

BIG DATA TECHNOLOGIES LABORATORY

NAME: CHAARVIKA.A

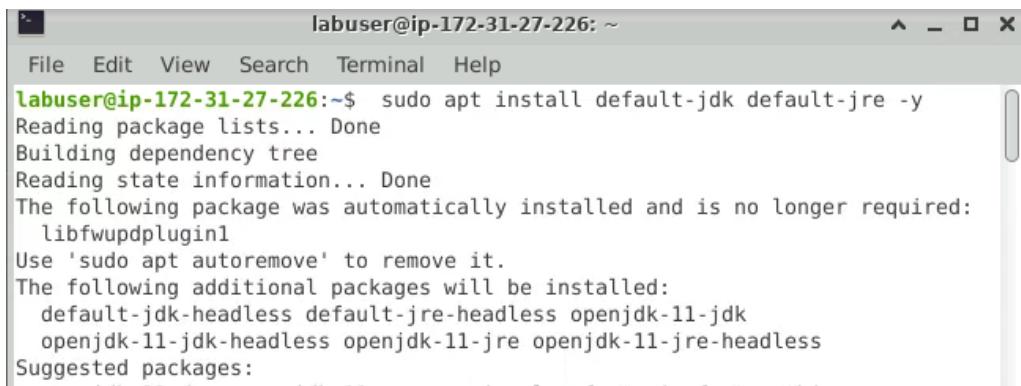
ROLL NUMBER: 20BIS012

LAB EX-1: Hadoop Installation

LAB EXERCISE-1

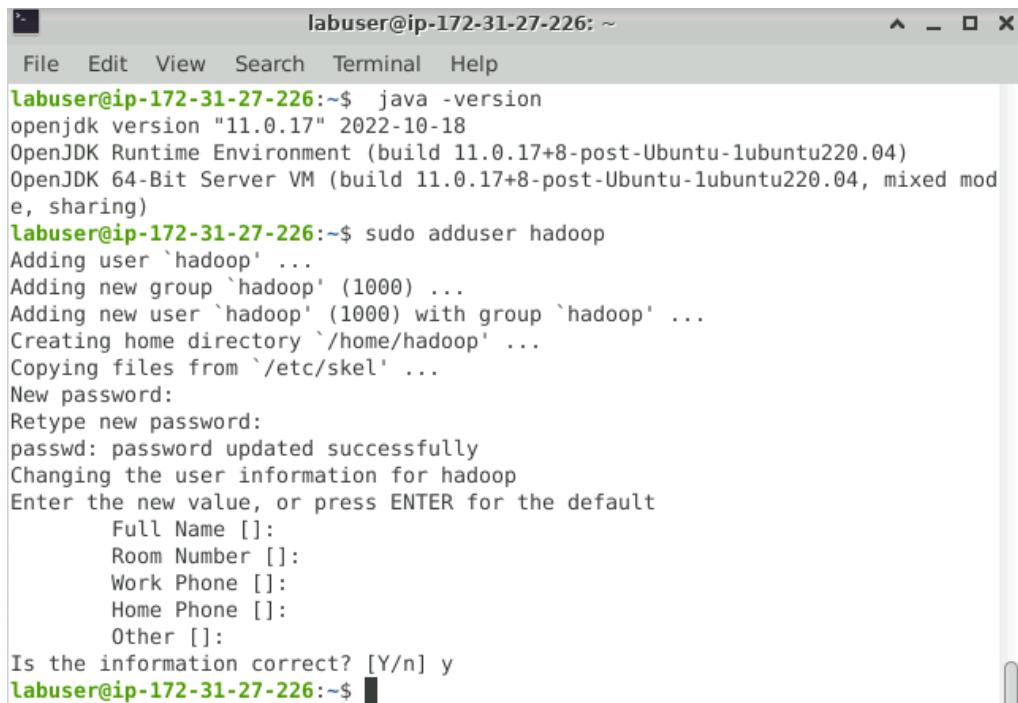
HADOOP INSTALLATION

STEP -1



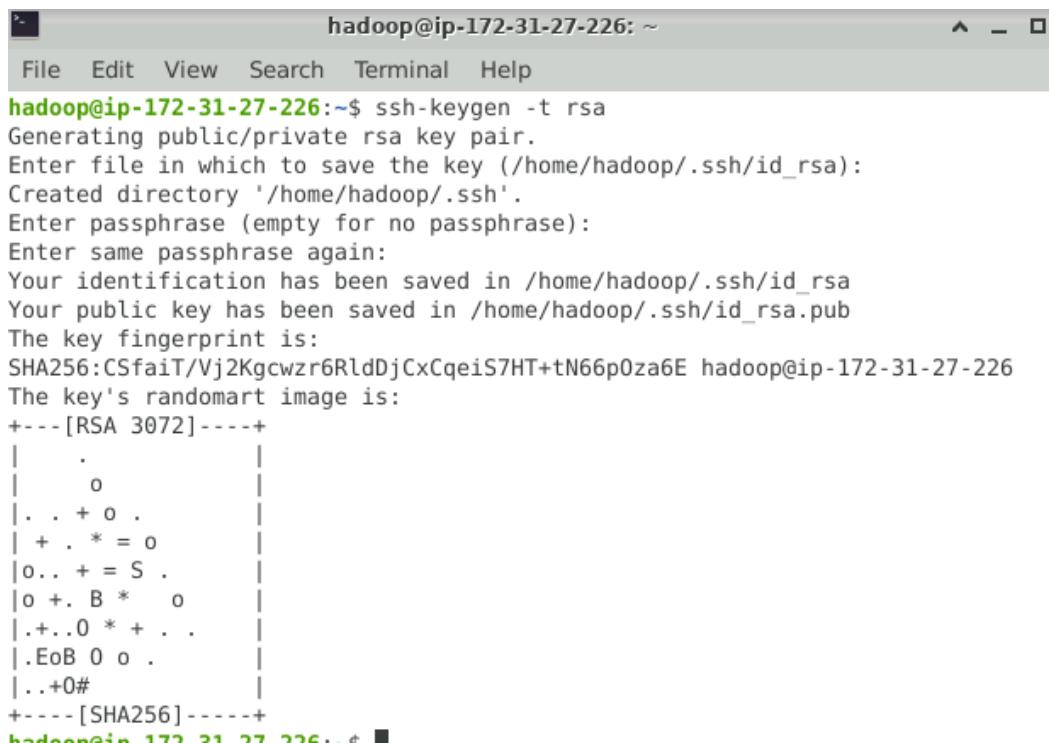
```
labuser@ip-172-31-27-226: ~
File Edit View Search Terminal Help
labuser@ip-172-31-27-226:~$ sudo apt install default-jdk default-jre -y
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following package was automatically installed and is no longer required:
  libfwupdplugin1
Use 'sudo apt autoremove' to remove it.
The following additional packages will be installed:
  default-jdk-headless default-jre-headless openjdk-11-jdk
  openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless
Suggested packages:
```

STEP - 2



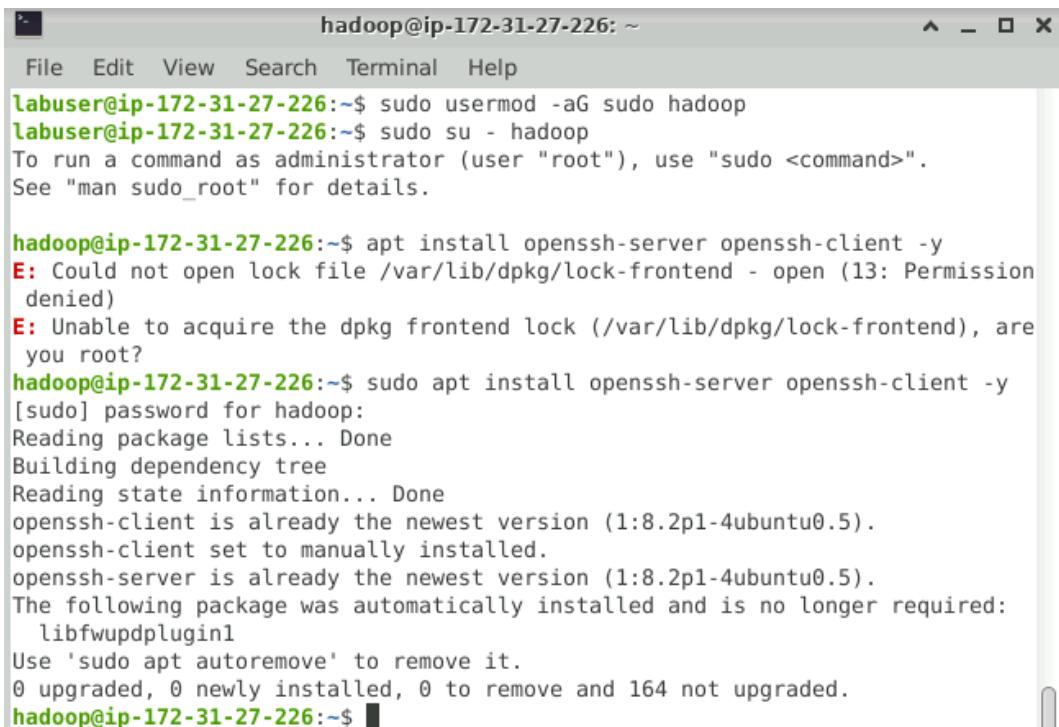
```
labuser@ip-172-31-27-226:~$ java -version
openjdk version "11.0.17" 2022-10-18
OpenJDK Runtime Environment (build 11.0.17+8-post-Ubuntu-1ubuntu220.04)
OpenJDK 64-Bit Server VM (build 11.0.17+8-post-Ubuntu-1ubuntu220.04, mixed mode, sharing)
labuser@ip-172-31-27-226:~$ sudo adduser hadoop
Adding user `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:~$ 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+tN66p0za6E hadoop@ip-172-31-27-226
The key's randomart image is:
+---[RSA 3072]----+
| . |
|  o |
| .. + o . |
| + . * = o |
|o.. + = S . |
|o +. B *   o |
|.+.0 * + . . |
|.EoB 0 o . |
|..+0#|
+---[SHA256]-----+
hadoop@ip-172-31-27-226:~$
```

STEP - 4



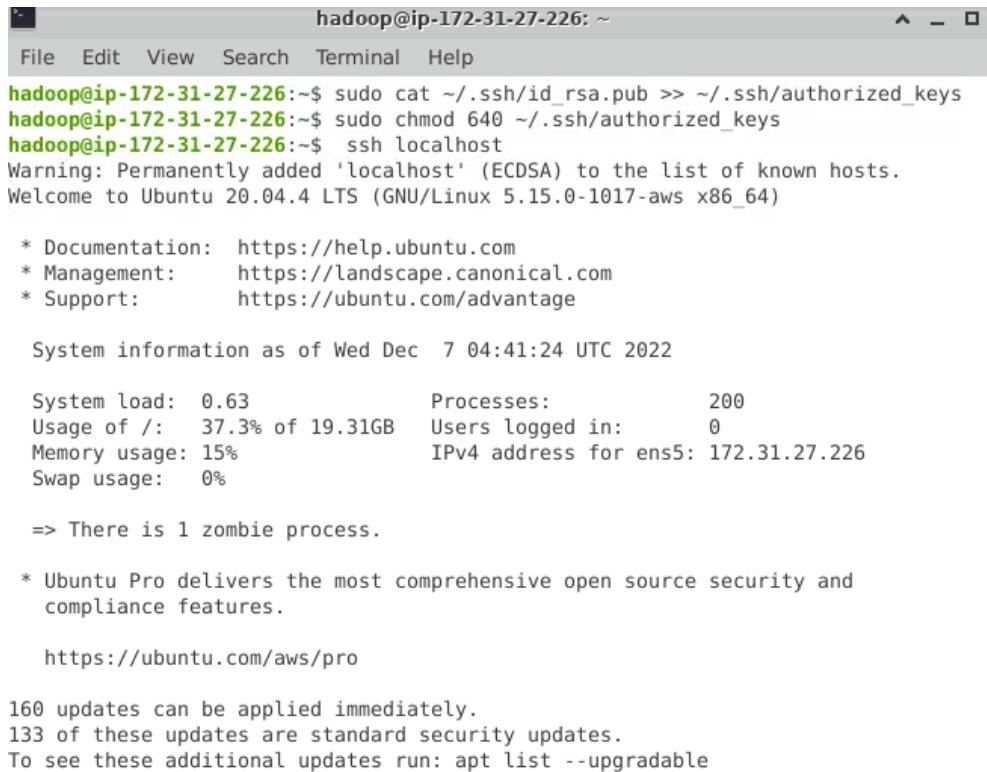
```

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

```

hadoop@ip-172-31-27-226:~$ wget https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
--2022-12-07 04:42:20-- https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.95.219, 2a01:4f9:3a:2c57::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected

