



M-4, 1st Floor, Old DLF Colony, Sector 14, Gurugram, Haryana 122001

# Project report on Big data

**Name**

Jairam J S

**Email ID**

jsjairam01@gmail.com

## Acknowledgement

Not all professionals do their work by themselves. Although they can be as prolific or as adept in their respective fields, they will still need assistance one way or another. For instance, writing a body of work takes a lot of research. They often depend on their assistants or subordinates to gather information about the subject matter. Aside from the research people, other individuals can also receive credit for their contributions to the writer's work.

# CONTENTS

S. no	Project name	Page. no
1	<b>Titanic data analysis</b> <ul style="list-style-type: none"> <li>1. find the average age of males and females tragedy.</li> <li>2. find the number of people died in each class with their genders and ages</li> </ul>	5-10     11-15
2	<b>Uber data analysis</b> <ul style="list-style-type: none"> <li>1. find the days on which each basement has more trips.</li> <li>2. find the days on which each basement has more number of active vehicles.</li> </ul>	16-21     22-27
3	<b>Pokemon data analysis</b>	28-38
4	<b>Time zone analysis</b>	39-40

5	<b>Word Count</b> 1. Word Count using java 2. Word Count using hive 3. Word Count using pig	41-47 47-48 49-53
6	<b>Data compression using Hive</b>	54-60
7	<b>YouTube data analysis</b>	61-67

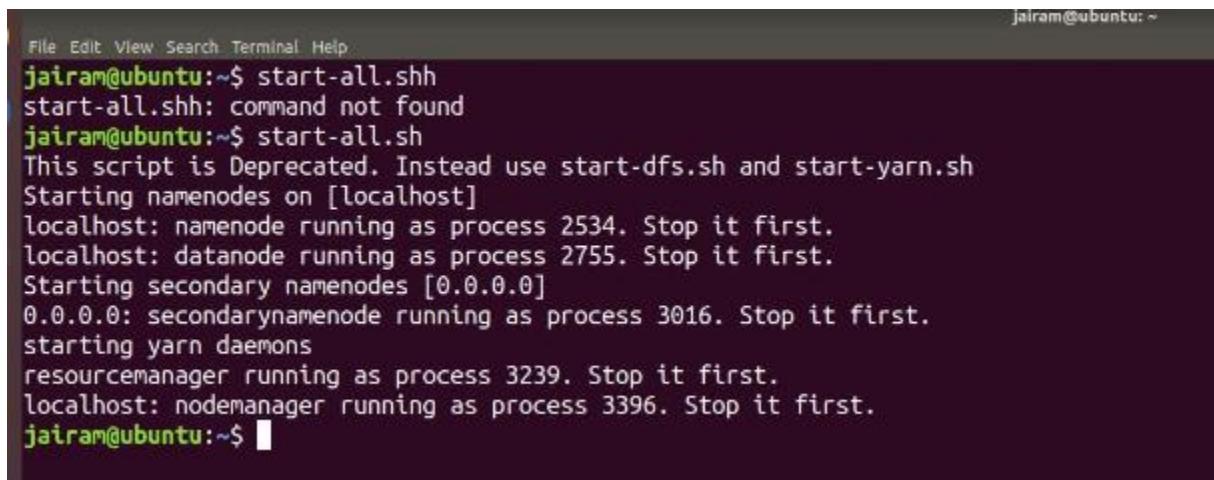
## Titanic data analysis

Problem 1:

In this problem statement, we will find the average age of males and females who died in the Titanic tragedy.

1. Open terminal and type

Command: ***start-all.sh***



```
File Edit View Search Terminal Help
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ █
```

2. Open bashrc file by ***nano ~/.bashrc*** and type

```
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
```

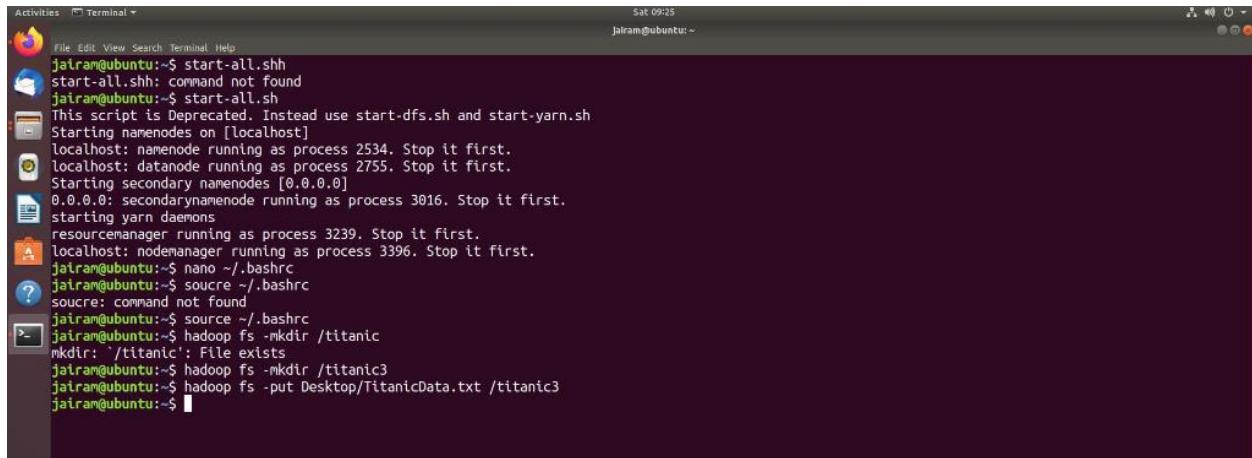
3. Then type ***source ~/.bashrc***

```
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$
```

command: *hadoop fs -mkdir /titanic3*

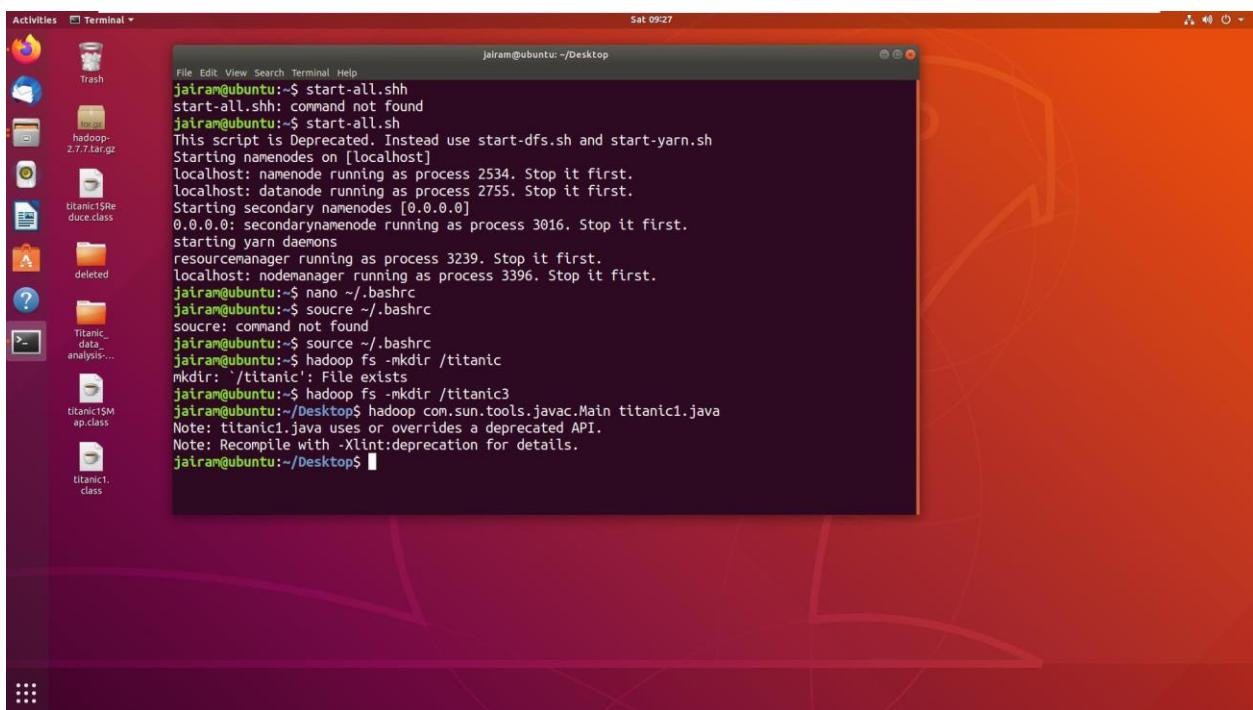
```
Activities Terminal * Sat 09:23 jairam@ubuntu:~$
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ hadoop fs -mkdir /titanic
mkdir: '/titanic': File exists
jairam@ubuntu:~$ hadoop fs -mkdir /titanic3
jairam@ubuntu:~$
```

#### 4. Command: *hadoop fs -put Desktop/TitanicData.txt /titanic3*



```
Activities Terminal Sat 09:25
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ hadoop fs -mkdir /titanic
mkdir: /titanic: File exists
jairam@ubuntu:~$ hadoop fs -mkdir /titanic3
jairam@ubuntu:~$ hadoop fs -put Desktop/TitanicData.txt /titanic3
jairam@ubuntu:~$
```

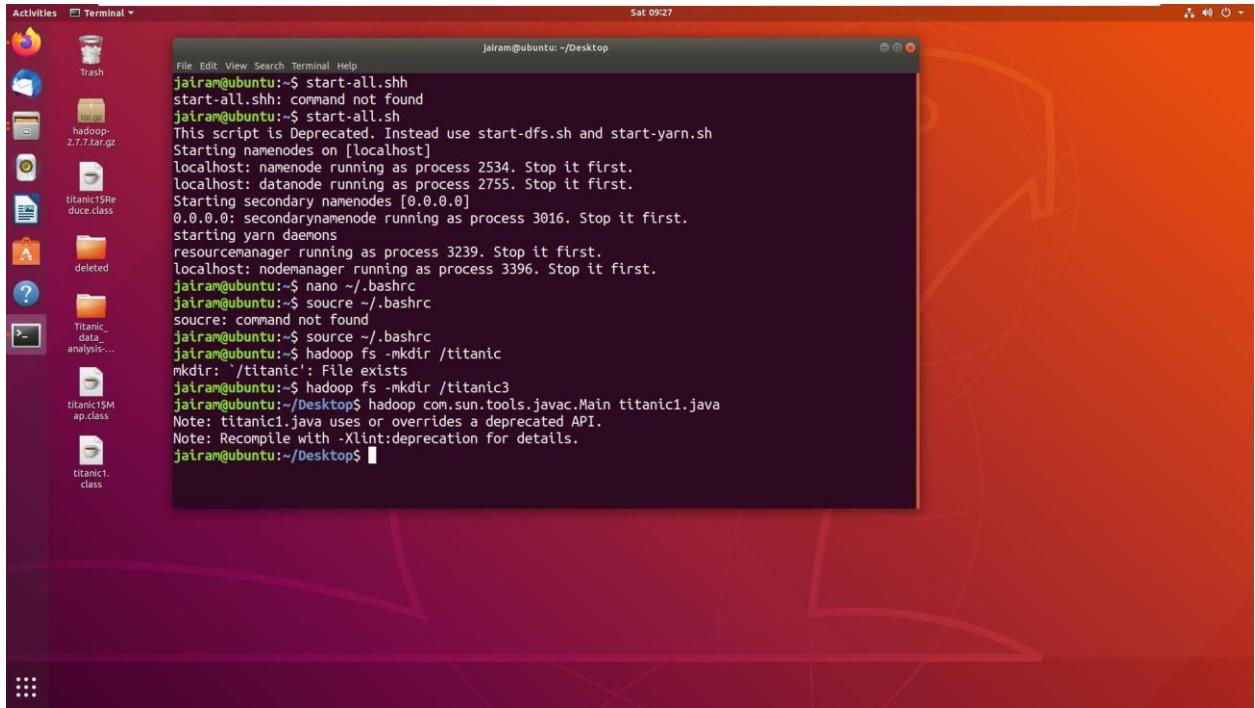
#### 5. command: *cd Desktop*



```
Activities Terminal Sat 09:27
jairam@ubuntu:~/Desktop
File Edit View Search Terminal Help
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ hadoop fs -mkdir /titanic
mkdir: /titanic: File exists
jairam@ubuntu:~$ hadoop fs -mkdir /titanic3
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic1.java
Note: titanic1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~$
```

