(30, 20, 40, 50) 6.scala> val df2 = data2.toDF("id2")
df2: org.apache.spark.sql.DataFrame = [id2: int]

scala> val dfJoined = df1.join(df2, $"id1" >= $"id2")
dfJoined: org.apache.spark.sql.DataFrame = [id1: int, id2: int]

scala> dfJoined.queryExecution.executedPlan scala> dfJoined.show