HTTP request sent, awaiting response... 200 OK
Length: 605187279 (577M) [application/x-gzip]
Saving to: 'hadoop-3.3.1.tar.gz'

hadoop-3.3.1.tar.gz      19%[=====]   5M 19.9MB/s eta 31s
hadoop-3.3.1.tar.gz      19%[=====]   114.51M 20.1MB/s eta 31s
hadoop-3.3.1.tar.gz      20%[=====]   115.98M 19.7MB/s eta 31s
hadoop-3.3.1.tar.gz      20%[=====]   117.46M 19.9MB/s eta

```

STEP - 7

```
hadoop@ip-172-31-27-226:~$ tar -xvzf hadoop-3.3.1.tar.gz
```

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

```

```
Dashboard | Nuvepro | Subscription Details | Nuvepro | i-0b33a58552a2ff9af | Install and Configure Apache Hadoop | +  
http://www.nuvepro.com/guacamole/#/client/a50wYjMzYTU4NTU7YTmZjh2gjGAS51mdVsW5r?hostname=3.110.111.151&protocol=rdp&port=3389&username=n...  
Google Gmail Mail T Y My Camu Canva Vidyard (59) ICICI SBI O M i S IS Miro yoga  
Applications DataNode Information ... hadoop@ip-172-31-19-1... hadoop@ip-172-31-19-153: ~  
File Edit View Search Terminal Help  
hadoop-3.3.1/include/TemplateFactory.hh  
hadoop-3.3.1/include/StringUtils.hh  
hadoop-3.3.1/include/hdfs.h  
hadoop-3.3.1/include/Pipes.hh  
hadoop@ip-172-31-19-153:~$ sudo mv hadoop-3.3.1 /usr/local/hadoop  
hadoop@ip-172-31-19-153:~$ sudo mkdir /usr/local/hadoop/logs  
hadoop@ip-172-31-19-153:~$ sudo chown -R hadoop:hadoop /usr/local/hadoop  
hadoop@ip-172-31-19-153:~$ sudo nano ~/.bashrc  
hadoop@ip-172-31-19-153:~$ source ~/.bashrc  
hadoop@ip-172-31-19-153:~$ source /home/hadoop/.bashrc  
hadoop@ip-172-31-19-153:~$ source /etc/profile  
hadoop@ip-172-31-19-153:~$ which javac  
/usr/bin/javac  
hadoop@ip-172-31-19-153:~$ readlink -f /usr/bin/javac  
/usr/lib/jvm/java-11-openjdk-amd64/bin/javac  
hadoop@ip-172-31-19-153:~$ sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh  
hadoop@ip-172-31-19-153:~$ cd /usr/local/hadoop/lib  
hadoop@ip-172-31-19-153:~/user/local/hadoop$ sudo wget https://jcenter.bintray.com/javax/activation/javax.activation-api/1.2.0/javax.activation-api-1.2.0.jar  
--2022-12-07 04:32:39 - https://jcenter.bintray.com/javax/activation/javax.activation-api/1.2.0/javax.activation-api-1.2.0.jar  
Resolving jcenter.bintray.com [jcenter.bintray.com]... 34.95.74.180  
Connecting to jcenter.bintray.com [jcenter.bintray.com]|34.95.74.180|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 56674 (55K) [application/java-archive]  
Saving to: 'javax.activation-api-1.2.0.jar'  
  
javax.activation-ap 100%[=====] 55.35K --.KB/s in 0.002s  
  
2022-12-07 04:32:39 (26.5 MB/s) - 'javax.activation-api-1.2.0.jar' saved [56674/56674]  
  
hadoop@ip-172-31-19-153:~/user/local/hadoop/lib$ hadoop version  
Hadoop 3.3.1  
Source code repository https://github.com/apache/hadoop.git - r a3b9c37a397ad4188041dd80621bdeefc46885f2
```

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

The screenshot shows the Mozilla Firefox interface with the address bar set to "localhost:8088/cluster". The main content area displays the Hadoop cluster management interface. On the left, there's a sidebar with links for Cluster (About, Nodes, Node Labels, Applications), Applications (NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), and Scheduler. The main panel shows "Cluster Metrics" with counts for Apps Submitted (0), Apps Pending (0), Apps Running (0), Apps Completed (0), and Containers Running (0). It also shows "Cluster Nodes Metrics" with Active Nodes (1), Decommissioning Nodes (0), and Decommissioned Nodes (0). Under "Scheduler Metrics", it shows Scheduler Type (Capacity Scheduler), Scheduling Resource Type ([memory-mb (unit=Mi), vcores]), and Minimum Allocation (<memory:1024, vCores:1>). A table titled "Show 20 entries" lists Scheduler entries with columns: ID, User, Name, Application Type, Application Tags, Queue, Application Priority, StartTime, LaunchTime, FinishTime, and State.

localhost:9864

The screenshot shows the Mozilla Firefox interface with the address bar set to "localhost:9864/datanode.html". The main content area displays the DataNode Information page. At the top, it says "DataNode on ip-172-31-19-153.ap-south-1.compute.internal:9866". Below this, there's a table with two rows: "Cluster ID: CID-d3be57eb-1737-4c72-84b7-121aa2f9a155" and "Version: 3.3.1, ra3b9c37a397ad4188041dd80621bdefc46885f2". Underneath, there's a section titled "Block Pools" with a table header: "Namemode Address", "Block Pool ID", "Actor State", "Last Heartbeat", "Last Block Report", and "Last Block Report Size (Max Size)".

localhost:9870

The screenshot shows a Mozilla Firefox window with the title "Namenode information" and the URL "localhost:9870/dfshealth.html#tab-overview". The browser interface includes tabs for "All Applications", "Namenode information", and "DataNode Information". A message bar at the top says, "It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!" with a "Refresh Firefox..." button.

The main content area is titled "Overview '0.0.0.0:9000' (active)". It contains a table with the following data:

Started:	Wed Dec 07 04:36:23 +0000 2022
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled:	Tue Jun 15 05:13:00 +0000 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-d3be57eb-1737-4c72-84b7-121aa2f9a155
Block Pool ID:	BP-568123600-172.31.19.153-1670387768213

Below the table, there is a link labeled "Summary".

LAB EXERCISE-2

STEP -1

```
$ hdfs dfs -mkdir /user/kct5thsemcidh036/mridula
npbdh login: kct5thsemcidh036
kct5thsemcidh036@npbdh.cloudloka.com's password:
Last login: Tue Dec 13 05:18:32 2022 from ec2-65-1-45-35.ap-south-1
[kct5thsemcidh036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/kct5thse
mcidh036/mridula
```

STEP -2

```
$ hdfs dfs -ls /
[kct5thsemcidh036@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
[kct5thsemcidh036@ip-10-1-1-204 ~]$
```

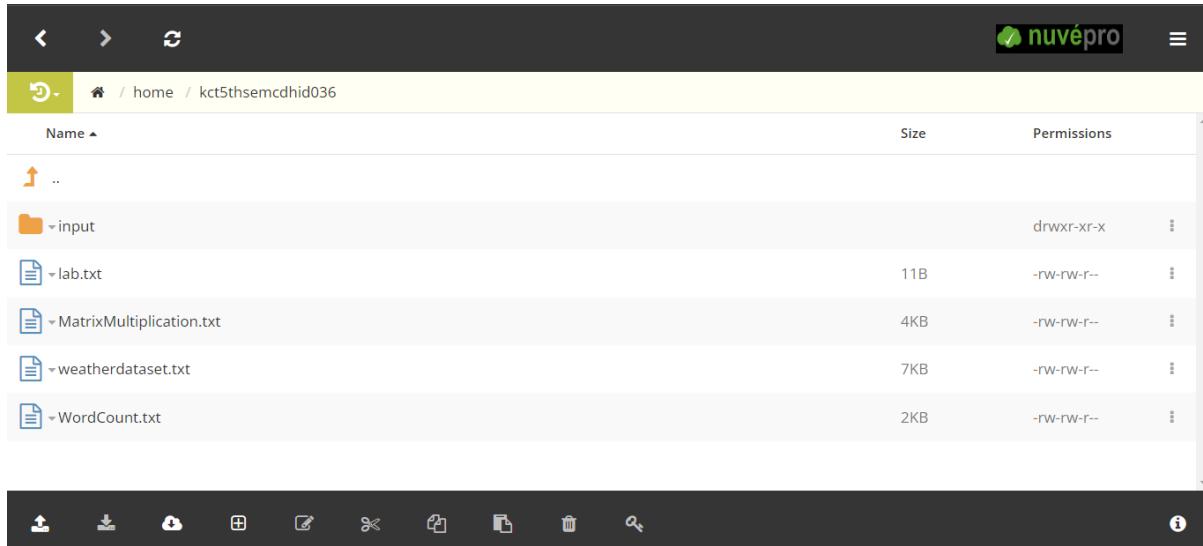
STEP -3

```
$ hdfs dfs -put '/home/kct5thsemcidh036/lab.txt'
/user/kct5thsemcidh036/mridula
[kct5thsemcidh036@ip-10-1-1-204 ~]$ hdfs dfs -put '/home/kct5thsemcidh036/lab.txt' /user/kct5thsemcidh036/mridula
[kct5thsemcidh036@ip-10-1-1-204 ~]$
```

STEP -4

```
$ hadoop fs -get /user/kct5thsemcidh036/mridula/lab.txt
/home/kct5thsemcidh036/
$ hdfs dfs -mkdir -p /user/kct5thsemcidh036/mridula/input/
$ hadoop fs -get /user/kct5thsemcidh036/mridula/input
/home/kct5thsemcidh036/
```

```
[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
/usr/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
```

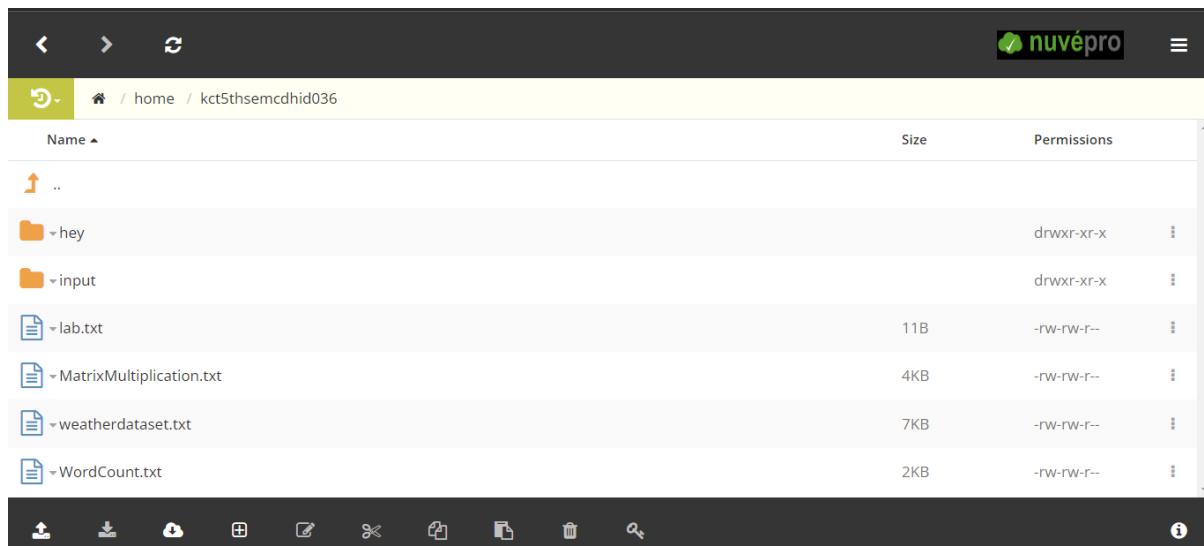
STEP -9

```
$ 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
```

```
[kct5thsemcidhid036@ip-10-1-1-204 ~]$ hdfs dfs -mkdir /user/kct5thsemcidhid036/hello  
[kct5thsemcidhid036@ip-10-1-1-204 ~]$ hdfs dfs -rmdir /user/kct5thsemcidhid036/hello  
[kct5thsemcidhid036@ip-10-1-1-204 ~]$ █
```

STEP -12

```
$ hadoop fs -tail /user/kct5thsemcidhid036/mridula/lab.txt/
```

```
[kct5thsemcidhid036@ip-10-1-1-204 ~]$ hadoop fs -tail /user/kct5thsemcidhid036/mridula/lab.txt/  
hi ...hello[kct5thsemcidhid036@ip-10-1-1-204 ~]$ █
```

STEP -13

```
$ hadoop fs -du /user/kct5thsemcidhid036/mridula/lab.txt/
```

```
[kct5thsemcidhid036@ip-10-1-1-204 ~]$ hadoop fs -du /user/kct5thsemcidhid036/mridula/lab.txt/  
11 33 /user/kct5thsemcidhid036/mridula/lab.txt  
[kct5thsemcidhid036@ip-10-1-1-204 ~]$ █
```