#### 6. Now compile titanic1.java file suppose that file is on Desktop

Command: *hadoop com.sun.tools.javac.Main titanic1.java*

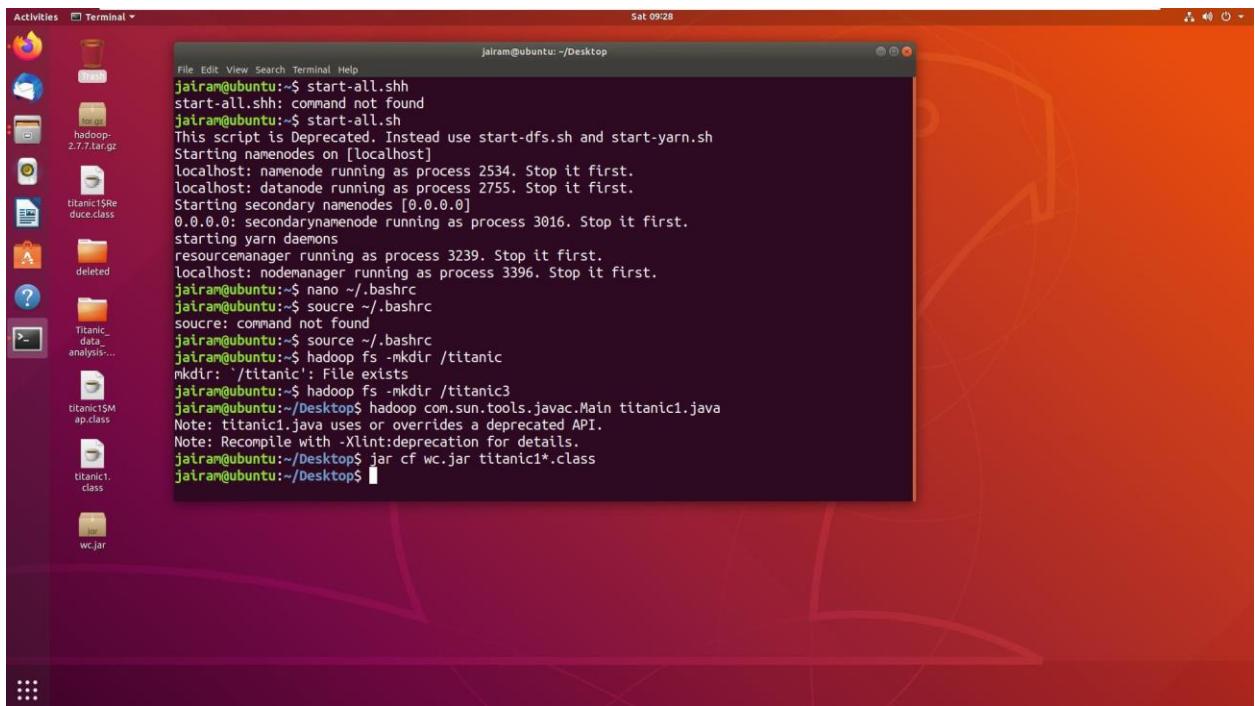


A screenshot of an Ubuntu desktop environment. On the left is a dock with icons for the Dash, Home, Activities, Terminal, and several application icons. A file browser window titled 'Activities' is open, showing a folder structure with files like 'hadoop-2.7.7.tar.gz', 'titanic1SRewe.class', 'deleted', 'Titanic\_data-analysis...', 'titanic1SMap.class', and 'titanic1.class'. In the center, a terminal window titled 'Terminal' is open, showing the following command-line session:

```
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ hadoop fs -mkdir /titanic
mkdir: '/titanic': File exists
jairam@ubuntu:~$ hadoop fs -mkdir /titanic3
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main titanic1.java
Note: titanic1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$
```

## 7. To combine all class files

Command: *jar cf wc.jar titanic1\*.class*

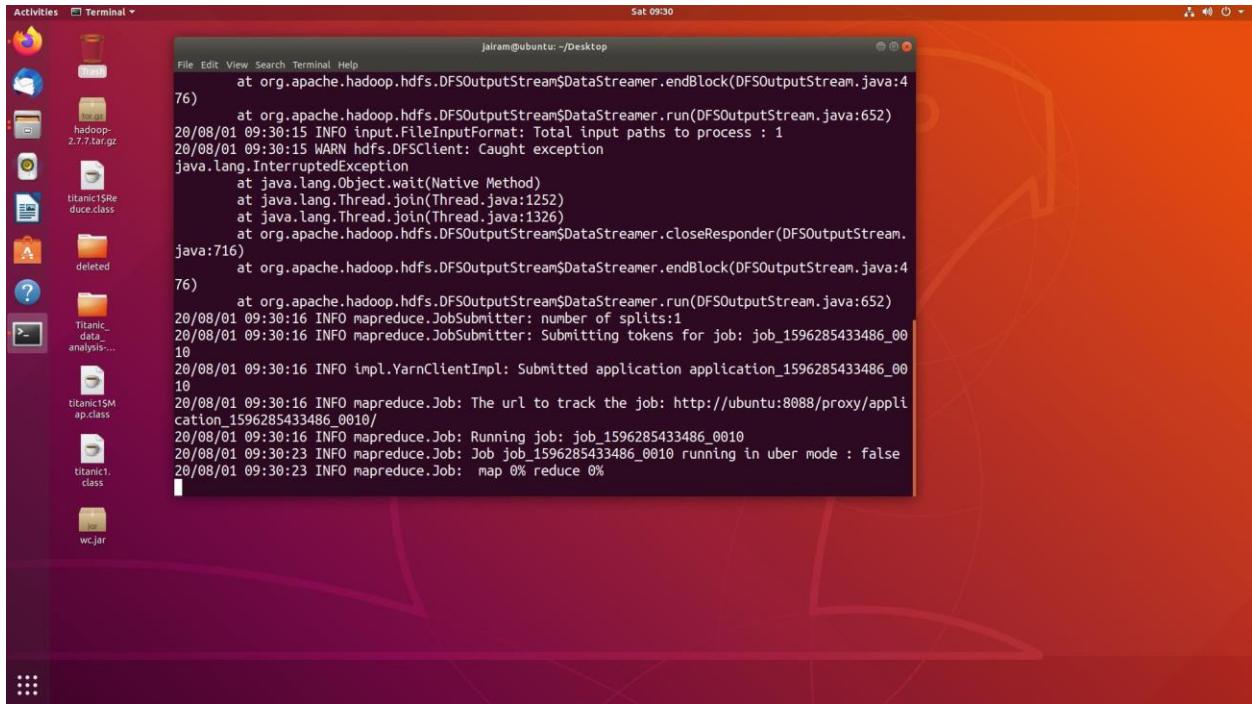


A screenshot of an Ubuntu desktop environment, identical to the one above. The terminal window shows the same command-line session, but the output ends with the command 'jar cf wc.jar titanic1\*.class' being entered:

```
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ hadoop fs -mkdir /titanic
mkdir: '/titanic': File exists
jairam@ubuntu:~$ hadoop fs -mkdir /titanic3
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main titanic1.java
Note: titanic1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$ jar cf wc.jar titanic1*.class
jairam@ubuntu:~/Desktop$
```

## 8. To execute

Command: ***hadoop jar wc.jar titanic1 /titanic/TitanicData.txt /average\_age***

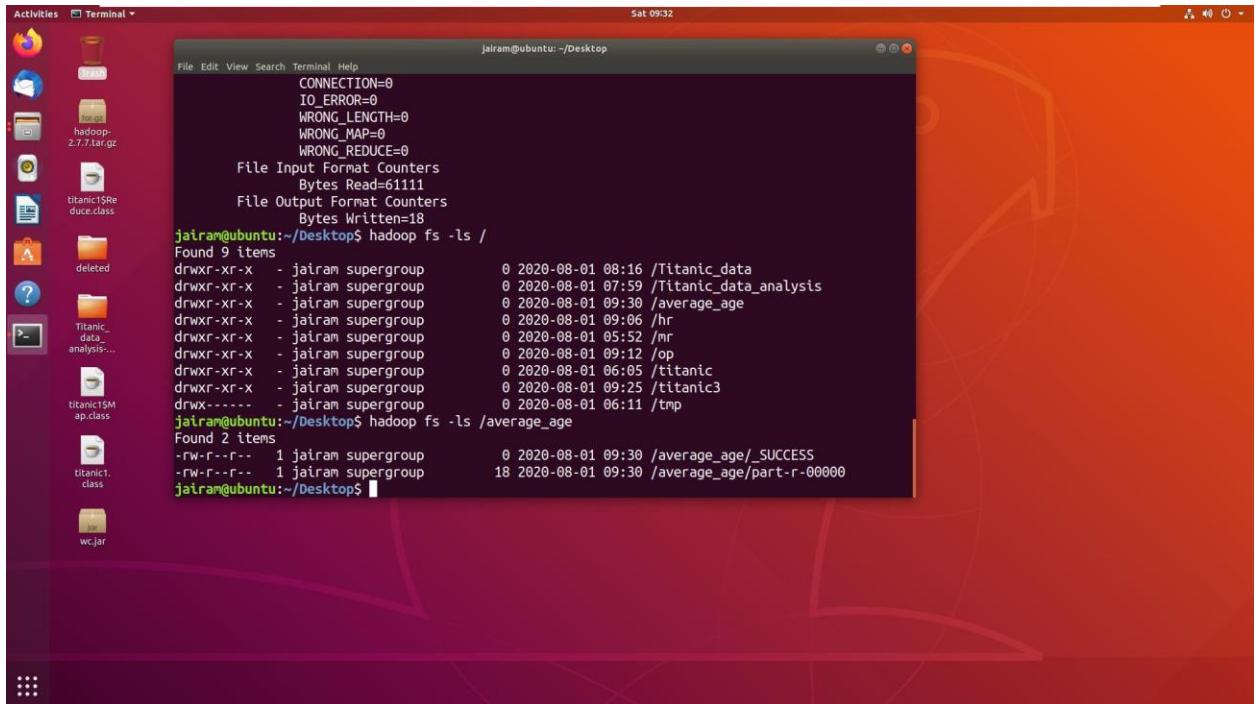


A screenshot of a Linux desktop environment. On the left is a dock with icons for a file manager, terminal, and other applications. The main area shows a terminal window titled 'Terminal' with the command 'hadoop jar wc.jar titanic1 /titanic/TitanicData.txt /average\_age'. The terminal output is as follows:

```
File Edit View Search Terminal Help
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:4
76)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:652)
20/08/01 09:30:15 INFO input.FileInputFormat: Total input paths to process : 1
20/08/01 09:30:15 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1252)
    at java.lang.Thread.join(Thread.java:1326)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.
java:716)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:4
76)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:652)
20/08/01 09:30:16 INFO mapreduce.JobSubmitter: number of splits:1
20/08/01 09:30:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1596285433486_00
10
20/08/01 09:30:16 INFO impl.YarnClientImpl: Submitted application application_1596285433486_00
10
20/08/01 09:30:16 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/appli
cation_1596285433486_0010/
20/08/01 09:30:16 INFO mapreduce.Job: Running job: job_1596285433486_0010
20/08/01 09:30:23 INFO mapreduce.Job: Job job_1596285433486_0010 running in uber mode : false
20/08/01 09:30:23 INFO mapreduce.Job: map 0% reduce 0%
```