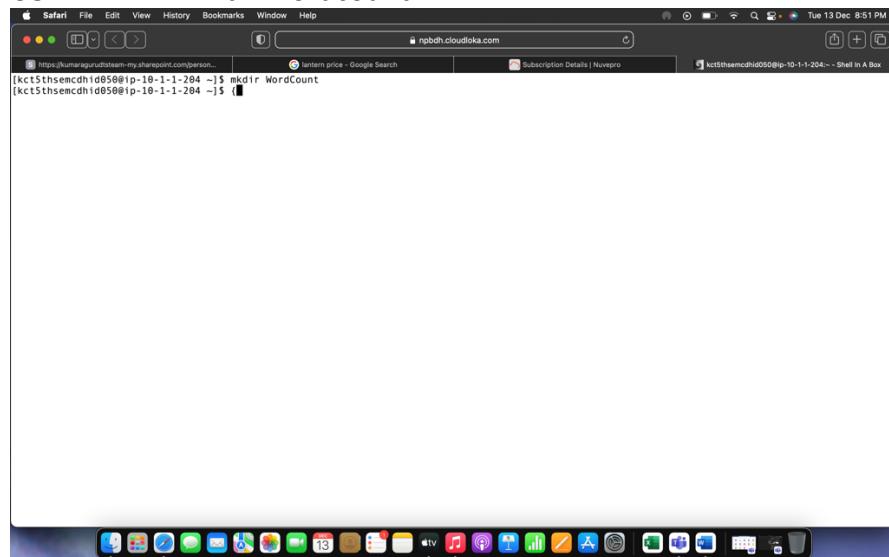

U18ISI5202-BIG DATA LABORATORY

**NAME: CHAARVIKA.A
ROLL NUMBER: 20BIS012**

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

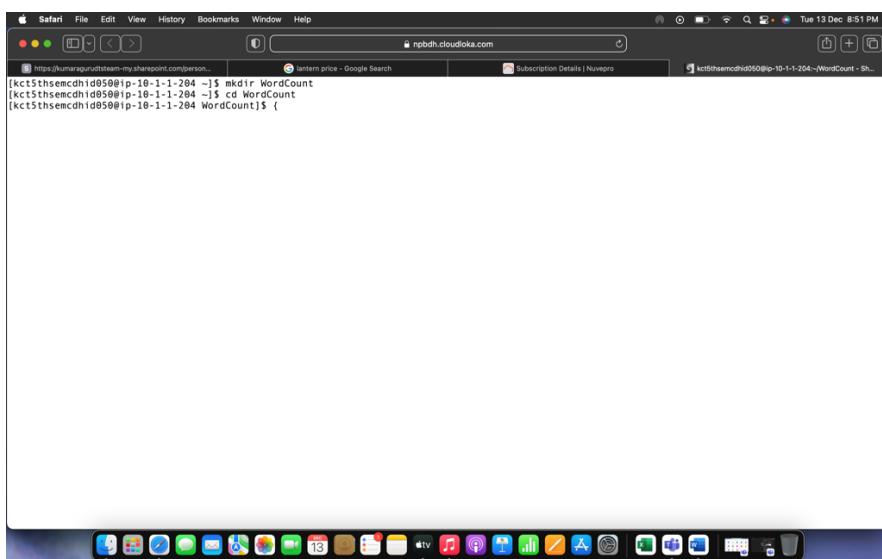
1. CREATE A FOLDER/ DIRECTORY FOR WORDCOUNT IN HOME

COMMAND: mkdir wordcount



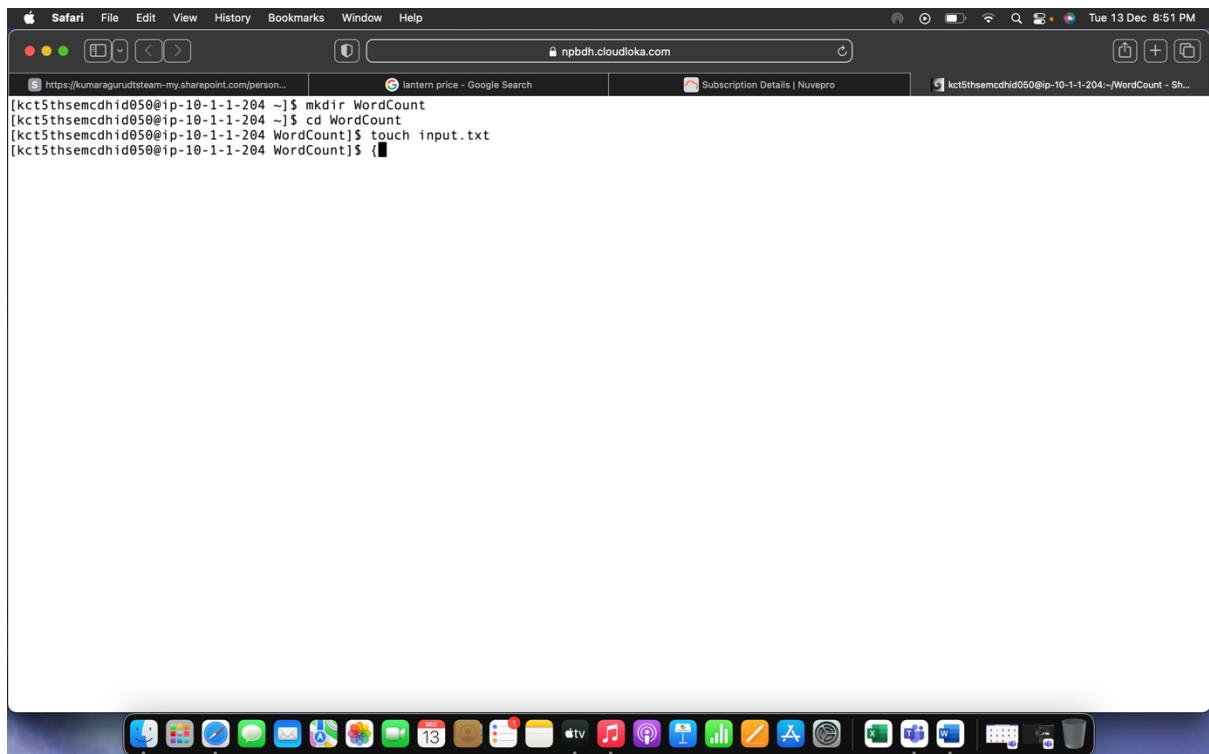
```
[kct5thsemcdhid050@ip-10-1-1-284 ~]$ mkdir WordCount
```

2. MOVE TO THAT FOLDER/DIRECTORY:

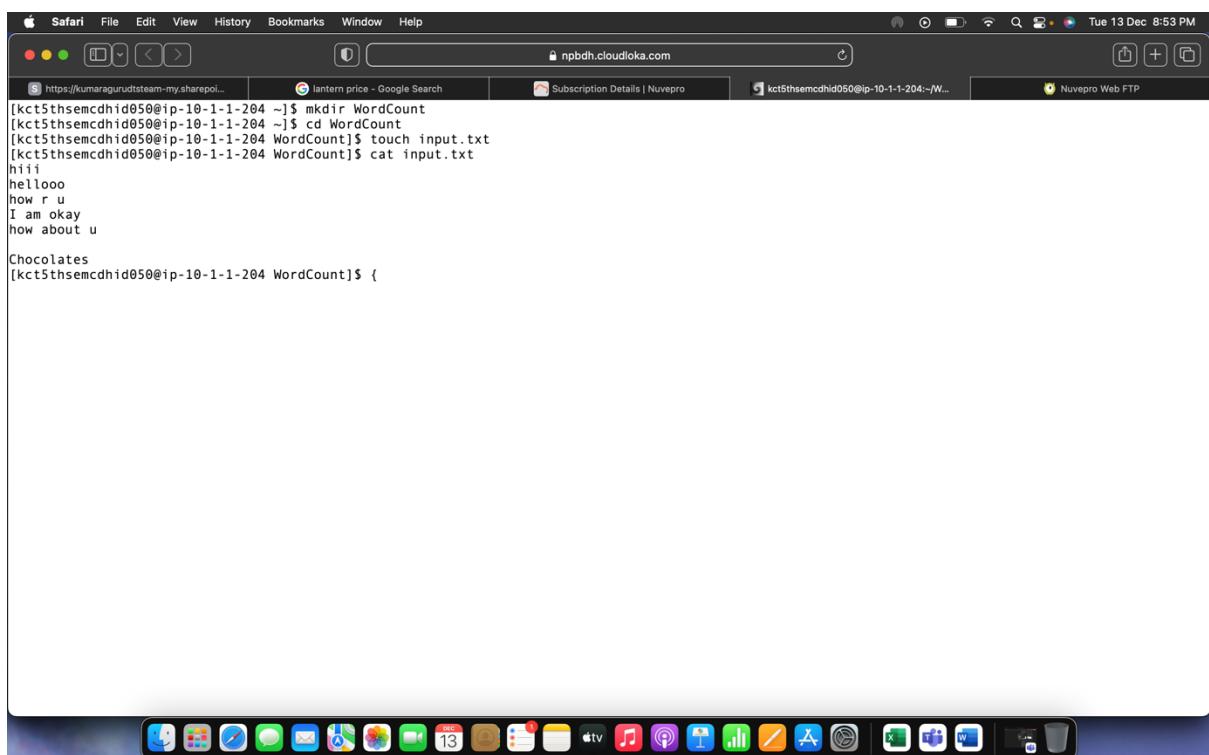


```
[kct5thsemcdhid050@ip-10-1-1-284 WordCount]$ cd WordCount
```

3. CREATE AN INPUT FILE INSIDE THE WORDCOUNT FOLDER:



```
[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]$
```



```
[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]$ {
```

4. CREATE A WORDCOUNT JAVA FILE:

```
1 import java.io.IOException;
2 import java.io.IntWritable;
3 import org.apache.hadoop.io.IntWritable;
4 import org.apache.hadoop.io.LongWritable;
5 import org.apache.hadoop.io.Text;
6 import org.apache.hadoop.mapreduce.Mapper;
7 import org.apache.hadoop.mapreduce.Reducer;
8 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
9 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
10 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
11 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
12 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
13 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
14 import org.apache.hadoop.fs.Path;
15 public class WordCount
16 {
17     public static class Map extends Mapper<LongWritable,Text,Text,IntWritable> {
18         public void map(LongWritable key, Text value,Context context) throws
19             IOException,InterruptedException{
20             StringTokenizer line = value.toString();
21             String[] tokens = line.toString().StringTokenizer(line);
22             while (tokens.hasMoreTokens()) {
23                 value.set(tokens.nextToken());
24                 context.write(value, new IntWritable(1));
25             }
26         }
27     }
28     public static class Reduce extends Reducer<Text,IntWritable,Text,IntWritable> {
29         public void reduce(Text key, Iterable<IntWritable> values,Context context)
30             throws IOException,InterruptedException {
31             int sum=0;
32             for(IntWritable x: values)
33             {
34                 sum+=x.get();
35             }
36             context.write(key, new IntWritable(sum));
37         }
38     }
39 }
40 //home/kct5thsemcdhd050@ip-10-1-204-7...WordCount.java
```