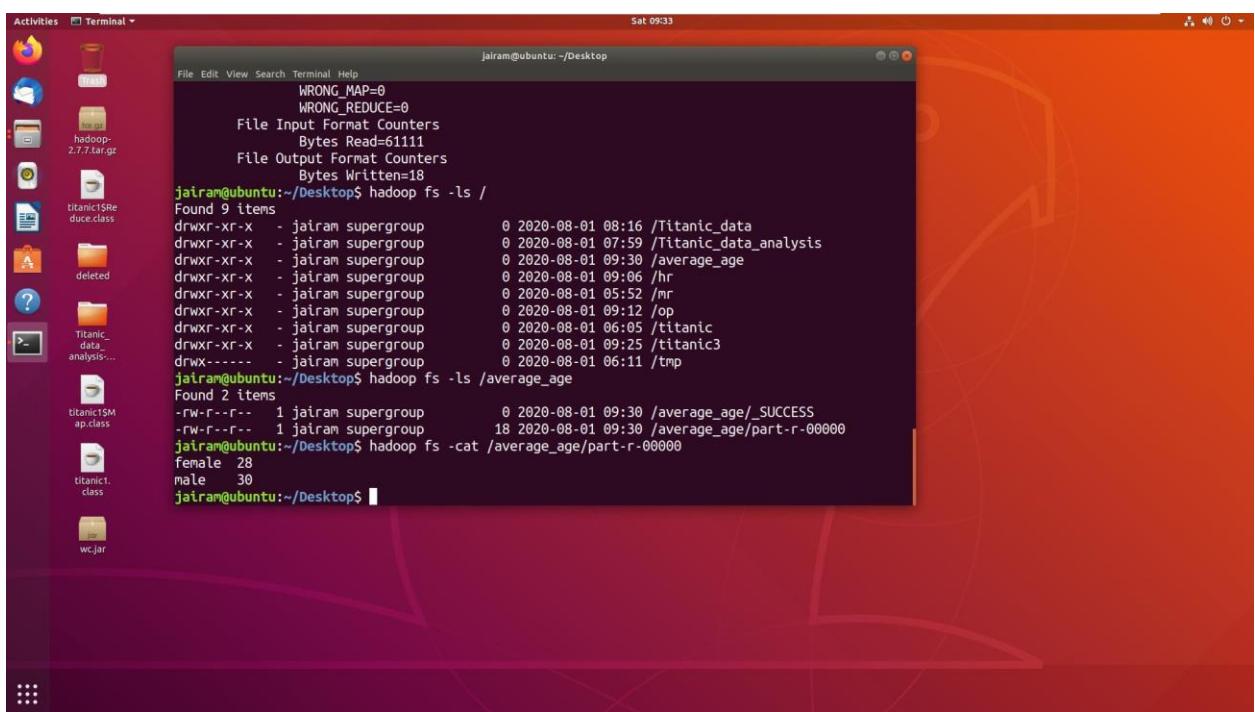
## 9. to check output folder is there or not

command: ***hadoop fs -ls /***



```
Activities Terminal Sat 09:32
jairam@ubuntu: ~/Desktop
File Edit View Search Terminal Help
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=6111
File Output Format Counters
Bytes Written=18
jairam@ubuntu:~/Desktop$ hadoop fs -ls /
Found 9 items
drwxr-xr-x  jairam supergroup 0 2020-08-01 08:16 /Titanic_data
drwxr-xr-x  jairam supergroup 0 2020-08-01 07:59 /Titanic_data_analysis
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:30 /average_age
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:06 /hr
drwxr-xr-x  jairam supergroup 0 2020-08-01 05:52 /mr
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:12 /op
drwxr-xr-x  jairam supergroup 0 2020-08-01 06:05 /titanic
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:25 /titanic3
drwx----- jairam supergroup 0 2020-08-01 06:11 /tmp
jairam@ubuntu:~/Desktop$ hadoop fs -ls /average_age
Found 2 items
-rw-r--r--  1 jairam supergroup 0 2020-08-01 09:30 /average_age/_SUCCESS
-rw-r--r--  1 jairam supergroup 18 2020-08-01 09:30 /average_age/part-r-00000
jairam@ubuntu:~/Desktop$
```

## 10. Command: *hadoop fs -cat /average\_age/part-r-00000*



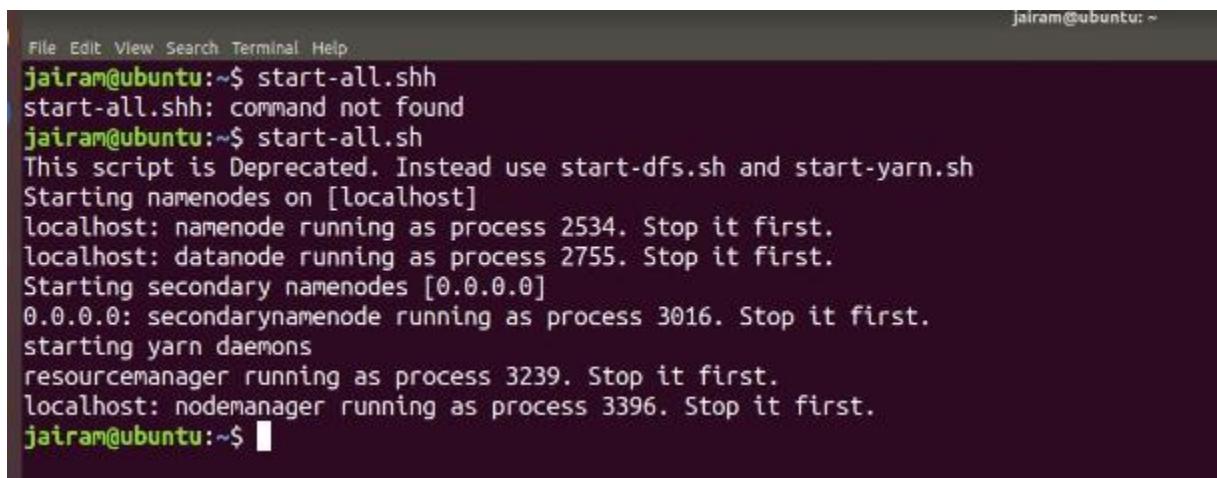
```
Activities Terminal Sat 09:33
jairam@ubuntu: ~/Desktop
File Edit View Search Terminal Help
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=6111
File Output Format Counters
Bytes Written=18
jairam@ubuntu:~/Desktop$ hadoop fs -ls /
Found 9 items
drwxr-xr-x  jairam supergroup 0 2020-08-01 08:16 /Titanic_data
drwxr-xr-x  jairam supergroup 0 2020-08-01 07:59 /Titanic_data_analysis
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:30 /average_age
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:06 /hr
drwxr-xr-x  jairam supergroup 0 2020-08-01 05:52 /mr
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:12 /op
drwxr-xr-x  jairam supergroup 0 2020-08-01 06:05 /titanic
drwxr-xr-x  jairam supergroup 0 2020-08-01 09:25 /titanic3
drwx----- jairam supergroup 0 2020-08-01 06:11 /tmp
jairam@ubuntu:~/Desktop$ hadoop fs -ls /average_age
Found 2 items
-rw-r--r--  1 jairam supergroup 0 2020-08-01 09:30 /average_age/_SUCCESS
-rw-r--r--  1 jairam supergroup 18 2020-08-01 09:30 /average_age/part-r-00000
jairam@ubuntu:~/Desktop$ hadoop fs -cat /average_age/part-r-00000
female 28
male 30
jairam@ubuntu:~/Desktop$
```

## Problem 2:

In this problem statement, we will find the number of people died or survived in each class with their genders and ages.

1. Open terminal and type

Command: ***start-all.sh***



```
File Edit View Search Terminal Help
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$
```

2. Open bashrc file by ***nano ~/.bashrc*** and type

```
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
```

3. Then type ***source ~/.bashrc***

```
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ █
```

Command: *hadoop fs -mkdir /titanic*

4. Command: *hadoop fs -put Desktop/TitanicData.txt /titanic*

```
jairam@ubuntu:~$ hadoop fs -put Desktop/TitanicData.txt /titanic
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic2.java
javac: file not found: titanic2.java
Usage: javac <options> <source_files>
use -help for a list of possible options
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic2.java
javac: file not found: titanic2.java
Usage: javac <options> <source_files>
use -help for a list of possible options
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic2.java
javac: file not found: titanic2.java
Usage: javac <options> <source_files>
use -help for a list of possible options
jairam@ubuntu:~$ cd Desktop
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main titanic2.java
Note: titanic2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$ █
```

5. Command: *cd Desktop*

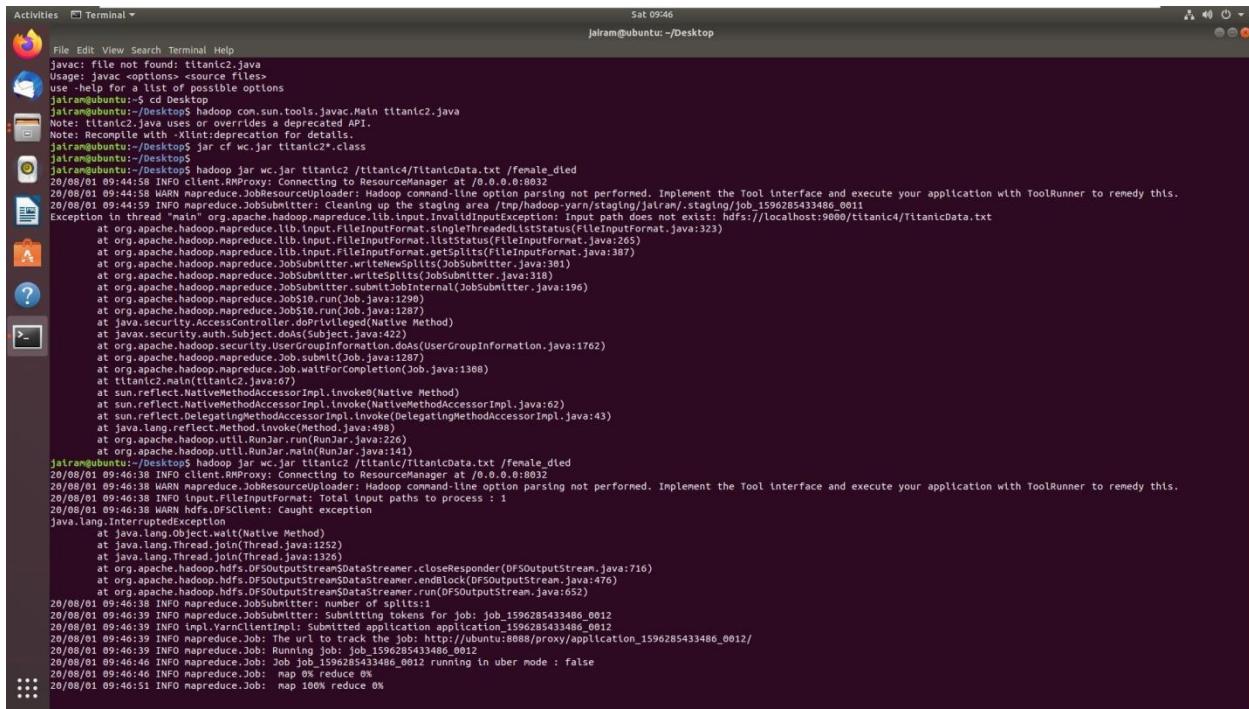
6. Now compile titanic2.java file suppose that file is on Desktop

Command: *hadoop com.sun.tools.javac.Main titanic2.java*

```
jairam@ubuntu:~$ hadoop fs -put Desktop/TitanicData.txt /titanic
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic2.java
javac: file not found: titanic2.java
Usage: javac <options> <source files>
use -help for a list of possible options
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic2.java
javac: file not found: titanic2.java
Usage: javac <options> <source files>
use -help for a list of possible options
jairam@ubuntu:~$ hadoop com.sun.tools.javac.Main titanic2.java
javac: file not found: titanic2.java
Usage: javac <options> <source files>
use -help for a list of possible options
jairam@ubuntu:~$ cd Desktop
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main titanic2.java
Note: titanic2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$
```

## 7. To combine all class files

Command: ***jar cf wc.jar titanic2\*.class***



```
Activities Terminal Sat 09:46 jairam@ubuntu: ~/Desktop
File Edit View Search Terminal Help
javac: file not found: titanic2.java
Usage: javac <options> <source files>
use -help for a list of possible options
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main titanic2.java
Note: titanic2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$ jar cf wc.jar titanic2*.class
jairam@ubuntu:~/Desktop$ hadoop jar wc.jar titanic2 /titanic4/TitanicData.txt /female_died
20/08/01 09:44:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8082
20/08/01 09:44:59 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/jairam-staging/job_1596285433486_0011
Exception in thread "main" org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://localhost:9000/titanic4/TitanicData.txt
at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:323)
at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:387)
at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:301)
at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:318)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:196)
at org.apache.hadoop.mapreduce.Job$1.run(Job.java:290)
at org.apache.hadoop.mapreduce.Job$1.run(Job.java:287)
at java.security.AccessController.doPrivileged(PrivilegedMethod)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1762)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1287)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1308)
at titanic2.main(titanic2.java:12)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:226)
at org.apache.hadoop.util.RunJar.main(RunJar.java:211)
jairam@ubuntu:~/Desktop$ hadoop jar wc.jar titanic2 /titanic4/TitanicData.txt /female_died
20/08/01 09:46:38 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8082
20/08/01 09:46:38 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/08/01 09:46:38 INFO InputFormat: Total input paths to process : 1
20/08/01 09:46:38 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
at java.lang.Object.wait(Native Method)
at java.lang.Thread.join(Thread.java:1252)
at java.lang.Thread.join(Thread.java:1320)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:716)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:476)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.close(DFSOutputStream.java:652)
20/08/01 09:46:38 INFO mapreduce.JobSubmitter: Number of splits:1
20/08/01 09:46:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1596285433486_0012
20/08/01 09:46:39 INFO impl.YarnClientImpl: Submitted application application_1596285433486_0012
20/08/01 09:46:39 INFO mapreduce.Job: Running job: http://ubuntu:8088/proxy/application_1596285433486_0012
20/08/01 09:46:46 INFO mapreduce.Job: Job job_1596285433486_0012 running in uber mode : false
20/08/01 09:46:46 INFO mapreduce.Job: map 0% reduce 0%
20/08/01 09:46:51 INFO mapreduce.Job: map 100% reduce 0%
```

## 8. To execute

Command: ***hadoop jar wc.jar titanic2 /titanic/TitanicData.txt /female\_died***

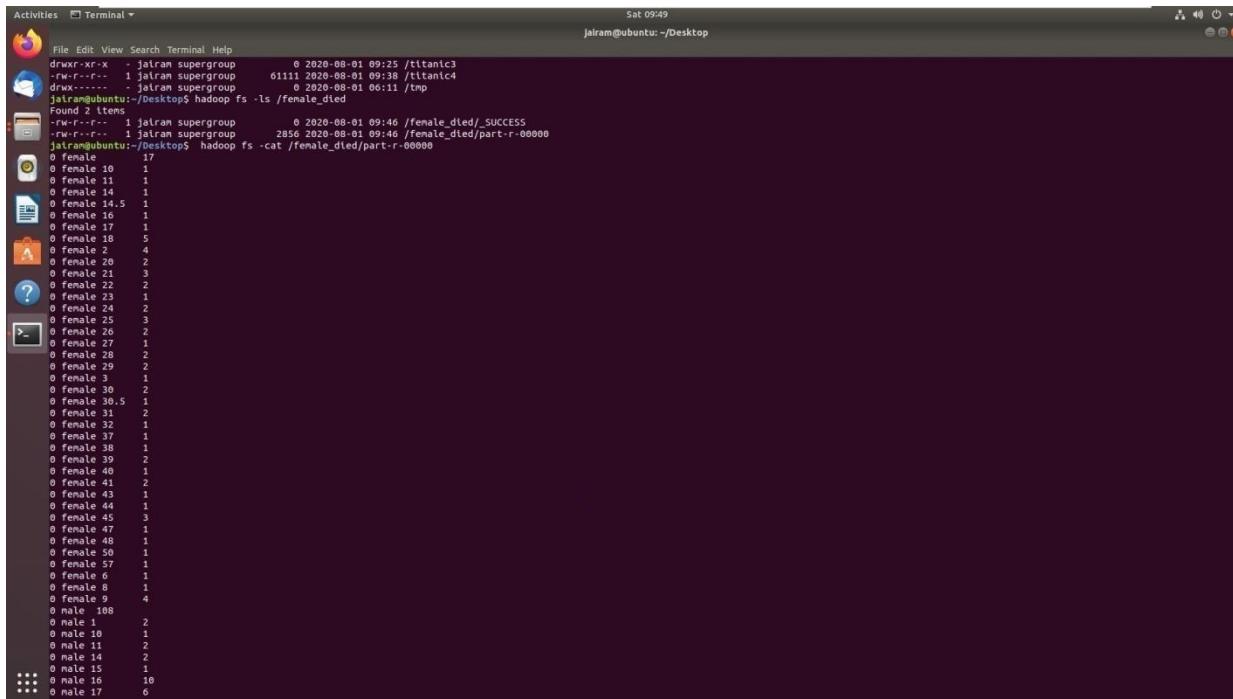
```
Activities Terminal Sat 09:46
jalram@ubuntu:~/Desktop
File Edit View Search Terminal Help
javac: file not found: titanic2.java
Usage: javac <options> <source files>
use -help for a list of possible options
jalram@ubuntu:~$ cd Desktop
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main titanic2.java
Note: titanic2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jalram@ubuntu:~/Desktop$ jar cf wc.jar titanic2*.class
jalram@ubuntu:~/Desktop$ hadoop jar wc.jar titanic2 /titanic4/TitanicData.txt /female_died
20/08/01 09:44:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8082
20/08/01 09:44:58 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Exception in thread "main" org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://localhost:9000/titanic4/TitanicData.txt
at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:323)
at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:265)
at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:387)
at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:301)
at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:318)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:196)
at org.apache.hadoop.mapreduce.Job$1.run(Job.java:129)
at org.apache.hadoop.mapreduce.Job$1.run(Job.java:1287)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1762)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1287)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1308)
at titanic2.main(titanic2.java:67)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:492)
at org.apache.hadoop.util.RunJar.run(RunJar.java:226)
at org.apache.hadoop.util.RunJar.main(RunJar.java:141)
jalram@ubuntu:~/Desktop$ hadoop jar wc.jar titanic2 /titanic4/TitanicData.txt /female_died
20/08/01 09:46:38 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8082
20/08/01 09:46:38 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/08/01 09:46:38 org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input paths to process : 1
20/08/01 09:46:38 WARN hdfs.DfsClient: Caught exception
java.lang.InterruptedIOException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1252)
    at java.lang.Thread.join(Thread.java:1326)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:716)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:476)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:592)
20/08/01 09:46:38 INFO mapreduce.JobSubmitter: number of splits:
20/08/01 09:46:39 INFO mapreduce.JobSubmitter: Submitting token for job: job_1596285433486_0012
20/08/01 09:46:39 INFO impl.YarnClientImpl: Submitted application application_1596285433486_0012
20/08/01 09:46:39 INFO mapreduce.Job: Job running user: jalram
20/08/01 09:46:39 INFO mapreduce.Job: Running job: job_1596285433486_0012
20/08/01 09:46:44 INFO mapreduce.Job: Job map 100% reduce 0%
20/08/01 09:46:51 INFO mapreduce.Job: map 100% reduce 0%
```

9. to check output folder is there or not

Command: *hadoop fs -ls /*

```
Activities Terminal Sat 09:47
joram@ubuntu: ~/Desktop
File Edit View Search Terminal Help
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=3039
Total time spent by all reduce tasks in occupied slots (ms)=2922
Total time spent by all map tasks (ms)=3039
Total time spent by all reduce tasks (ms)=2922
Total vcore-milliseconds taken by all map tasks=3039
Total vcore-milliseconds taken by all reduce tasks=2922
Total negabyte-milliseconds taken by all map tasks=3111936
Total negabyte-milliseconds taken by all reduce tasks=2992128
Map-Reduce Framework
  Map input records=891
  Map output records=891
  Map output bytes=12743
  Map output materialized bytes=14531
  Map output blocks=1
  Combined Input=891
  Combine input records=0
  Combine output records=0
  Reduce input groups=226
  Reduce shuffle bytes=14531
  Reduce input records=891
  Reduce output records=226
  Spilled Records=1782
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=93
  CPU time spent (ms)=0
  Physical memory (bytes) snapshot=306671616
  Virtual memory (bytes) snapshot=3772000128
  Total committed heap usage (bytes)=170004480
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=111
File Output Format Counters
  Bytes Written=2856
joram@ubuntu: ~/Desktop$ hadoop fs -ls /
Found 11 items
drwxr-x--x  - joram supergroup  0 2020-08-01 08:16 /titanic_data
drwxr-x--x  - joram supergroup  0 2020-08-01 09:59 /titanic_data_analysis
drwxr-x--x  - joram supergroup  0 2020-08-01 09:46 /titanic_disease
drwxr-x--x  - joram supergroup  0 2020-08-01 09:46 /female_died
drwxr-x--x  - joram supergroup  0 2020-08-01 09:06 /hr
drwxr-x--x  - joram supergroup  0 2020-08-01 05:52 /mr
drwxr-x--x  - joram supergroup  0 2020-08-01 09:12 /op
drwxr-x--x  - joram supergroup  0 2020-08-01 09:46 /titanic
drwxr-x--x  - joram supergroup  0 2020-08-01 09:25 /titanic3
drwxr----  1 joram supergroup  61111 2020-08-01 09:38 /titanic4
drwxr----  - joram supergroup  0 2020-08-01 06:11 /tmp
joram@ubuntu: ~/Desktop$
```

## 10.command: *hadoop fs -cat /female\_died/part-r-00000*



```
Activities Terminal Sat 09:49
jairam@ubuntu:~/Desktop$ hadoop fs -ls /female_died
Found 2 items
drwxr-xr-x  - jairam supergroup  0 2020-08-01 09:25 /titanic3
-rw-r--r--  1 jairam supergroup  01111 2020-08-01 09:38 /titanic4
drwx-----  - jairam supergroup  0 2020-08-01 06:11 /tmp
jairam@ubuntu:~/Desktop$ hadoop fs -cat /female_died/_SUCCESS
2856 2020-08-01 09:46 /female_died/part-r-00000
jairam@ubuntu:~/Desktop$ hadoop fs -cat /female_died/part-r-00000
0 female
0 female 1
0 female 10
0 female 11
0 female 14
0 female 14.5
0 female 16
0 female 17
0 female 18
0 female 2
0 female 20
0 female 21
0 female 22
0 female 23
0 female 24
0 female 25
0 female 26
0 female 27
0 female 28
0 female 29
0 female 3
0 female 30
0 female 30.5
0 female 31
0 female 32
0 female 37
0 female 38
0 female 39
0 female 40
0 female 41
0 female 43
0 female 44
0 female 45
0 female 47
0 female 48
0 female 50
0 female 57
0 female 6
0 female 8
0 male 9
0 male 108
0 male 1
0 male 10
0 male 11
0 male 14
0 male 15
0 male 16
0 male 17
```

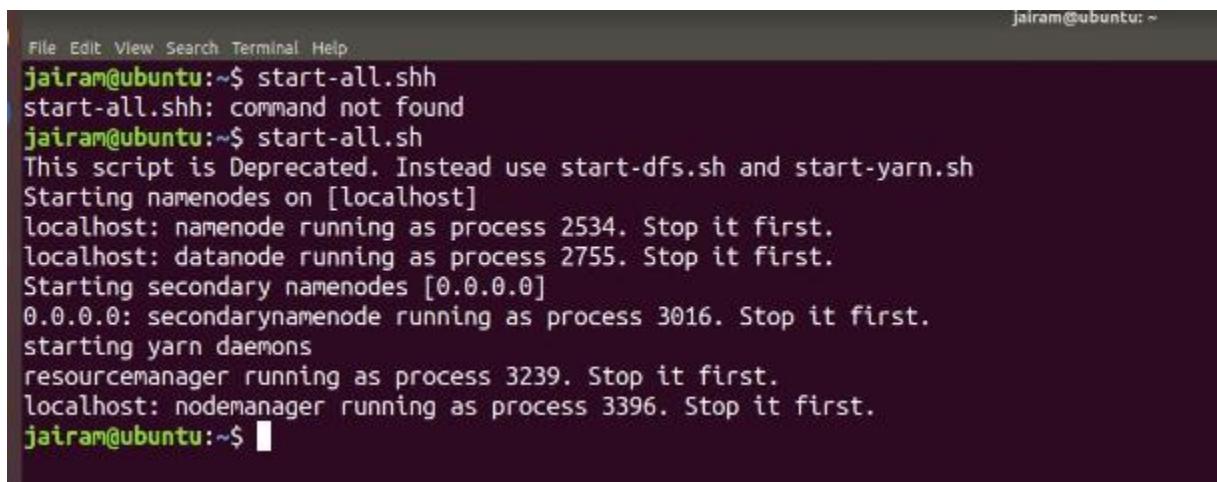
## Uber data analysis

Problem 1:

In this problem statement, we will find the days on which each basement has more trips.