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 {

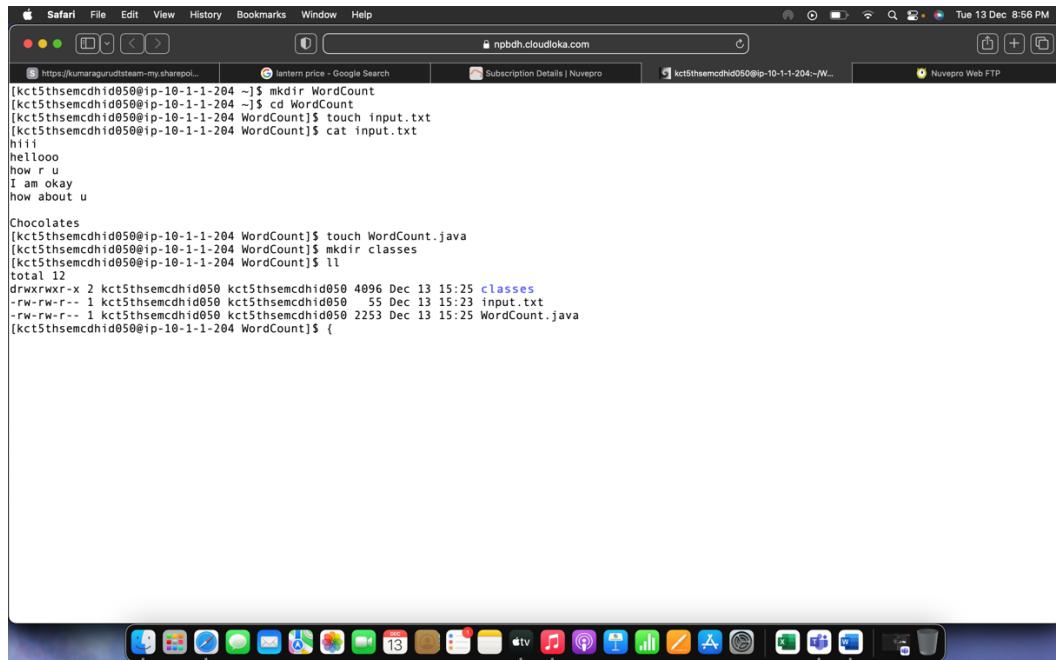
        Configuration conf= new Configuration();
        Job job = new Job(conf,"My Word Count Program");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        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:



```

Safari File Edit View History Bookmarks Window Help
https://kumaraguru007team-my.sharepoint.com/ 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
hi
helloo
how r u
I am okay
how about u

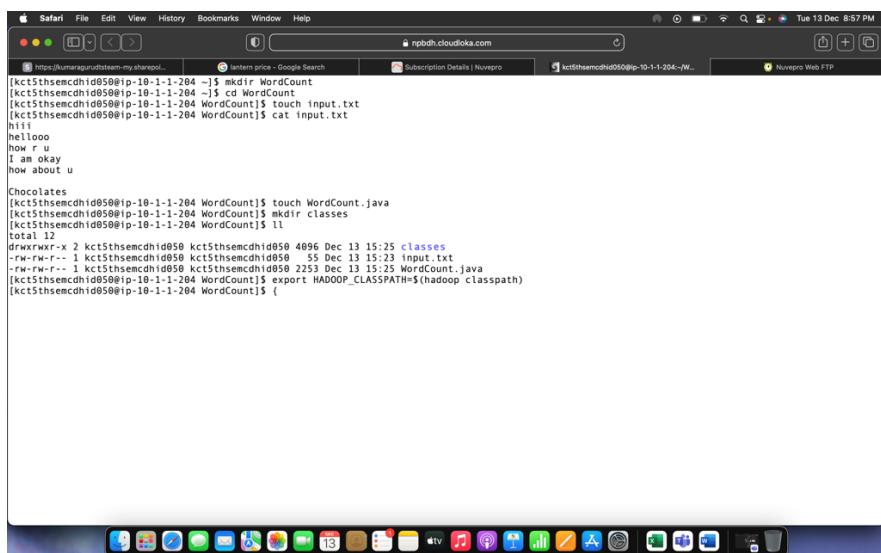
Chocolate
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 12
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:25 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 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]$ 

```

6. SET THE PATH FOR JAVA FILE

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

`echo $HADOOP_CLASSPATH`

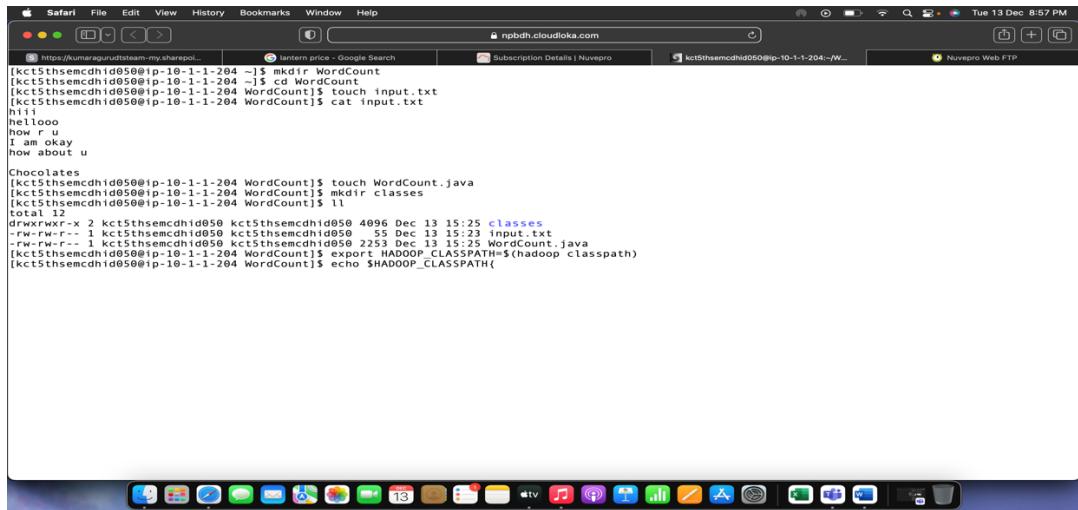


```

Safari File Edit View History Bookmarks Window Help
https://kumaraguru007team-my.sharepoint.com/ 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
hi
helloo
how r u
I am okay
how about u

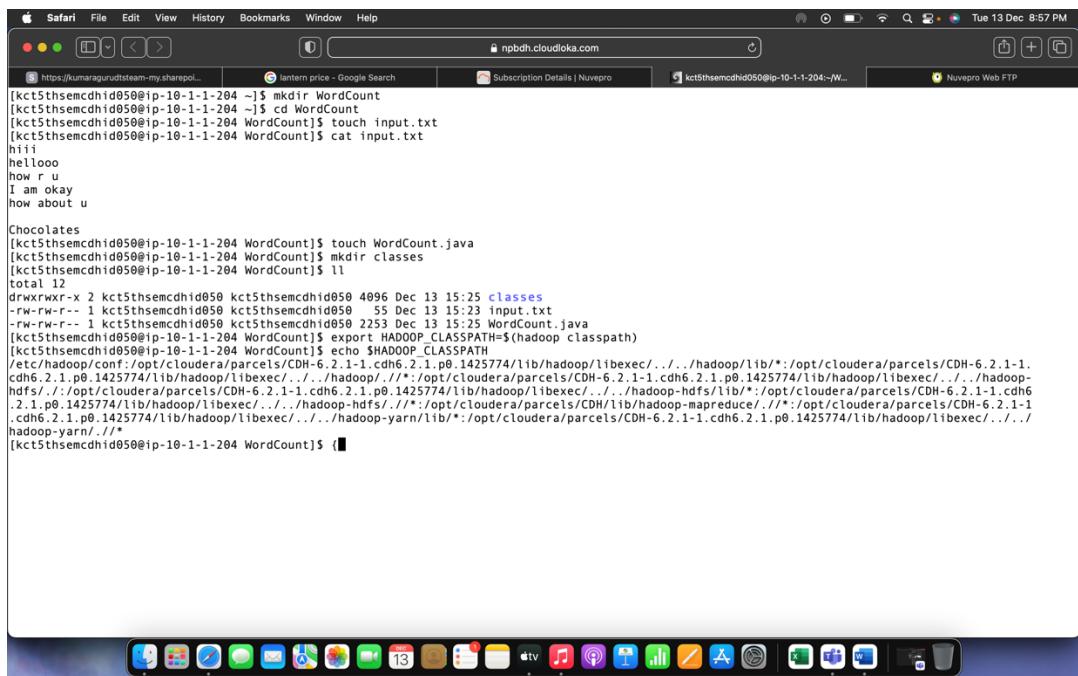
Chocolate
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ ll
total 12
drwxrwxr-x 2 kct5thsemcdhid050 kct5thsemcdhid050 4096 Dec 13 15:25 classes
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 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=$(hadoop classpath)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ 

```



```
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ mkdir classes
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ touch input.txt
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hiiii
hellooo
how r u
I am okay
now 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
```

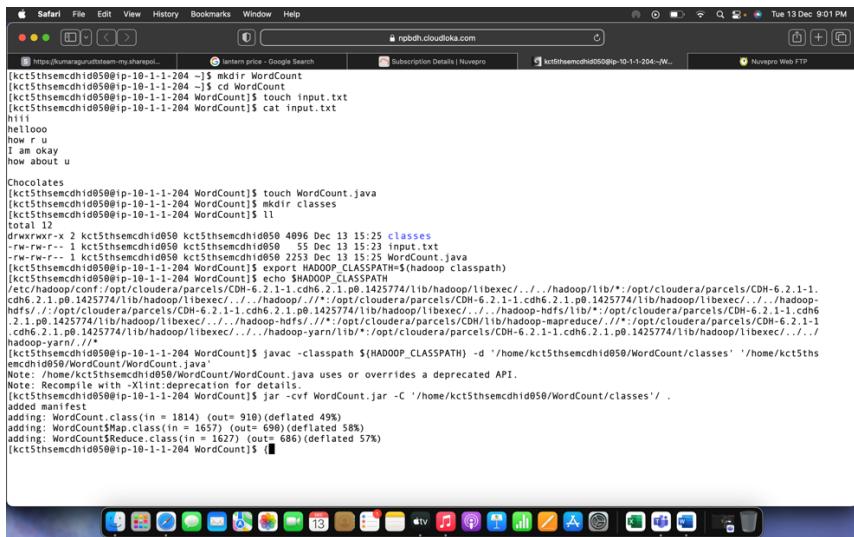


```
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ mkdir WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cd WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch input.txt
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hiiii
hellooo
how r u
I am okay
now 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-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-mapreduce/../../hadoop-yarn/lib/*:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop-yarn/../../hadoop
```

7. COMPILE THE JAVAFILE:

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

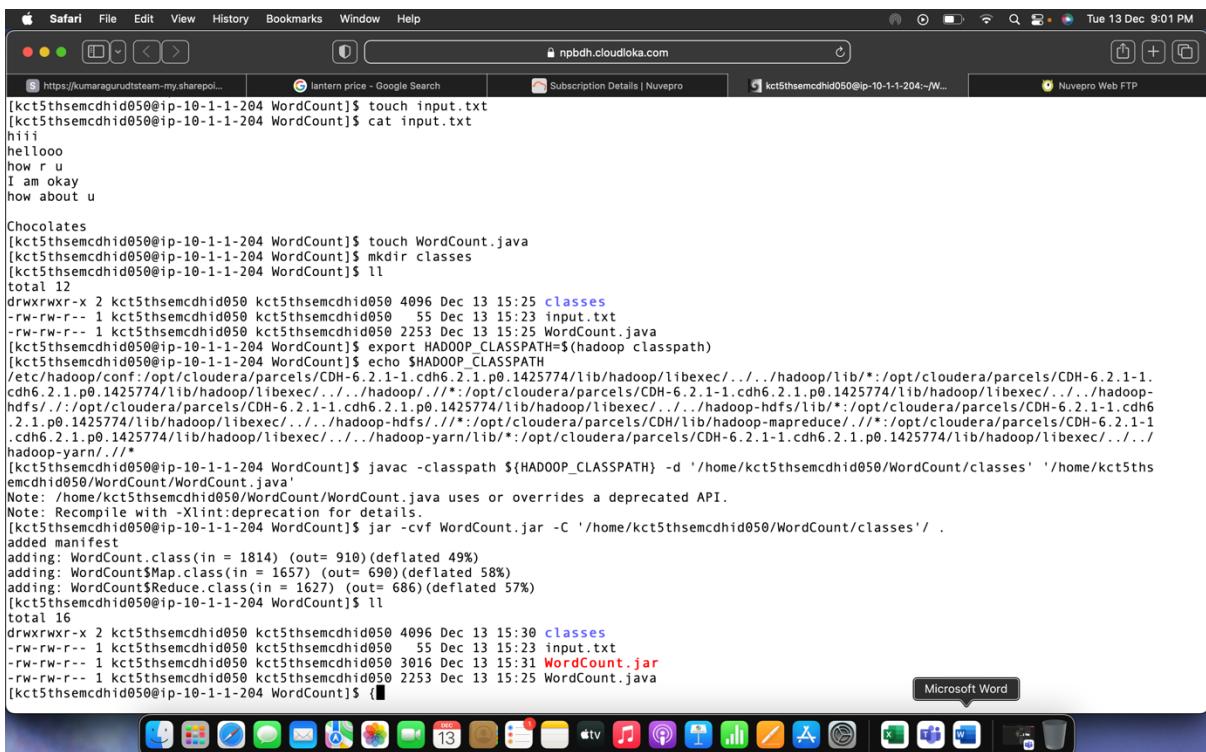


```
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ touch WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 ~]$ cd WordCount
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch input.txt
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hi
hellooo
how r u
I am okay
how about u

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 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=$(/usr/local/hadoop/bin/hadoop classpath)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ javac -classpath ${HADOOP_CLASSPATH} -d '/home/kct5thsemcdhid050/WordCount/classes' '/home/kct5thsemcdhid050/WordCount.java'
Note: /home/kct5thsemcdhid050/WordCount.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ jar -cvf WordCount.jar -C '/home/kct5thsemcdhid050/WordCount/classes' .
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]$
```

8. CREATE A JAR FILE:

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



```
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ touch input.txt
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ cat input.txt
hi
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=$(/usr/local/hadoop/bin/hadoop classpath)
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ javac -classpath ${HADOOP_CLASSPATH} -d '/home/kct5thsemcdhid050/WordCount/classes' '/home/kct5thsemcdhid050/WordCount.java'
Note: /home/kct5thsemcdhid050/WordCount.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ jar -cvf WordCount.jar -C '/home/kct5thsemcdhid050/WordCount/classes' .
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.jar
-rw-rw-r-- 1 kct5thsemcdhid050 kct5thsemcdhid050 2253 Dec 13 15:25 WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$
```