1. Open terminal and type

Command: ***start-all.sh***



A screenshot of a terminal window titled "jaliram@ubuntu: ~". The window shows the command "start-all.sh" being run, followed by a series of messages indicating the startup of various Hadoop components like namenodes, datanodes, secondary namenodes, and yarn daemons. The terminal is dark-themed with white text.

```
File Edit View Search Terminal Help
jaliram@ubuntu:~$ start-all.shh
start-all.shh: command not found
jaliram@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jaliram@ubuntu:~$
```

2. Open bashrc file by ***nano ~/.bashrc*** and type

```
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
```

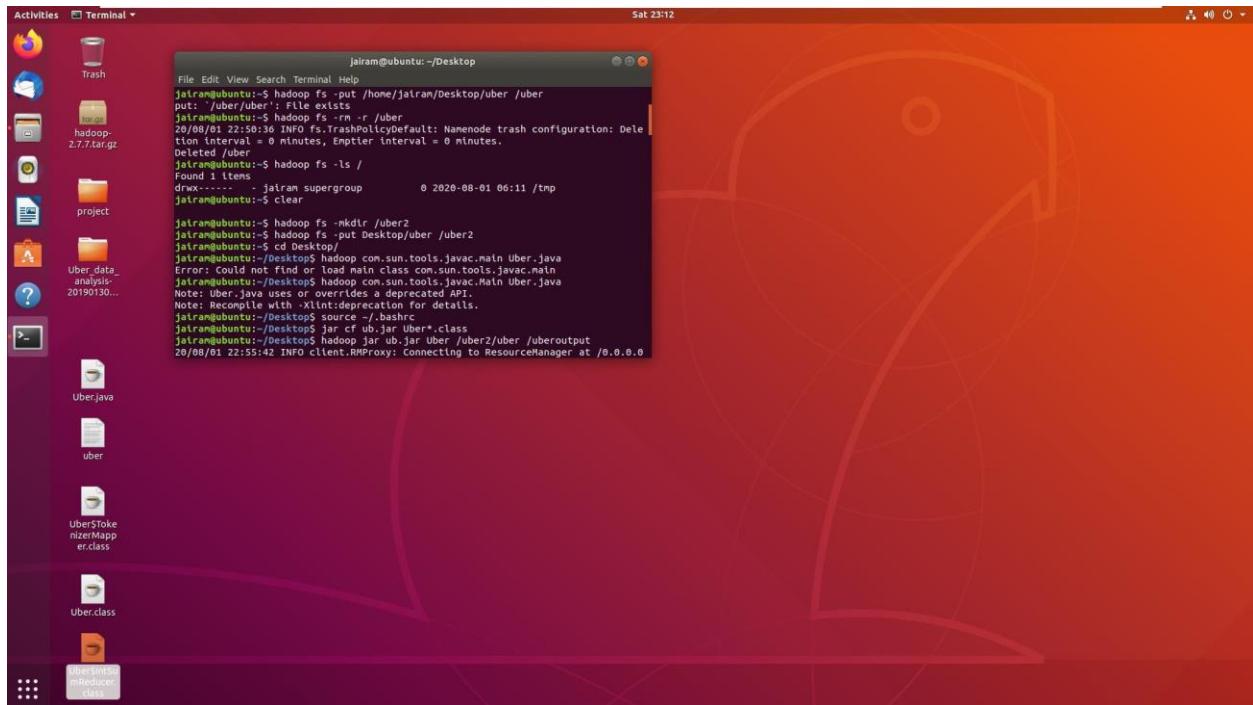
3. Then ***source ~/.bashrc***

```

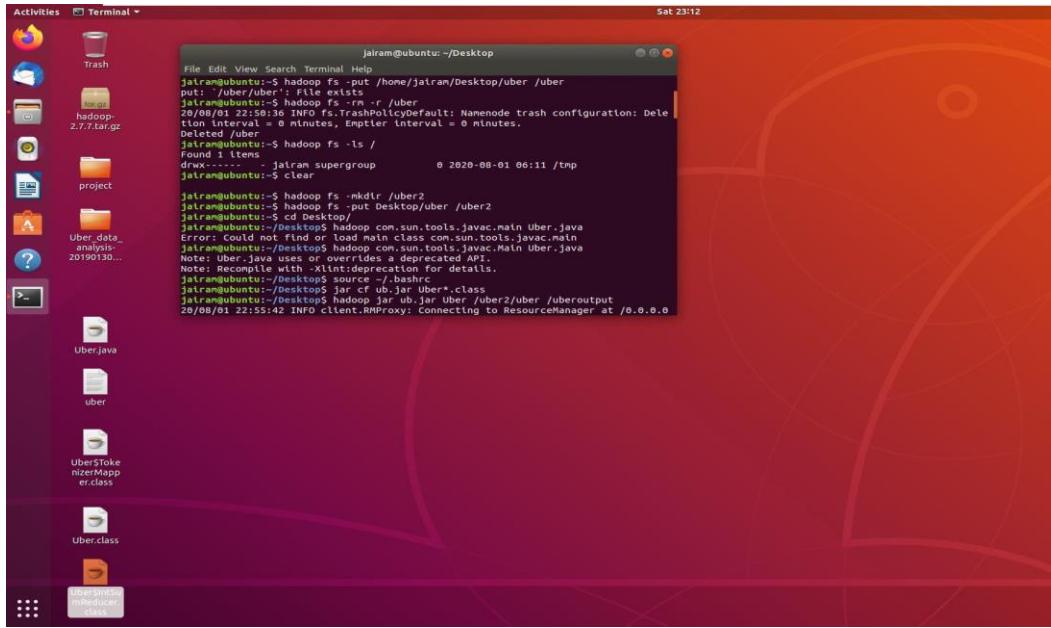
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ soucre ~/.bashrc
soucre: command not found
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ 

```

#### 4. command: *hadoop fs -mkdir /uber2*



#### 5. command: *hadoop fs -put Desktop/Uber.txt /uber2*



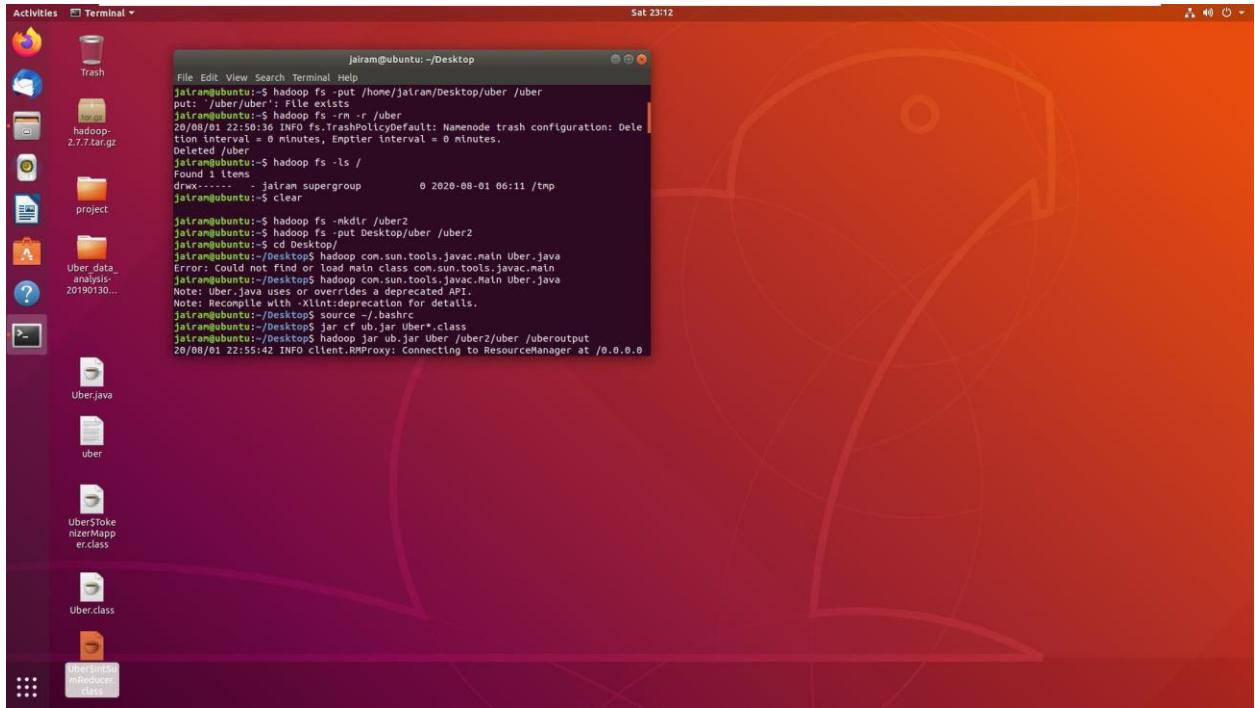
A screenshot of an Ubuntu desktop environment. In the top left corner, there's an 'Activities' button and a 'Terminal' window. The terminal window shows a session with the user 'jalram' at 'jalram@ubuntu'. The user is executing Hadoop commands to manage files in the '/uber' directory. The desktop background is orange, and various icons are visible in the dock, including a Firefox icon, a trash icon, and several application icons like LibreOffice and a file manager.

```
jalram@ubuntu:~$ hadoop fs -put /home/jalram/Desktop/uber /uber
 jalram@ubuntu:~$ hadoop fs -rm -r /uber
20/08/01 22:50:36 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion grace period = 6 minutes, Emptier interval = 6 minutes.
Deleted /uber
jalram@ubuntu:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x 2 jalram supergroup      0 2020-08-01 06:11 /tmp
jalram@ubuntu:~$ hadoop fs -mkdir /uber2
jalram@ubuntu:~$ hadoop fs -put Desktop/uber /uber2
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main Uber.java
Error: Could not find or load main class com.sun.tools.javac.Main
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main Uber.java
Note: Uber.java uses or overrides a deprecated API.
      Note: Recompile with -Xlint:deprecation for details.
jalram@ubuntu:~/Desktop$ jar cf ub.jar Uber*.class
jalram@ubuntu:~/Desktop$ hadoop jar ub.jar Uber /uber2/uber /uberoutput
20/08/01 22:55:42 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
```

6. command: ***cd Desktop***

7. Now compile titanic1.java file suppose that file is on Desktop

Command: ***hadoop com.sun.tools.javac.Main uber1.java***

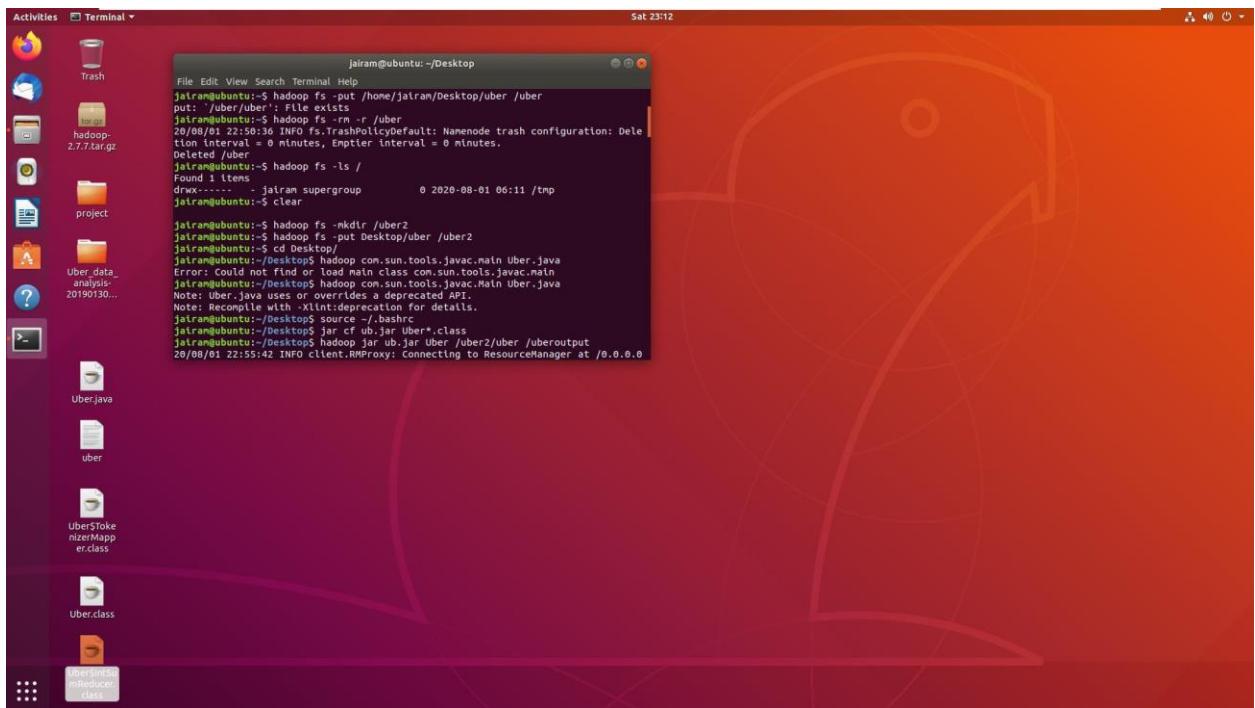


```
File Edit View Search Terminal Help
jalram@ubuntu:~/Desktop
jalram@ubuntu:~$ hadoop fs -put /home/jalram/Desktop/uber /uber
put: '/uber/uber': File exists
jalram@ubuntu:~$ hadoop fs -rm -r /uber
20/08/01 22:50:36 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /uber
Deleted /uber
jalram@ubuntu:~$ hadoop fs -ls /
Found 1 items
drwx----- jalram supergroup 0 2020-08-01 06:11 /tmp
jalram@ubuntu:~$ clear

jalram@ubuntu:~$ hadoop fs -mkdir /uber2
jalram@ubuntu:~$ hadoop fs -put Desktop/uber /uber2
jalram@ubuntu:~$ cd Desktop/
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.main Uber.java
Error: Could not find or load main class com.sun.tools.javac.main
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main Uber.java
Note: Uber.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jalram@ubuntu:~/Desktop$ source ~/.bashrc
jalram@ubuntu:~/Desktop$ jar cf ub.jar Uber*.class
jalram@ubuntu:~/Desktop$ hadoop jar ub.jar Uber /uber2/uber /uberoutput
20/08/01 22:55:42 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
```

## 8. To combine all class files

Command: *jar cf wc.jar uber1\*.class*



```
File Edit View Search Terminal Help
jalram@ubuntu:~/Desktop
jalram@ubuntu:~$ hadoop fs -put /home/jalram/Desktop/uber /uber
put: '/uber/uber': File exists
jalram@ubuntu:~$ hadoop fs -rm -r /uber
20/08/01 22:50:36 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /uber
Deleted /uber
jalram@ubuntu:~$ hadoop fs -ls /
Found 1 items
drwx----- jalram supergroup 0 2020-08-01 06:11 /tmp
jalram@ubuntu:~$ clear

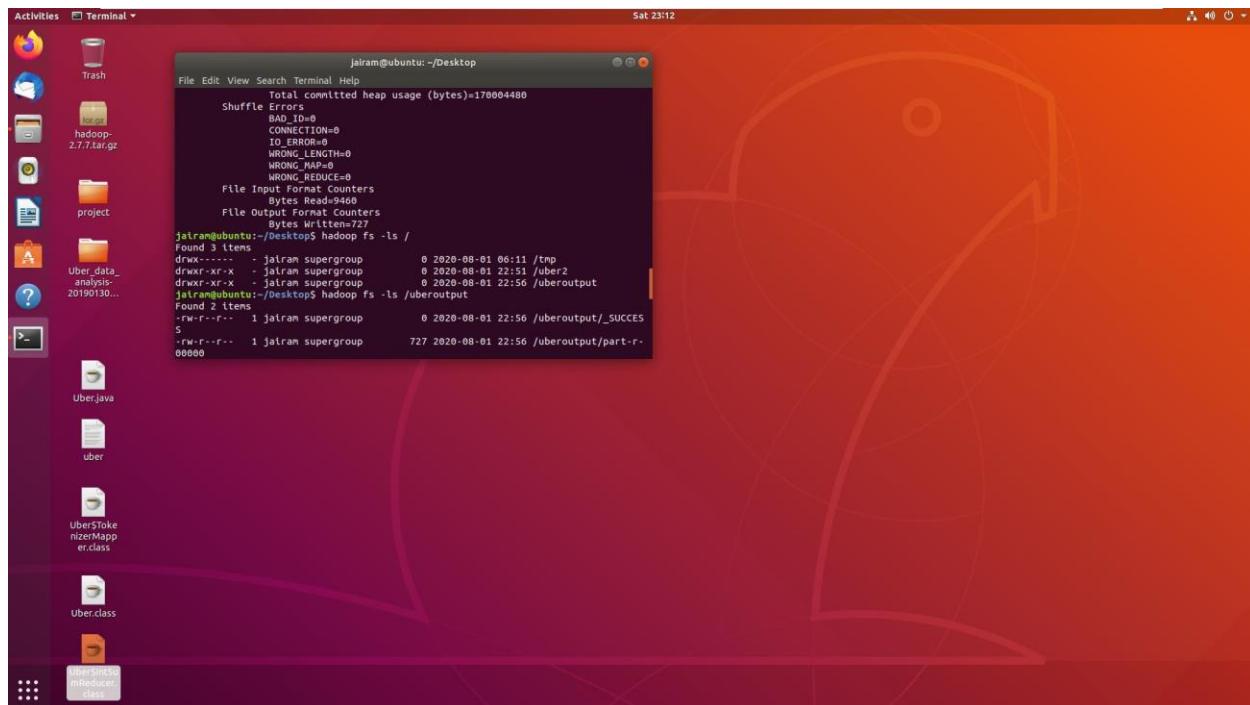
jalram@ubuntu:~$ hadoop fs -mkdir /uber2
jalram@ubuntu:~$ hadoop fs -put Desktop/uber /uber2
jalram@ubuntu:~$ cd Desktop/
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.main Uber.java
Error: Could not find or load main class com.sun.tools.javac.main
jalram@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main Uber.java
Note: Uber.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jalram@ubuntu:~/Desktop$ source ~/.bashrc
jalram@ubuntu:~/Desktop$ jar cf ub.jar Uber*.class
jalram@ubuntu:~/Desktop$ hadoop jar ub.jar Uber /uber2/uber /uberoutput
20/08/01 22:55:42 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
```

## 9. To execute

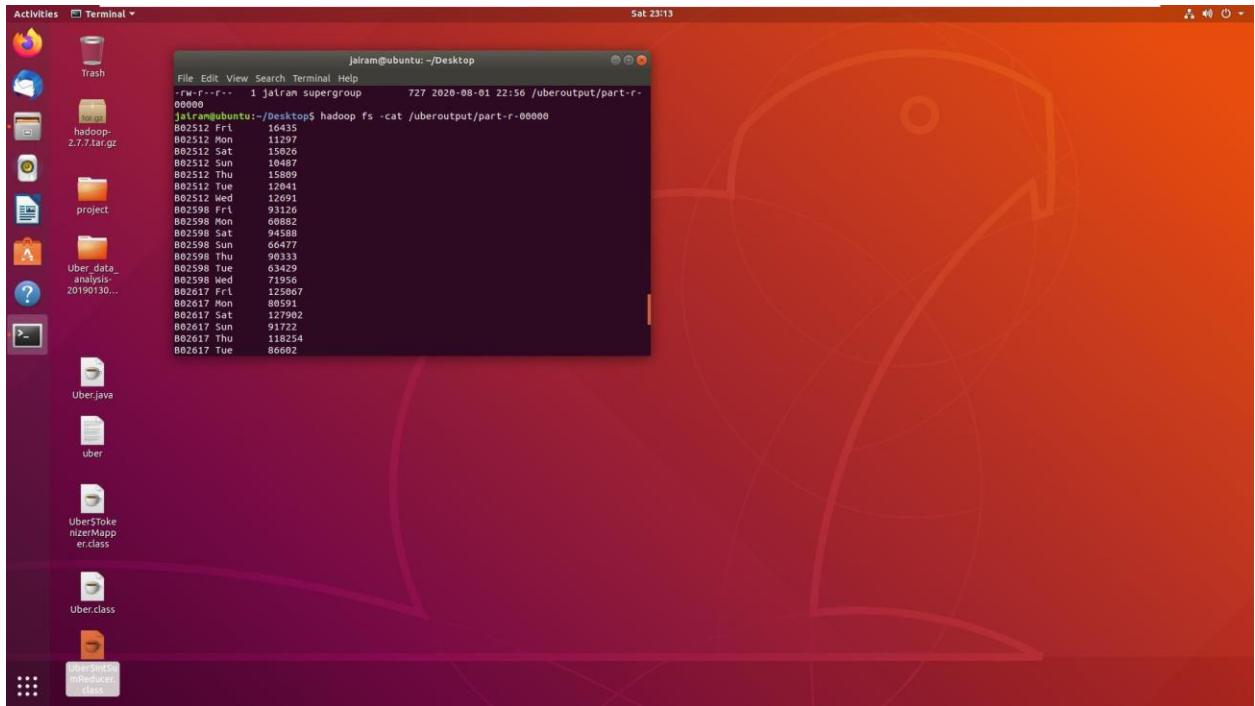
Command: ***hadoop jar wc.jar uber1 /uber2/Uber.txt /uberoutput***

11.to check output folder is there or not

Command: ***hadoop fs -ls /***



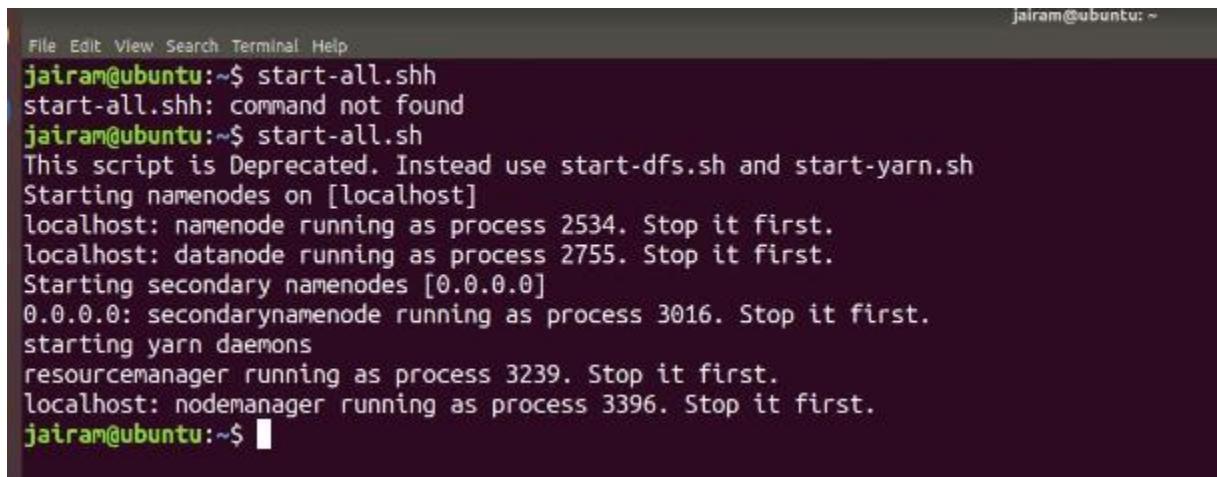
11.command: ***hadoop fs -cat /uberoutput/part-r-00000***



## Problem 2:

In this problem statement, we will find the days on which each basement has more number of active vehicles.

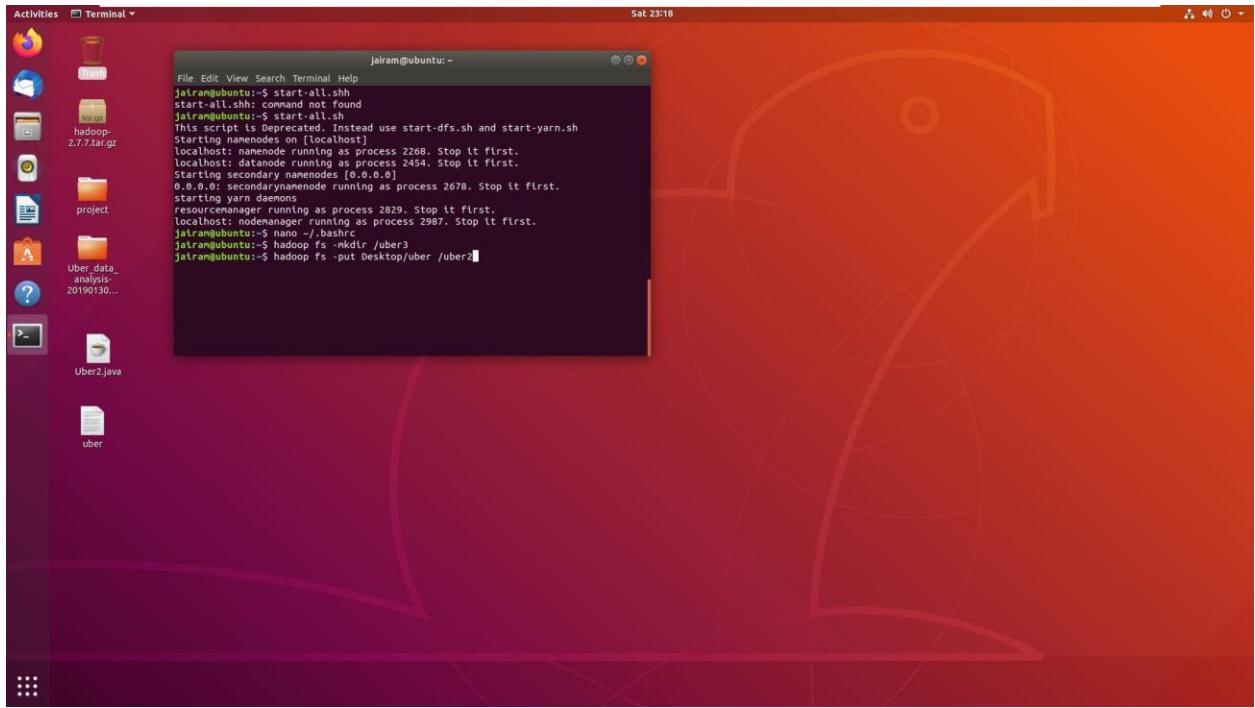
1. Open terminal and type Command: ***start-all.sh***