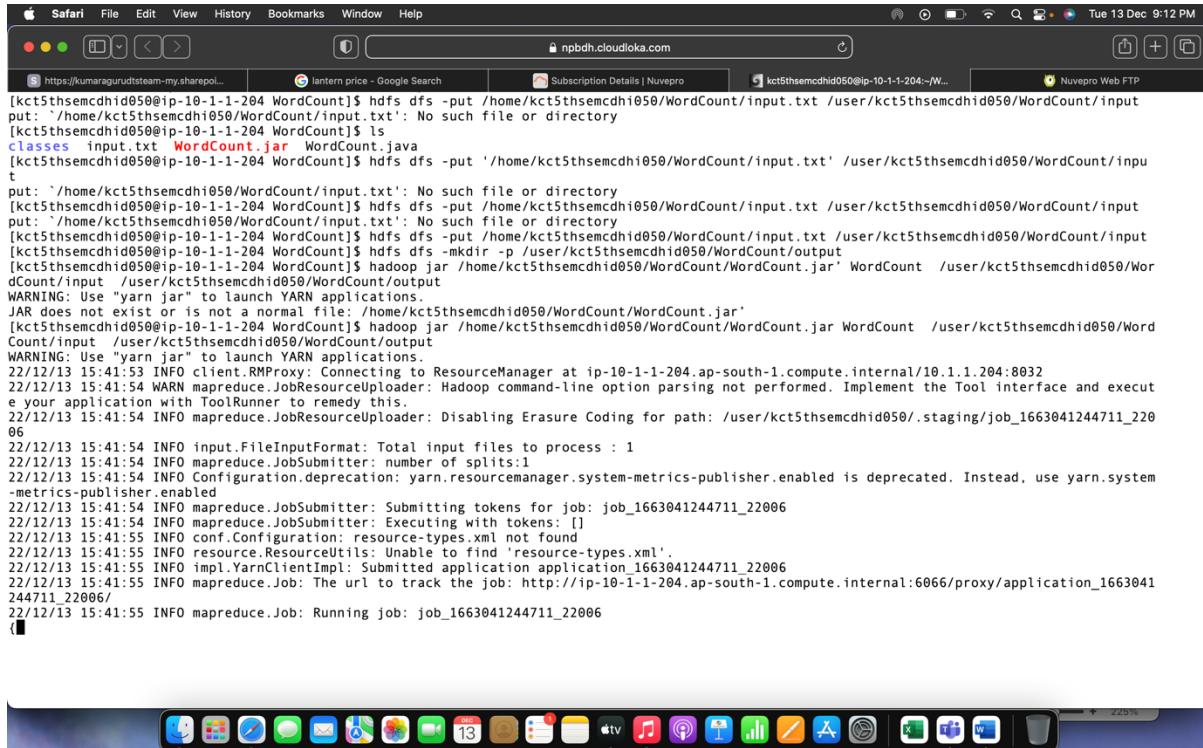
9. CREATE A DIRECTORY IN HADOOP

10. CREATE A DIRECTORY IN HADOOP

```
Safari File Edit View History Bookmarks Window Help npbdh.cloudloka.com Tue 13 Dec 9:08 PM
https://kumaraguruudteam-my.sharepo... lantern price - Google Search Subscription Details | Nuvepro kct5thsemcdhid050@ip-10-1-1-204:/~W... Nuvepro Web FTP
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdh1050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: '/home/kct5thsemcdh1050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdh1050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
classes input.txt WordCount.jar WordCount.java
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put '/home/kct5thsemcdh1050/WordCount/input.txt' /user/kct5thsemcdhid050/WordCount/input
put: '/home/kct5thsemcdh1050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put '/home/kct5thsemcdh1050/WordCount/input.txt' /user/kct5thsemcdhid050/WordCount/input
put: '/home/kct5thsemcdh1050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put '/home/kct5thsemcdh1050/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/kct5thsemcdhid050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: '/home/kct5thsemcdhid050/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/kct5thsemcdhid050/WordCount/input.txt' /user/kct5thsemcdhid050/WordCount/input
put: '/home/kct5thsemcdhid050/WordCount/input.txt': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhid050/WordCount/input.txt /user/kct5thsemcdhid050/WordCount/input
put: '/home/kct5thsemcdhid050@ip-10-1-1-204 WordCount': No such file or directory
[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -put /home/kct5thsemcdhid050/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/WordCount/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/WordCount/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 execute 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_1663041244711_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_1663041244711_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=1
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

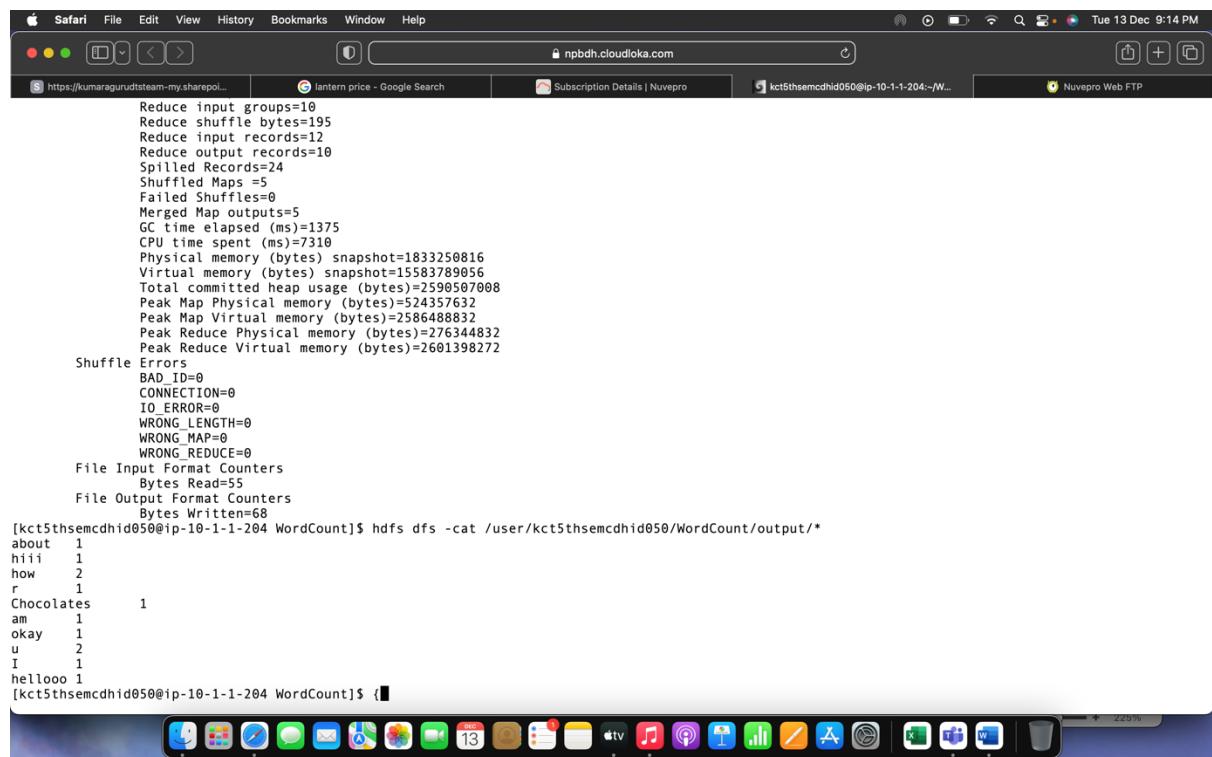
Map-Reduce Framework
    Map input records=7
    Map output records=12
    Map output bytes=102
    Map output materialized bytes=195
    Input split bytes=133
    Combine input records=0
    Combine output records=0
    Reduce input groups=10
    Reduce shuffle bytes=195
    Reduce input records=12
    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/*



```

Safari File Edit View History Bookmarks Window Help
npbdh.cloudloka.com
Tue 13 Dec 9:14 PM
https://kumaragurudhsteam-my.sharepo... Lantern price - Google Search Subscription Details | Nuvepro
kct5thsemcdhid050@ip-10-1-1-204:~/W...
Nuvepro Web FTP

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
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=55
File Output Format Counters
Bytes Written=68

[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ hdfs dfs -cat /user/kct5thsemcdhid050/WordCount/output/*
about      1
hiii      1
how       2
r         1
Chocolates   1
am         1
okay      1
u          2
I          1
hellooo   1

[kct5thsemcdhid050@ip-10-1-1-204 WordCount]$ 

```

U18ISI5202-BIG DATA LABORATORY

NAME: CHAARVIKA.A

ROLL NUMBER: 20BIS012

S.NO: 4

TOPIC : Implement a Map Reduce Program to analyse time-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

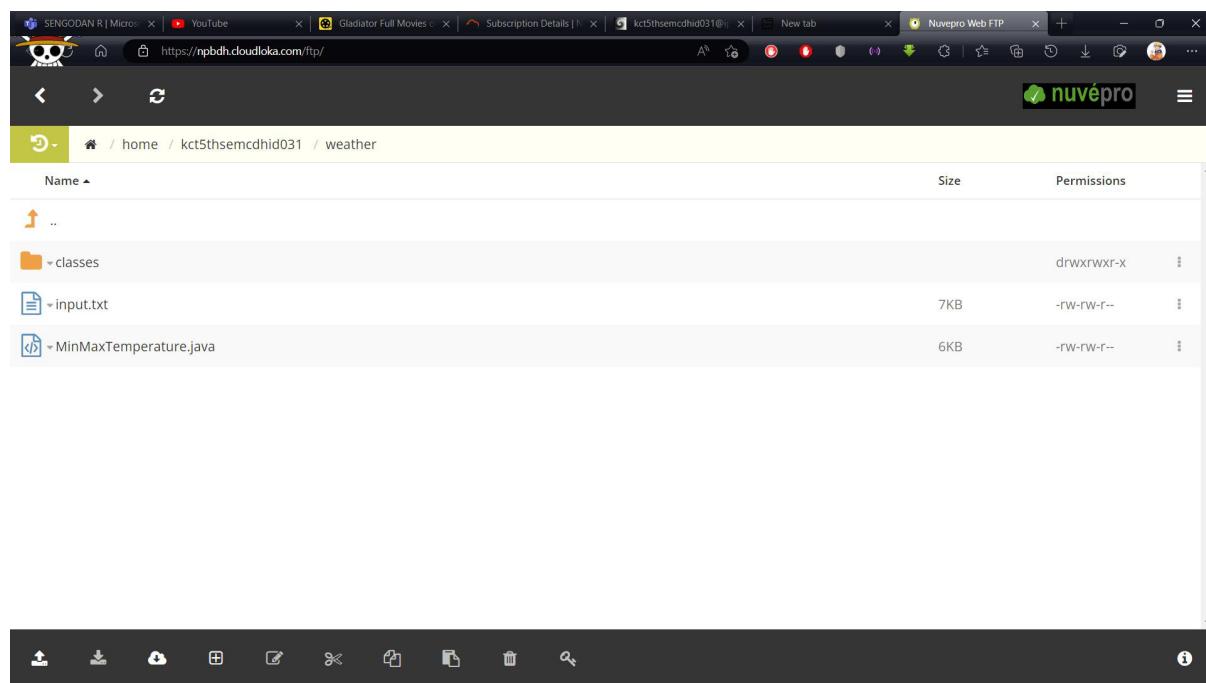
PROCEDURE:

STEP1-LOGIN TO THE WEB SHELL USING THE USERNAME AND PASSWORD

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

```

1 CA_25-Jan-2014 00:12:345 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 -22.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
2 CA_26-Jan-2014 00:54:245 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
3 CA_27-Jan-2014 00:14:045 35.7 01:19:345 23.1 02:34:542 -22.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
4 CA_28-Jan-2014 00:22:315 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
5 CA_29-Jan-2014 00:15:345 15.7 01:19:345 23.1 02:34:542 52.9 03:12:187 16 04:00:093 -14 05:12:345 45.0 06:19:345 23.1 07:34:542 -2.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -17 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
6 NJ_29-Jan-2014 00:15:345 15.7 01:19:345 23.1 02:34:542 52.9 03:12:187 16 04:00:093 -14 05:12:345 45.0 06:19:345 23.1 07:34:542 -2.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -17 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
7 CA_30-Jan-2014 00:22:445 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 39.6 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
8 CA_31-Jan-2014 00:42:245 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 49.2 06:19:345 23.1 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
9 NY_29-Jan-2014 00:15:345 15.7 01:19:345 23.1 02:34:542 52.9 03:12:187 16 04:00:093 -14 05:12:345 45.0 06:19:345 23.1 07:34:542 -2.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -17 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
10 NY_30-Jan-2014 00:22:445 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 39.6 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
11 NY_31-Jan-2014 00:42:245 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 49.2 06:19:345 23.1 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
12 NJ_30-Jan-2014 00:22:445 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 39.6 07:34:542 12.3 08:12:187 16
09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16
22:00:093 -7
13 ALIS_25-Jan-2014 00:12:345 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16

```

Auto-save Save Close

STEP6-PASTE THE SOURCE CODE IN THE FILE