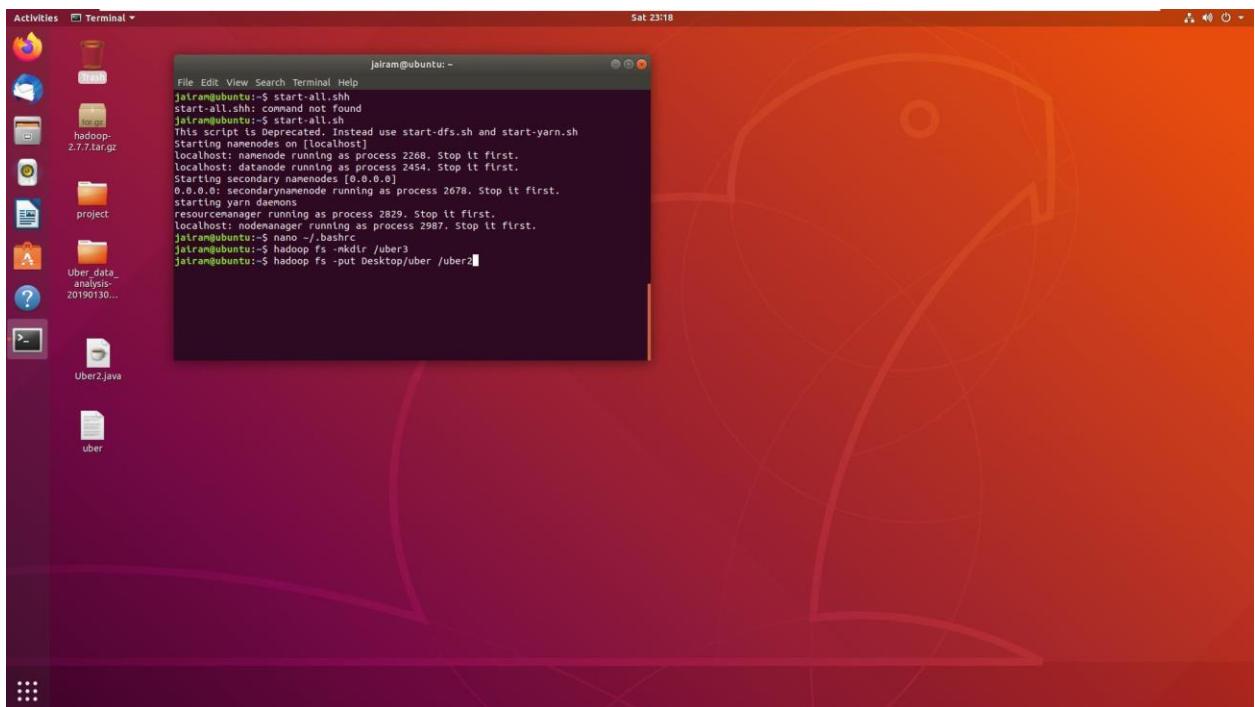
```
File Edit View Search Terminal Help
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$
```

2. Open bashrc file by ***nano ~/.bashrc*** and type

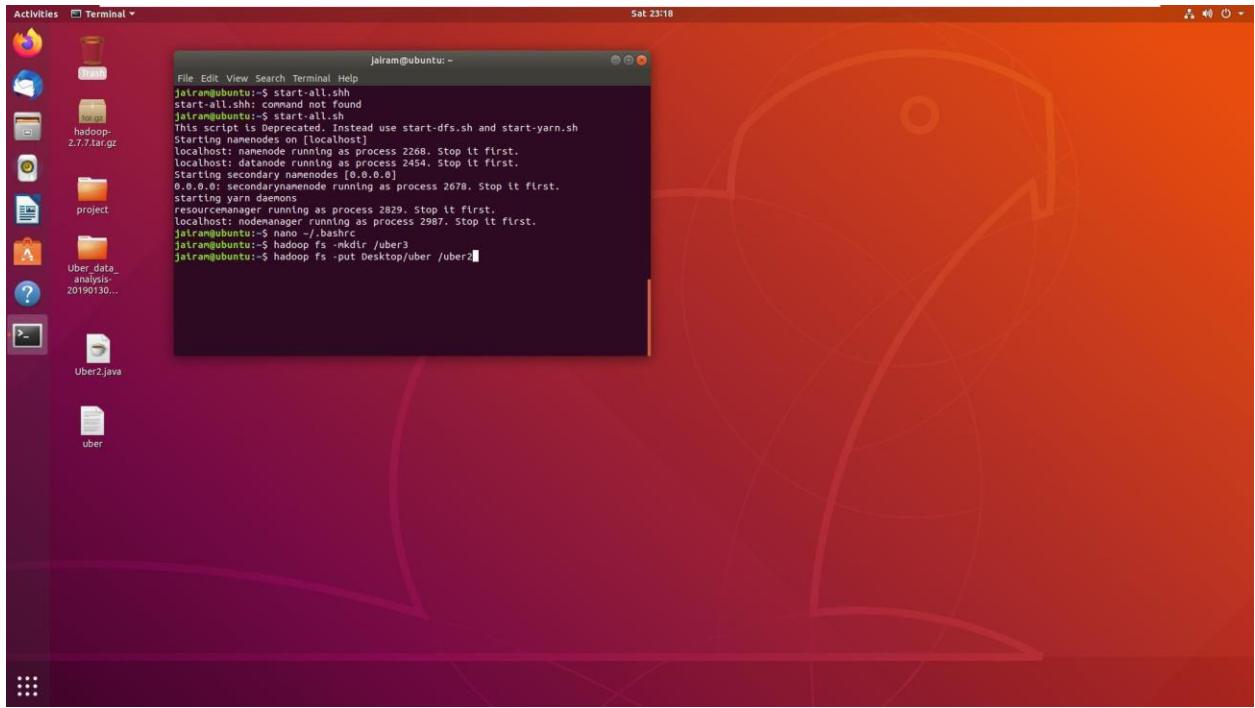
```
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
```



### 3. Command: *hadoop fs -mkdir /uber3*



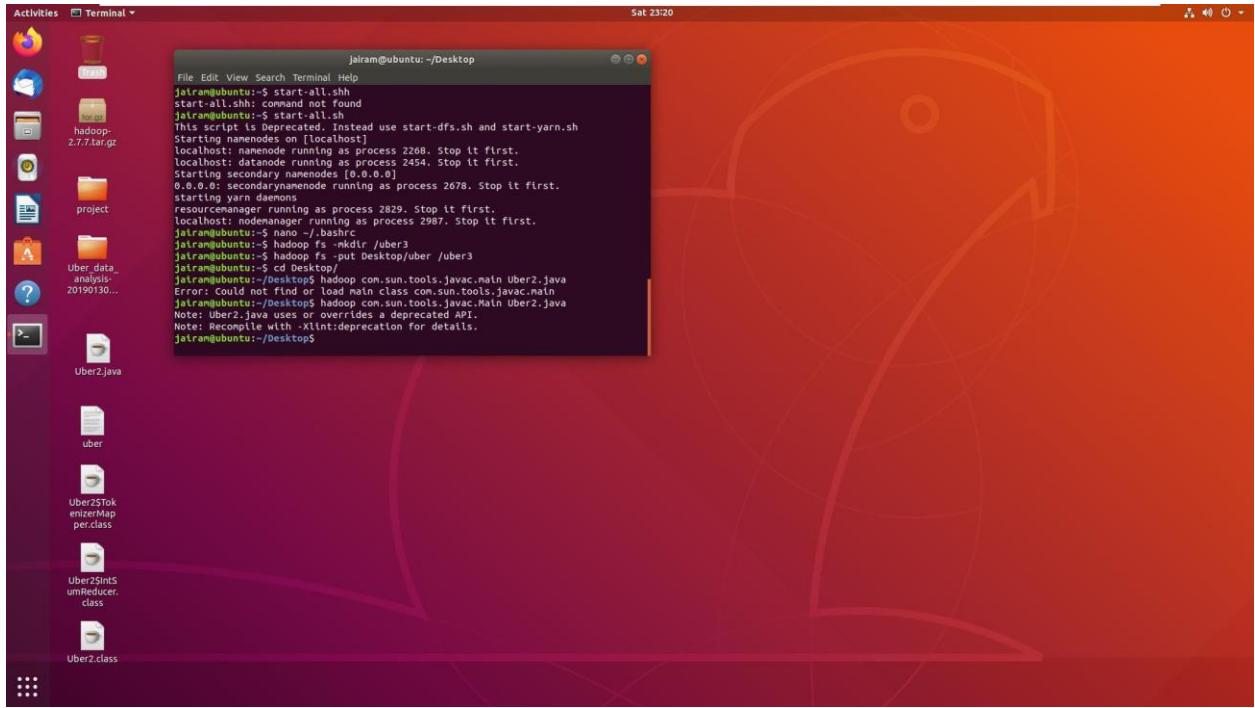
### 4. Command: *hadoop fs -put Desktop/Uber.txt /uber3*



5. Command: ***cd Desktop***

6. Now compile titanic1.java file suppose that file is on Desktop

***hadoop com.sun.tools.javac.Main uber2.java***



7. Then ***source ~/.bashrc***
8. To combine all class files

***jar cf wc.jar uber2\*.class***

A screenshot of a Linux desktop environment, likely Ubuntu, featuring a Unity interface. The desktop background is orange and features a large, stylized white owl logo. In the top left corner, there's a dock with icons for various applications. On the right side, there's a vertical panel with more icons. A terminal window is open in the top center, showing a command-line session as follows:

```
jairam@ubuntu:~/Desktop
File Edit View Search Terminal Help
jairam@ubuntu:~$ start-all.sh
This script is deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on localhost...
localhost: datanode running as process 2268. Stop it first.
localhost: datanode running as process 2454. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 2678. Stop it first.
starting yarn daemons
resource manager running as process 2829. Stop it first.
localhost: resourcemanager running as process 2987. Stop it first.
jairam@ubuntu:~$ nano ./bashrc
jairam@ubuntu:~$ hadoop fs -mkdirr /uber3
jairam@ubuntu:~$ cd Desktop
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main Uber2.java
Error: Could not find or load main class com.sun.tools.javac.Main
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main Uber2.java
Note: Uber2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$ source ./bashrc
bash: /home/jairam/.bashrc: No such file or directory
jairam@ubuntu:~/Desktop$ source ./bashrc
jairam@ubuntu:~/Desktop$
```

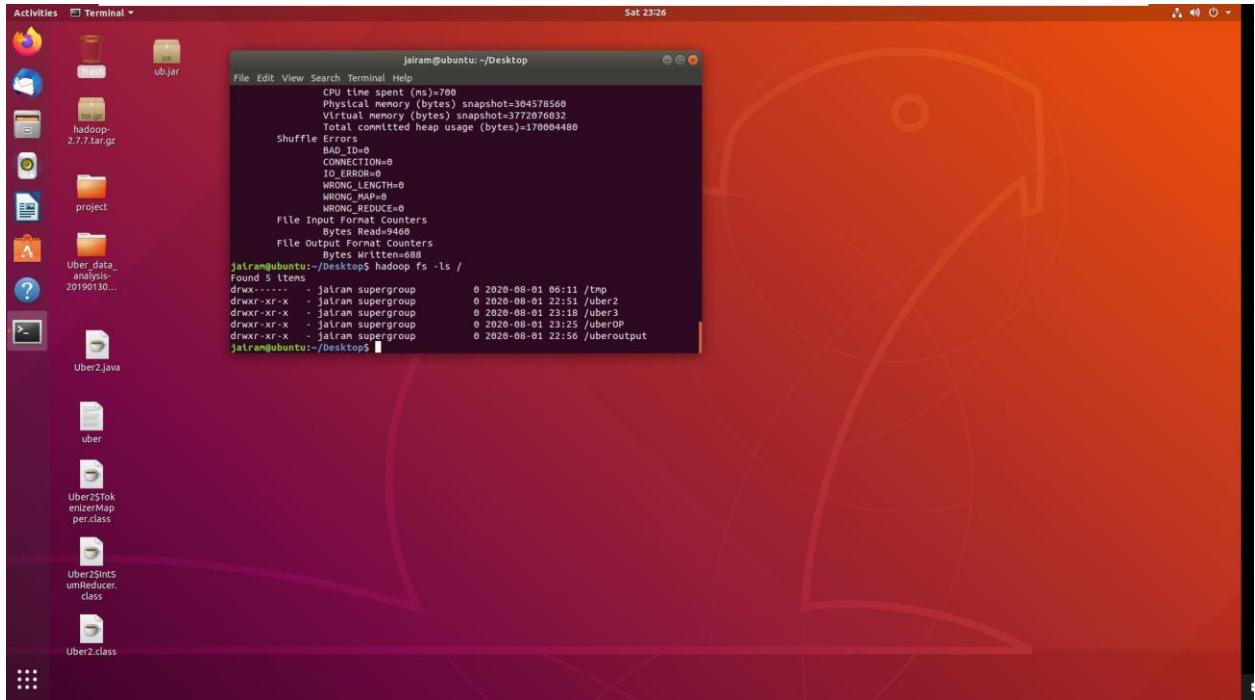
The terminal window has a dark background with light-colored text. The user's home directory is ~/Desktop.