```

23
24 public static class WhetherForecastMapper extends Mapper<Object, Text, Text, Text> {
25     public void map(Object keyOffSet, Text dayReport, Context con) throws IOException, InterruptedException {
26         StringTokenizer strTokens = new StringTokenizer(dayReport.toString(), "\t");
27         StringTokenizer strTokens = new StringTokenizer(dayReport.toString(), "\t");
28         int counter = 0;
29         Float currnetTemp = null;
30         Float minTemp = Float.MAX_VALUE;
31         Float maxTemp = Float.MIN_VALUE;
32         String date = null;
33         String currentTime = null;
34         String minTempANDTime = null;
35         String maxTempANDTime = null;
36
37         while (strTokens.hasMoreElements()) {
38             if (counter == 0) {
39                 date = strTokens.nextToken();
40             } else {
41                 if (counter % 2 == 1) {
42                     currentTime = strTokens.nextToken();
43                 }
44                 else {
45
46                     currnetTemp = Float.parseFloat(strTokens.nextToken());
47                     /* DecimalFormat df = new DecimalFormat();
48                     currnetTemp = df.parse(strTokens.nextToken()).floatValue(); */
49                     if (minTemp > currnetTemp) {
50                         minTemp = currnetTemp;
51                         minTempANDTime = minTemp + " AND " + currentTime;
52                     }
53                     if (maxTemp < currnetTemp) {
54                         maxTemp = currnetTemp;
55                         maxTempANDTime = maxTemp + " AND " + currentTime;
56                     }
57                 }
58             }
59             counter++;
60
61         }
62     }
63 }

```

Auto-save Save Close

```
59     counter++;
60 }
61 // Write to context - MinTemp, MaxTemp and corresponding time
62 Text temp = new Text();
63 temp.set(maxTempANDTime);
64 Text dateText = new Text();
65 dateText.set(date);
66 try {
67     con.write(dateText, temp);
68 } catch (Exception e) {
69     e.printStackTrace();
70 }
71
72 temp.set(minTempANDTime);
73 dateText.set(date);
74 con.write(dateText, temp);
75 }
76 }
77
78 public static class WhetherForcastReducer extends Reducer<Text, Text, Text, Text> {
79     MultipleOutputs<Text, Text> mos;
80
81     public void setup(Context context) {
82         mos = new MultipleOutputs<Text, Text>(context);
83     }
84
85     public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
86         int counter = 0;
87         String reducerInputStr[] = null;
88         String f1Time = "";
89         String f2Time = "";
90         String f1 = "", f2 = "";
91         Text result = new Text();
92         for (Text value : values) {
93
94             if (counter == 0) {
95                 reducerInputStr = value.toString().split("AND");
96             }
97             if (counter == 1) {
98                 f1Time = value.toString();
99             }
100            if (counter == 2) {
101                f2Time = value.toString();
102            }
103            if (counter == 3) {
104                f1 = value.toString();
105            }
106            if (counter == 4) {
107                f2 = value.toString();
108            }
109            if (counter == 5) {
110                break;
111            }
112        }
113        if (f1Time != "" && f2Time != "") {
114            if (f1Time.compareTo(f2Time) > 0) {
115                f1Time = f2Time;
116            }
117            if (f1.compareTo(f2) > 0) {
118                f1 = f2;
119            }
120            result.set(f1 + " " + f1Time);
121            mos.write(key, result);
122        }
123    }
124 }
```

```
107 }  
108 if (Float.parseFloat(f1) > Float.parseFloat(f2)) {  
109  
110     result = new Text("Time: " + f2Time + " MinTemp: " + f2 + "\t" + "Time: " + f1Time + " MaxTemp: " + f1);  
111 } else {  
112     result = new Text("Time: " + f1Time + " MinTemp: " + f1 + "\t" + "Time: " + f2Time + " MaxTemp: " + f2);  
113 }  
114 String fileName = "";  
115 if (key.toString().substring(0, 2).equals("CA")) {  
116     fileName = MinMaxTemperature.caloutputName;  
117 } else if (key.toString().substring(0, 2).equals("NY")) {  
118     fileName = MinMaxTemperature.nyoutputname;  
119 } else if (key.toString().substring(0, 2).equals("NJ")) {  
120     fileName = MinMaxTemperature.njoutputname;  
121 } else if (key.toString().substring(0, 3).equals("AUS")) {  
122     fileName = MinMaxTemperature.ausoutputName;  
123 } else if (key.toString().substring(0, 3).equals("BOS")) {  
124     fileName = MinMaxTemperature.bosoutputName;  
125 } else if (key.toString().substring(0, 3).equals("BAL")) {  
126     fileName = MinMaxTemperature.baloutputName;  
127 }  
128 String strArr[] = key.toString().split("_");  
129 key.set(strArr[1]); //key is date value  
130 mos.write(fileName, key, result);  
131 }  
132  
133 @Override  
134 public void cleanup(Context context) throws IOException, InterruptedException {  
135     mos.close();  
136 }  
137 }  
138  
139 public static void main(String[] args) throws IOException,  
140     ClassNotFoundException, InterruptedException {  
141     Configuration conf = new Configuration();  
142     Job job = Job.getInstance(conf, "Weather Statistics of USA");  
143     job.setJarByClass(MinMaxTemperature.class);  
144 }
```

The screenshot shows a web browser window with multiple tabs open. The active tab displays the source code of a Java file named 'MinMaxTemperature.java'. The code is a MapReduce job configuration for processing weather statistics. It includes imports for various Hadoop classes, configuration setup, multiple output paths, and a main class definition.

```
141 Configuration conf = new Configuration();
142 Job job = Job.getInstance(conf, "Weather Statistics of USA");
143 job.setJarByClass(MinMaxTemperature.class);
144
145 job.setMapperClass(WeatherForcastMapper.class);
146 job.setReducerClass(WeatherForcastReducer.class);
147
148 job.setOutputKeyClass(Text.class);
149 job.setOutputValueClass(Text.class);
150
151 job.setOutputKeyClass(Text.class);
152 job.setOutputValueClass(Text.class);
153
154 Multipleoutputs.addNamedOutput(job, calOutputName,TextOutputFormat.class, Text.class, Text.class);
155 Multipleoutputs.addNamedOutput(job, nyOutputName,TextOutputFormat.class, Text.class, Text.class);
156 Multipleoutputs.addNamedOutput(job, njOutputName,TextOutputFormat.class, Text.class, Text.class);
157 Multipleoutputs.addNamedOutput(job, bosOutputName,TextOutputFormat.class, Text.class, Text.class);
158 Multipleoutputs.addNamedOutput(job, ausOutputName,TextOutputFormat.class, Text.class, Text.class);
159 Multipleoutputs.addNamedOutput(job, balOutputName,TextOutputFormat.class, Text.class, Text.class);
160
161 // FileInputFormat.addInputPath(job, new Path(args[0]));
162 // FileOutputFormat.setOutputPath(job, new Path(args[1]));
163 Path outputPath = new Path(args[1]);
164 //Configuring the input/output path from the filesystem into the job
165 FileInputFormat.addInputPath(job, new Path(args[0]));
166 FileOutputFormat.setOutputPath(job, new Path(args[1]));
167 //deleting the output path automatically from hdfs so that we don't have to
168 //delete it explicitly
169 outputPath.getFileSystem(conf).delete(outputPath);
170 //Exiting the job only if the flag value becomes false
171 try {
172     System.exit(job.waitForCompletion(true) ? 0 : 1);
173 } catch (Exception e) {
174     // TODO Auto-generated catch block
175     e.printStackTrace();
176 }
```

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.caOutputName;
    } 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'

```

SENGODAN R | Micro: | YouTube | Gladiator Full Movies | Subscription Details | kct5thsemcdhid031@ip-10-1-1-204 | New tab | Nuvepro Web FTP | + | - | x | ... | https://npbdh.cloudloka.com:4200

npbdh login: kct5thsemcdhid031
kct5thsemcdhid031@npbdh.cloudloka.com's password:
Last login: Tue Dec 13 06:07:19 2022 from ec2-65-1-45-35.ap-south-1.compute.amazonaws.com
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ ls
matrix weather wordcount
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ cd weatherweather
-bash: cd: weatherweather: No such file or directory
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ cd weather
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll
total 20
drwxrwxr-x 2 kct5thsemcdhid031 kct5thsemcdhid031 4096 Dec 13 16:11 classes
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 7262 Dec 13 16:10 input.txt
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 6317 Dec 13 16:10 MinMaxTemperature.java
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ export HADOOP_CLASSPATH=$(hadoop classpath)
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop/libexec/../../../hadoop/../../../opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/p0.1425774/lib/hadoop/libexec/../../hadoop-hdfs/lib:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.ceis/CDH/lib/hadoop-mapreduce/../../opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hado
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ javac -classpath ${HADOOP_CLASSPATH} -d '/home/kct5thsemcdhid031/weather/classes' '/home/kct5thsemcdhid031/weather/MinMaxTemperature.java'
Note: /home/kct5thsemcdhid031/weather/MinMaxTemperature.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll

```

STEP10-CREATE A JAR FILE

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

```

SENGODAN R | Micro: | YouTube | Gladiator Full Movies | Subscription Details | kct5thsemcdhid031@ip-10-1-1-204 | New tab | Nuvepro Web FTP | + | - | x | ... | https://npbdh.cloudloka.com:4200

npbdh login: kct5thsemcdhid031
kct5thsemcdhid031@npbdh.cloudloka.com's password:
Last login: Tue Dec 13 06:07:19 2022 from ec2-65-1-45-35.ap-south-1.compute.amazonaws.com
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ ls
matrix weather wordcount
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ cd weatherweather
-bash: cd: weatherweather: No such file or directory
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ cd weather
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll
total 20
drwxrwxr-x 2 kct5thsemcdhid031 kct5thsemcdhid031 4096 Dec 13 16:11 classes
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 7262 Dec 13 16:10 input.txt
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 6317 Dec 13 16:10 MinMaxTemperature.java
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ export HADOOP_CLASSPATH=$(hadoop classpath)
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/libexec/../../hadoop/libexec/../../../hadoop/../../../opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/p0.1425774/lib/hadoop/libexec/../../hadoop-hdfs/lib:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.ceis/CDH/lib/hadoop-mapreduce/../../opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hado
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ javac -classpath ${HADOOP_CLASSPATH} -d '/home/kct5thsemcdhid031/weather/classes' '/home/kct5thsemcdhid031/weather/MinMaxTemperature.java'
Note: /home/kct5thsemcdhid031/weather/MinMaxTemperature.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ jar -cvf WordCount.jar -C '/home/kct5thsemcdhid031/weather/classes' .
added manifest
adding: MinMaxTemperature$WeatherForcastReducer.class(in = 3651) (out= 1546)(deflated 57%)
adding: MinMaxTemperature$WeatherForcastMapper.class(in = 2437) (out= 1143)(deflated 53%)
adding: MinMaxTemperature.class(in = 2734) (out= 1349)(deflated 50%)
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll
total 28
drwxrwxr-x 2 kct5thsemcdhid031 kct5thsemcdhid031 4096 Dec 13 16:16 classes
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 7262 Dec 13 16:10 input.txt
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 4880 Dec 13 16:31 MinMaxTemperature.jar
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 6317 Dec 13 16:10 MinMaxTemperature.java
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll

```

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

```
npbdu login: kct5thsemcdhid031
kct5thsemcdhid031@npbdu.cloudloka.com's password:
Last login: Tue Dec 13 06:07:19 2022 from ec2-65-1-45-35.ap-south-1.compute.amazonaws
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ ls
matrix weather wordcount
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ cd weatherweather
-bash: cd: weatherweather: No such file or directory
[kct5thsemcdhid031@ip-10-1-1-204 ~]$ cd weather
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll
total 20
drwxrwxr-x 2 kct5thsemcdhid031 kct5thsemcdhid031 4096 Dec 13 16:11 classes
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 7262 Dec 13 16:10 input.txt
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 6317 Dec 13 16:10 MinMaxTemperatur
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ export HADOOP_CLASSPATH=$(hadoop classpath)
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/1
adoop/libexec/../../../../hadoop/*:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425
774/lib/hadoop/libexec/../../../../hadoop-hdfs/lib/*:/opt/cloudera/parcels/CDH-6.
cecs/CDH/lib/hadoop-mapreduce/../../../../hadoop/parcels/CDH-6.2.1-1.cdh6.2.1.p0.14
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ javac -classpath ${HADOOP_CLASSPATH} -d
/home/kct5thsemcdhid031/weather/MinMaxTemperature.java uses or overrides a de
Note: /home/kct5thsemcdhid031/weather/MinMaxTemperature.java uses or overrides a de
Note: Recompile with -Xlint:deprecation for details.
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ jar -cvf MinMaxTemperature.jar -C '/hom
added manifest
adding: MinMaxTemperature$WeatherForcastReducer.class(in = 3651) (out= 1546)(deflat
adding: MinMaxTemperature$WeatherForcastMapper.class(in = 2437) (out= 1143)(deflate
adding: MinMaxTemperature.class(in = 2734) (out= 1349)(deflated 50%)
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ ll
total 28
drwxrwxr-x 2 kct5thsemcdhid031 kct5thsemcdhid031 4096 Dec 13 16:16 classes
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 7262 Dec 13 16:10 input.txt
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 4888 Dec 13 16:31 MinMaxTemperatur
-rw-rw-r-- 1 kct5thsemcdhid031 kct5thsemcdhid031 6317 Dec 13 16:10 MinMaxTemperatur
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid0
31/weather
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ hdfs dfs -mkDir -p /user/kct5thsemcdhid0
31/weather/input/
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ hdfs dfs -put /home/kct5thsemcdhid031/we
ather/input.txt /user/kct5thsemcdhid031/weather/input/
[kct5thsemcdhid031@ip-10-1-1-204 weather]$ hdfs dfs -mkdir -p /user/kct5thsemcdhid0
31/weather/output/
[kct5thsemcdhid031@ip-10-1-1-204 weather]$
```

STEP14-RUN THE MAP REDUCE PROGRAM

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

```

SENGODAN R | Micro: X | YouTube X | Gladiator Full Movies ... X | Subscription Details | X | kct5thsemcdhid031@... X | New tab X | Nuvelpro Web FTP X | + ...
https://npbdh.cloudloka.com:4200
[1kct5thsemcdhid031@ip-10-1-1-204 weather]$ hadoop jar /home/kct5thsemcdhid031/weather/MinMaxTemperature.jar MinMaxTemperature /user/kct5thsemcdhid031/weather/input /u
ser/kct5thsemcdhid031/weather/output
WARNING: Use "Yarn jar" to launch YARN applications.
22/12/13 16:41:45 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 16:41:46 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with T
oolRunner to remedy this
22/12/13 16:41:46 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/kct5thsemcdhid031/.staging/job_1663041244711_22058
22/12/13 16:41:46 INFO input.FileInputFormat: Total input files to process: 1
22/12/13 16:41:46 INFO mapreduce.JobSubmitter: number of splits:1
22/12/13 16:41:46 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enable
d
22/12/13 16:41:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1663041244711_22058
22/12/13 16:41:47 INFO mapreduce.JobSubmitter: Executing with tokens: []
22/12/13 16:41:47 INFO conf.Configuration: resource-types.xml not found
22/12/13 16:41:47 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/12/13 16:41:47 INFO impl.YarnHttpClientImpl: Submitted application application_1663041244711_22058
22/12/13 16:41:47 INFO mapreduce.Client: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041244711_22058/
22/12/13 16:41:47 INFO mapreduce.Job: Running job: job_1663041244711_22058
22/12/13 16:42:32 INFO mapreduce.Job: map 0% reduce 0%
22/12/13 16:42:32 INFO mapreduce.Job: map 100% reduce 100%
22/12/13 16:43:23 INFO mapreduce.Job: Job job_1663041244711_22058 completed successfully
22/12/13 16:43:25 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=906
  FILE: Number of bytes written=1364303
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=7393
  HDFS: Number of bytes written=1752
  HDFS: Number of read operations=40
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=34
  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)=32288
  Total time spent by all reduces in occupied slots (ms)=57813
  Total time spent by all map tasks (ms)=52288
Total time spent by all reduces in occupied slots (ms)=57813
Total time spent by all map tasks (ms)=32288
Total time spent by all reduce tasks (ms)=57813
Total vcore-milliseconds taken by all map tasks=32288
Total vcore-milliseconds taken by all reduce tasks=57813
Total megabyte-milliseconds taken by all map tasks=33062912
Total megabyte-milliseconds taken by all reduce tasks=59200512
Map-Reduce Framework
  Map input records=24
  Map output records=48
  Map output bytes=1584
  Map output materialized bytes=886
  Input split bytes=131
  Combine input records=0
  Combine output records=0
  Reduce input groups=24
  Reduce shuffle bytes=886
  Reduce input records=48
  Reduce output records=0
  Spilled Records=96
  Shuffled Maps=5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=1640
  CPU time spent (ms)=8140
  Physical memory (bytes) snapshot=1813049344
  Virtual memory (bytes) snapshot=15573565440
  Total committed heap usage (bytes)=2592604160
  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
[1kct5thsemcdhid031@ip-10-1-1-204 weather]$

```

STEP15-VERIFY THE OUTPUT:

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

SENGODAN R | Micro: X | YouTube X | Gladiator Full Movies X | Subscription Details X | kct5thsemcdhid031@... X | New tab X | Nuvepro Web FTP X | +

Total Committed Heap Usage (bytes)=2392684168
Peak Map Physical memory (bytes)=457826384
Peak Map Virtual memory (bytes)=2578415616
Peak Reduce Physical memory (bytes)=278921216
Peak Reduce Virtual memory (bytes)=2600792064

Shuffle Errors
BAD_TDE=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  
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: 53.9  
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
wai.nuvepro.com
Start Page Subscription Details | Nuvepro
Applications hadoop@ip-172-31-17-9...
hadoop@ip-172-31-17-9: ~/spark-3.2.0-bin-hadoop3.2/bin
File Edit View Search Terminal Help
22/12/13 05:47:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://ip-172-31-17-9.ap-south-1.compute.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1670910452273).
Spark session available as 'spark'.
Welcome to
  / \ / \ / \ / \
 / \ \ / \ / \ / \ / \
 / \ / \ \ / \ / \ / \ / \
 / \ / \ / \ \ / \ / \ / \ / \
version 3.2.0

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 11.0.17)
Type in expressions to have them evaluated.
Type :help for more information.

scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold")
res0: String = 1048576B

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)

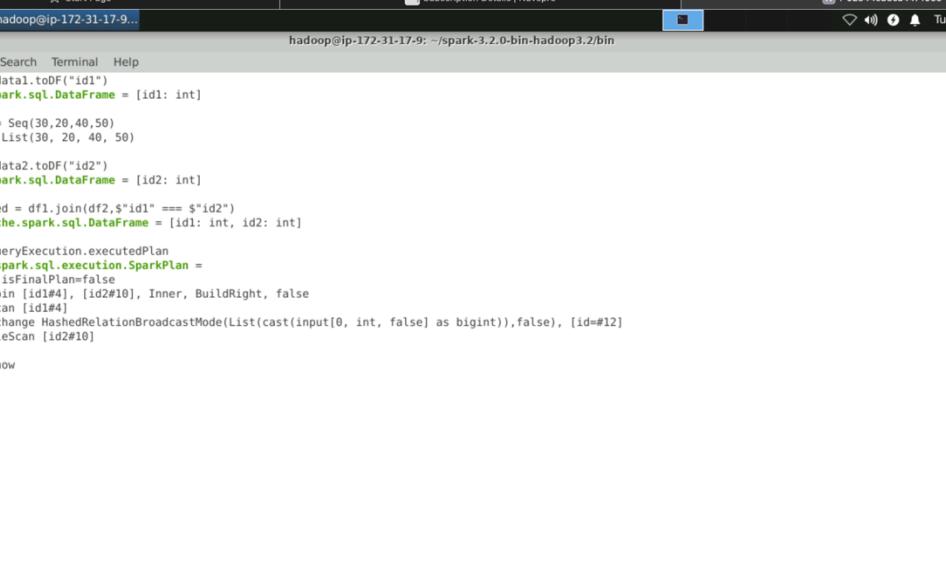
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int]

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
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]
```



A screenshot of a Mac OS X desktop environment. At the top, a Safari browser window is open to wai.nuvepro.com, showing 'Subscription Details | Nuvepro'. Below it, a terminal window titled 'hadoop@ip-172-31-17-9: /spark-3.2.0-bin-hadoop3.2/bin' displays a Scala session. The session starts with creating DataFrames from Seqs, performing a join operation, and then printing the resulting DataFrame. The terminal window has tabs for 'Applications' and 'hadoop@ip-172-31-17-9...'. The system status bar at the top right shows the date as 'Tue 13 Dec 11:22 AM' and a process ID 'i-02344eb3ea4471065'. The dock at the bottom contains icons for various Mac applications like Mail, Calendar, and Finder.

```
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.explain()
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|
| 50| 50|
+---+---+
scala>
```

Microsoft Teams Edit View Window Help

Amazon.in wai.nuvepro.com

Applications : Install Apache Spark on... hadoop@ip-172-31-22-6...

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.executedPlan
<console>:24: error: value excutedPlan is not a member of org.apache.spark.sql.execution.QueryExecution
  dfJoined.queryExecution.executedPlan
^
```

```
Safari File Edit View History Bookmarks Window Help
wai.nuvepro.com
Applications [Install Apache Spark on... hadoop@ip-172-31-22-6...
File Edit View Search Terminal Help
scala> spark.conf.get("spark.sql.join.preferSortMergeJoin")
res1: String = false

scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold")
res2: 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.execution.executedPlan
<console>:24: error: value executedPlan is not a member of org.apache.spark.sql.execution.QueryExecution
           dfJoined.execution.executedPlan
                           ^

scala> dfJoined.execution.executedPlan
res3: 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
wai.nuvepro.com
Applications [Install Apache Spark on... hadoop@ip-172-31-22-6...
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.execution.executedPlan
<console>:24: error: value executedPlan is not a member of org.apache.spark.sql.execution.QueryExecution
           dfJoined.execution.executedPlan
                           ^

scala> dfJoined.execution.executedPlan
res5: 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|
| 50| 50|
+---+---+
```

U18ISI5202-BIG DATA LABORATORY

NAME: CHAARVIKA.A

ROLL NUMBER: 20BIS012

S.NO: 9

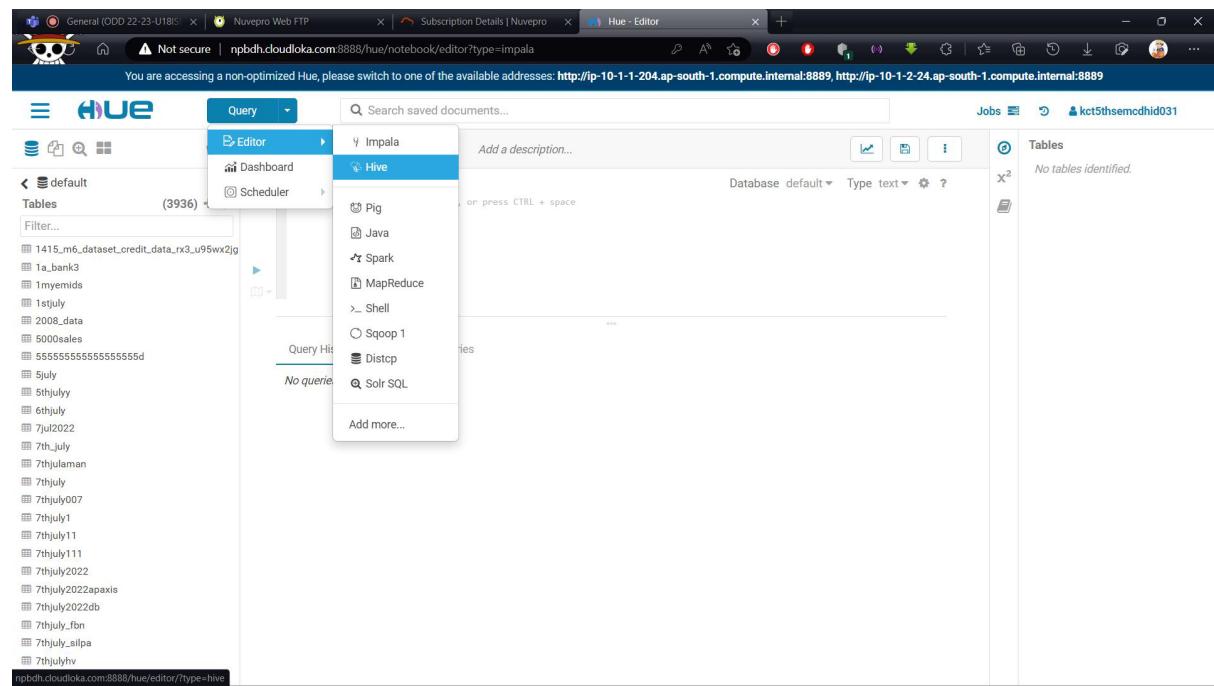
TOPIC : Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes

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



The screenshot shows the Hue Editor interface. At the top, there are tabs for 'General (ODD 22-23-U1B1)', 'Nuvepro Web FTP', 'Subscription Details | Nuvepro', and 'Hue - Editor'. The URL is npbdh.cloudloka.com:8888/hue/editor?type=hive. A message at the top says: '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'. The main area has a 'Query' tab selected, showing a search bar 'Search saved documents...'. Below it is a 'Hive' section with 'Add a name...' and 'Add a description...' fields, and a dropdown 'Unset from default application'. To the right, there are buttons for 'Database', 'Type', and 'Text'. A 'Tables' section shows 'No tables identified.' Below the query area, there are tabs for 'Query History' and 'Saved Queries', both of which are empty.

STEP2: Create database

create database 20bis031bigdatahive; show databases; use 20bis031bigdatahive;

The screenshot shows the Hue Editor interface after running the commands. The 'Query' tab is selected, displaying the following SQL code and its execution results:

```
1 show databases;
2 create database 20bis031bigdatahive;
3 show databases;
4 use 20bis031bigdatahive;
5 |
```

Output:

```
1.47s Database default Type text
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20221213072020_0b5001fa-cc2f-478c-932b-d2281fc252c2); Time taken: 0.004 seconds
INFO : OK
```

The 'Tables' section on the right shows 'Statement 4/4' and 'No tables identified.'. Below the query area, there are tabs for 'Query History' and 'Saved Queries', with several entries listed:

- 2 minutes ago ✓ use 20bis031bigdatahive
- 12 minutes ago ✓ show databases
- 12 minutes ago show databases
- 13 minutes ago ✓ show databases; create database 20bis031bigdatahive;
- 14 minutes ago ✓ show databases;

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>

HUE

Query Search saved documents...

Tables (3938)

Filter...

default

- Tables (3938) +
- Filter...
- ipl_team1
- demog
- user3
- partitionprac
- cause_of_deaths_partitioned
- products_buc
- ptable
- respatent
- act1_par_buc2_ananth
- status_data
- mysalary2
- pokemon_ukj
- hbase_table_emp1
- demog_g1_t5
- airlines1
- newsales_records6
- iptable
- '7thJuly007'
- hostvn
- cust_table
- maxi
- txnrecordsdha
- stocks_task3
- students1
- cquery4

14thJuly18

14thJuly_silpa

14thJulydb

14thJulySandb

16jan2022anand

18bd1a0420

1carddb

1stJulyPritam

20bis031bigdatahive

21Jun2022

21June2022test

22JulyManit

22Jun2022

22jun22dsuraj

22juneCards

23jj

23July_pkg

23JulyManit

25June

26march

2mich_retail_db

29_E05341

Jobs

Tables Statement 3/3
No tables identified.

STEP3:Create table

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>

HUE

Query Search saved documents...

Hive Add a name... Add a description...

Tables

Add

7.4s Database default Type text ?

```
create table studentskct(rollno varchar(10), subject varchar(10), mark int, grade varchar(10));
```

INFO : Starting task [Stage-0:DDL] in serial mode

INFO : Completed executing command(queryId=hive_20221213073116_5b34a723-c3d5-4379-9585-44a03966753b); Time taken: 0.059 seconds

INFO : OK

Success.

Query History Saved Queries

a few seconds ago	create table studentskct(rollno varchar(10), subject varchar(10), mark int, grade varchar(10))
a minute ago	create table studentkct(rollno varchar(10), subject varchar(10), mark int, grade varchar(10))
3 minutes ago	show databases; create database 20bis031bigdatahive; show databases; use 20bis031bigdatahive; create table studentise(rollno varchar(10), subject varchar(10), mark int, grade varchar(10)); select * from studentise;
3 minutes ago	show databases; create database 20bis031bigdatahive; show databases; use 20bis031bigdatahive; create table studentise(rollno varchar(10), subject varchar(10), mark int, grade varchar(10));

Tables No tables identified.

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;

```

1 create table studentskct(rollno varchar(10) , subject varchar(10) , mark int , grade varchar(10));
2 insert into studentskct values('20bis100','cs',90,'A'),
3 ('20bis101','cs',90,'A'),
4 ('20bis102','is',80,'B'),
5 ('20bis103','cs',80,'B'),
6 ('20bis104','is',60,'D'),
7 ('20bis105','is',70,'C');
8 select * from studentskct;

```

studentskct.rollno	studentskct.subject	studentskct.mark	studentskct.grade
1 20bis100	cs	90	A
2 20bis101	cs	90	A
3 20bis102	is	80	B
4 20bis103	cs	80	B
5 20bis104	is	60	D
6 20bis105	is	70	C

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

```

1 select * from studentskct where mark<80;
2

```

studentskct.rollno	studentskct.subject	studentskct.mark	studentskct.grade
1 20bis104	is	60	D
2 20bis105	is	70	C

STEP6:Drop table
drop table studentskct;

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>

HUE Query Search saved documents...

Hive Add a name... Add a description...

default Tables (3945) +

1| drop table studentskct;

1.25s Database default Type text ?

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 sec
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...
x² default.studentskct

The screenshot shows the Hue Hive Editor interface. On the left, there's a sidebar with a list of tables under the 'default' database, including 'ipl_team', 'demog', 'user3', 'partitionprac', 'cause_of_deaths_partitioned', 'products_buc', 'ptable', 'respatient', 'sc1_par_buc2_ananth', 'status_data', 'nyseDaily2', 'pokemon_ujk', 'hbase_table_emp1', 'demog_g1_t5', 'airlines1', 'newsales_records6', 'iptable', '7thJuly007', 'hostvn', 'cust_table', 'maxi', 'txnrecordsdha', 'stocks_task3', 'students1', and 'cqquery4'. The main area has a query editor with the command 'drop table studentskct;'. Below it, a log window shows the execution details: 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 sec, and INFO : OK. A success message 'Success.' is displayed. At the bottom, there are tabs for 'Query History' and 'Saved Queries', with several recent queries listed. A sidebar on the right shows a table of tables with a filter input field.

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
```

```
scala>
scala>
scala>
scala>
scala>
scala>
scala> dfJoined.queryExecution.executedPlan
res4: 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|
+---+---+
```

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)
```

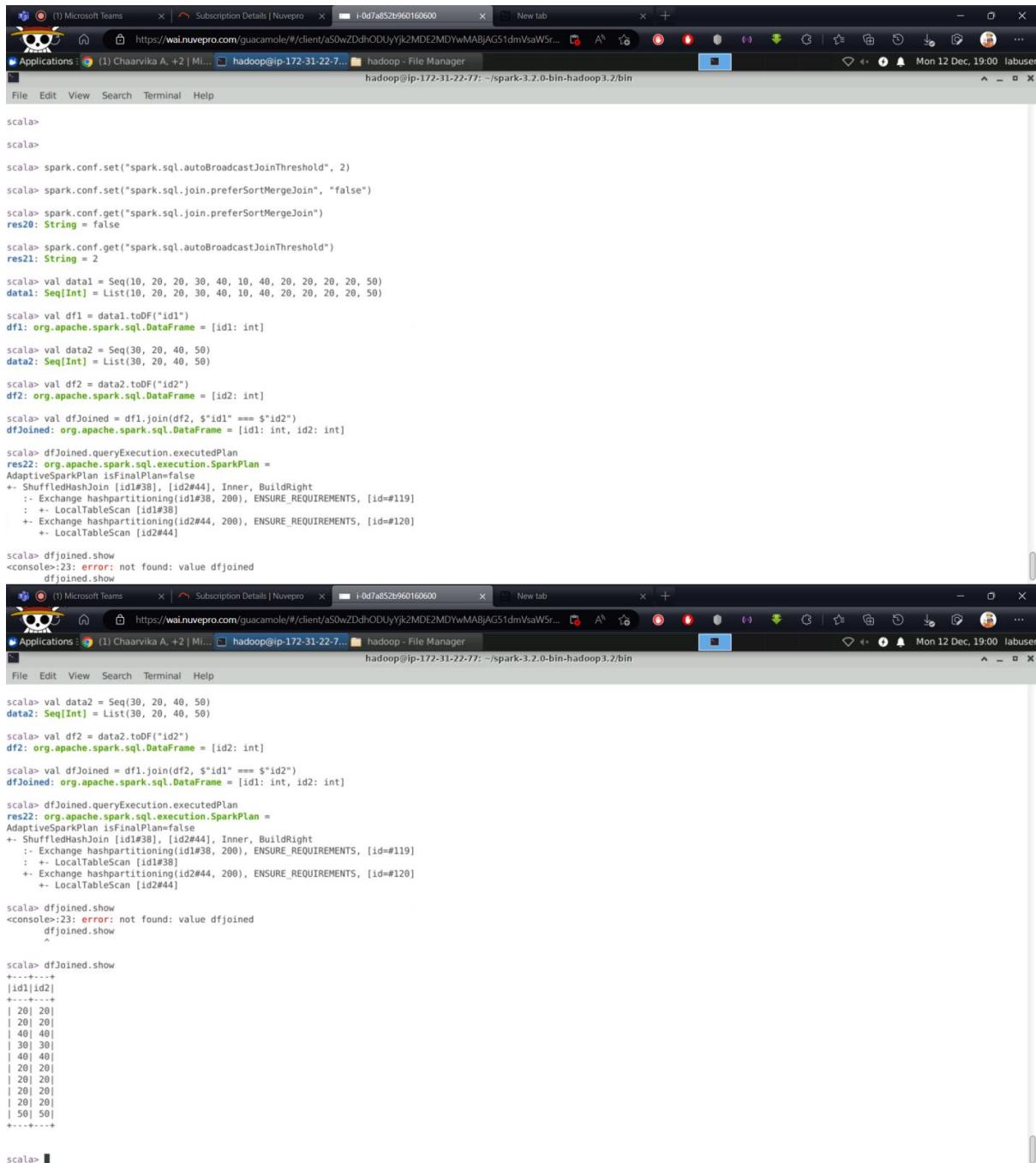
```
scala> val df1 = data1.toDF("id1")
df1: org.apache.spark.sql.DataFrame = [id1: int]
```

```
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
```



```

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")
res20: String = false
scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold")
res21: 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.executionPlan(res22: org.apache.spark.sql.execution.SparkPlan =
AdaptiveSparkPlan.isFinalPlan=false
+- ShuffledHashJoin [id1#38], [id2#44], Inner, BuildRight
  :- Exchange hashpartitioning(id1#38, 200), ENSURE_REQUIREMENTS, [id=#119]
  :+ LocalTableScan [id1#38]
  +- Exchange hashpartitioning(id2#44, 200), ENSURE_REQUIREMENTS, [id=#120]
  +- LocalTableScan [id2#44]
scala> dfJoined.show
<console>:23: error: not found: value dfJoined
      dfJoined.show

```



```

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.executionPlan(res22: org.apache.spark.sql.execution.SparkPlan =
AdaptiveSparkPlan.isFinalPlan=false
+- ShuffledHashJoin [id1#38], [id2#44], Inner, BuildRight
  :- Exchange hashpartitioning(id1#38, 200), ENSURE_REQUIREMENTS, [id=#119]
  :+ LocalTableScan [id1#38]
  +- Exchange hashpartitioning(id2#44, 200), ENSURE_REQUIREMENTS, [id=#120]
  +- LocalTableScan [id2#44]
scala> dfJoined.show
<console>:23: error: not found: value dfJoined
      dfJoined.show

```

+---+---+

id1	id2
10	20
10	20
20	20
30	30
40	40
40	20
50	20
50	50

+---+---+

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
```

The screenshot shows two terminal windows side-by-side. Both windows have a title bar with tabs for Microsoft Teams, Subscription Details | Nuvepro, and a file path starting with i-0d7a852b960160600. The second tab in both windows is labeled hadoop@ip-172-31-22-7... / hadoop - File Manager.

The left terminal window contains the following Scala code:

```
scala>
scala>
scala>
scala>
scala>
scala> spark.conf.get("spark.sql.join.preferSortMergeJoin")
res25: String = false
scala> spark.conf.get("spark.sql.autoBroadcastJoinThreshold")
res26: 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
res27: org.apache.spark.sql.execution.SparkPlan =
CartesianProduct (id1#63 >= id2#69)
:- LocalTableScan [id1#63]
+- LocalTableScan [id2#69]
scala> dfJoined.show
```

The right terminal window shows the output of the `dfJoined.show` command, displaying the joined data as a table:

```
+---+---+
|id1|id2|
+---+---+
| 20| 20|
| 20| 20|
| 30| 30|
| 30| 20|
| 40| 30|
| 40| 20|
| 40| 40|
| 40| 30|
| 40| 20|
| 20| 20|
| 20| 20|
| 20| 20|
| 50| 30|
| 50| 20|
| 40| 40|
| 50| 40|
| 50| 50|
+---+---+
```