### 9. To execute

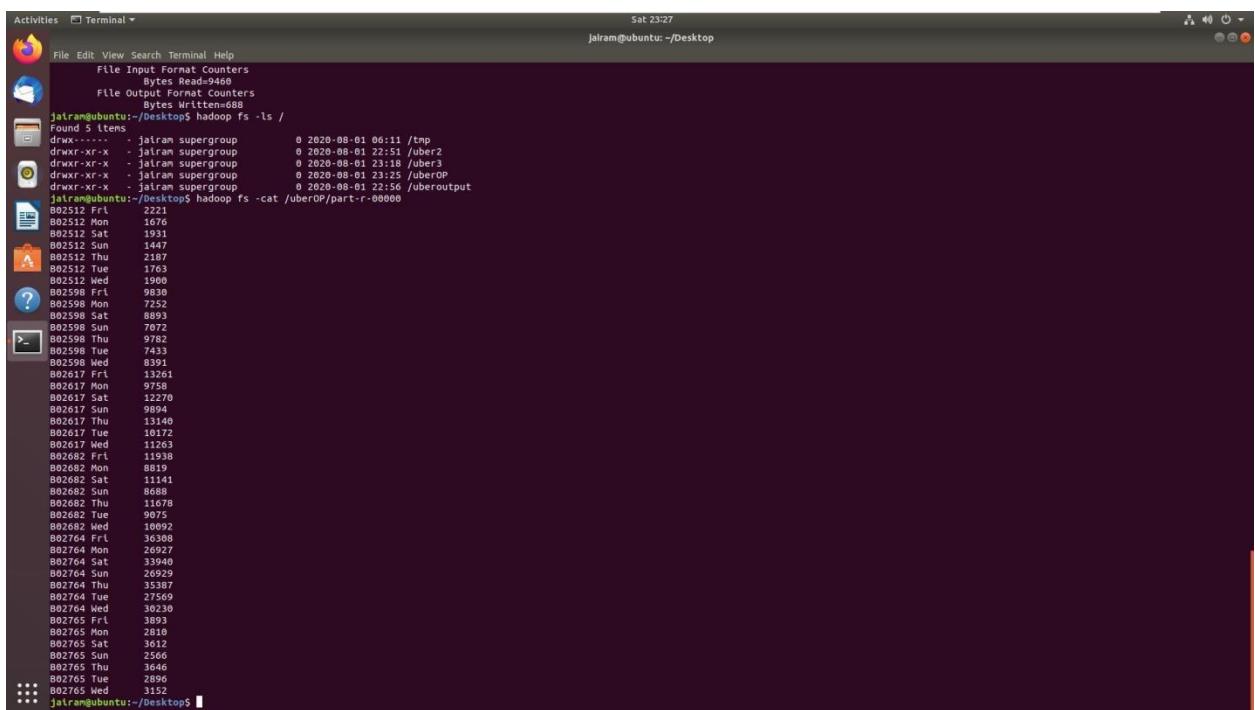
```
hadoop jar wc.jar uber2 /uber3/Uber.txt /uberop
```

10.to check output folder is there or not

*hadoop fs -ls /*

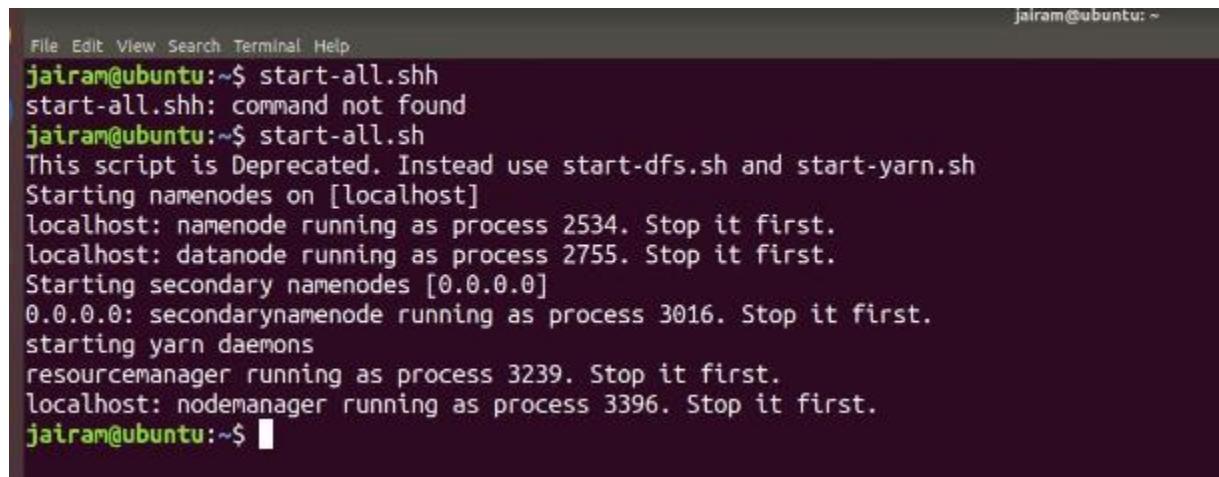


## 11. command: *hadoop fs -cat /uberop/part-r-00000*



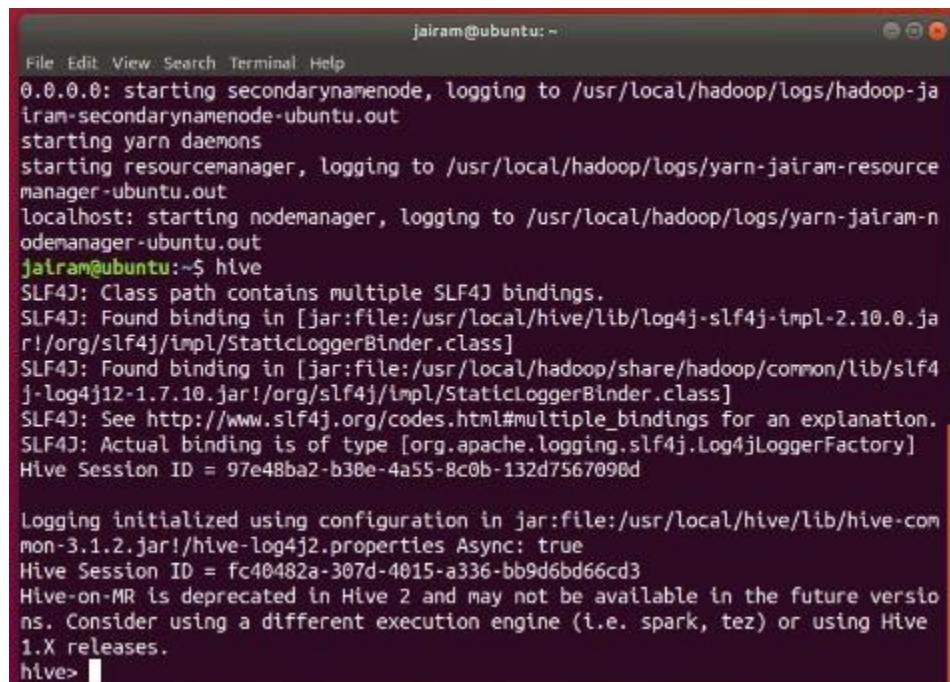
## Pokemon data analysis

1. Open terminal and type Command: *start-all.sh*



```
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$
```

2. Command: *hive* for lauch hive

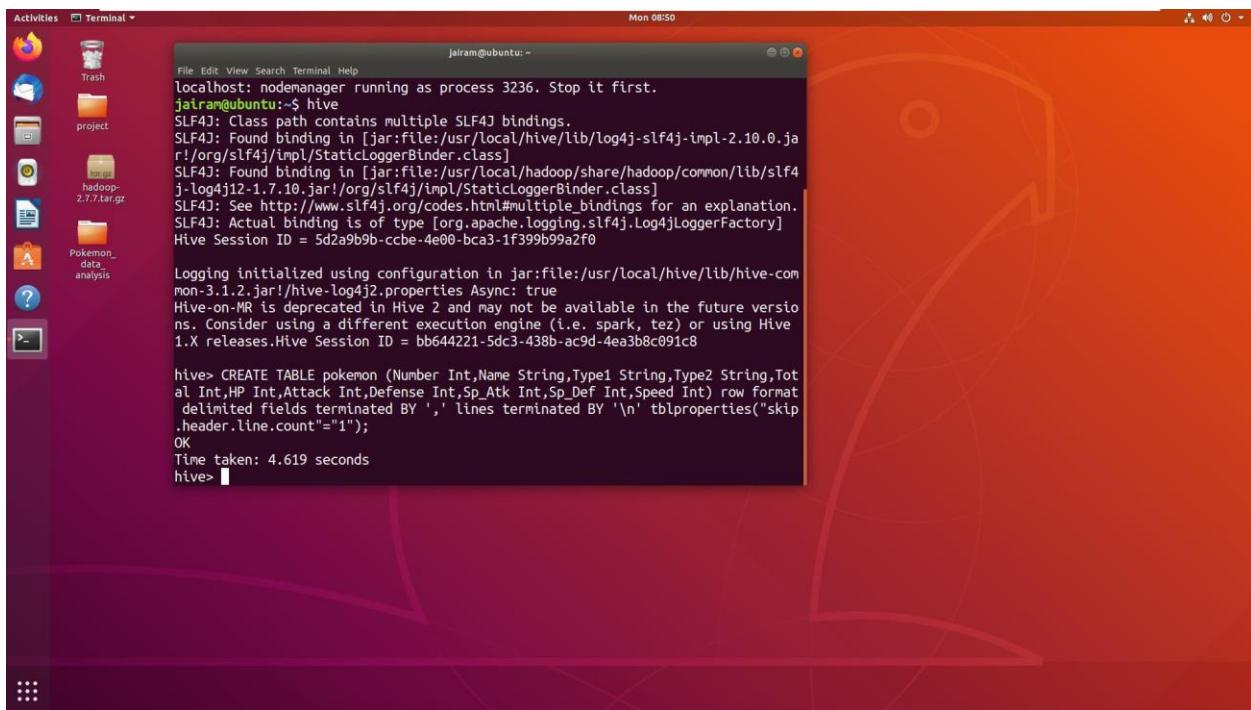


```
jairam@ubuntu:~$ 
File Edit View Search Terminal Help
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-jairam-secondarynamenode-ubuntu.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-jairam-resource
manager-ubuntu.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-jairam-n
odemanager-ubuntu.out
jairam@ubuntu:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.ja
r!/:org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4
j-log4j12-1.7.10.jar!/:org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 97e48ba2-b30e-4a55-8c0b-132d7567090d

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-com
mon-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fc40482a-307d-4015-a336-bb9d6bd66cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive>
```

### 3. to create table and type

```
CREATE TABLE pokemon (Number Int,Name String,Type1 String,Type2 String,Total Int,HP Int,Attack Int,Defense Int,Sp_Atk Int,Sp_Def Int,Speed Int)  
row format delimited fields terminated BY ',' lines terminated BY '\n'  
tblproperties("skip.header.line.count"="1");
```



```
localhost: nodemanager running as process 3236. Stop it first.  
jairam@ubuntu:~$ hive  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]  
Hive Session ID = 5d2a9b9b-ccbe-4e00-bca3-1f399b99a2f0  
Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true  
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.Hive Session ID = bb644221-5dc3-438b-ac9d-4ea3b8c091c8  
hive> CREATE TABLE pokemon (Number Int,Name String,Type1 String,Type2 String,Total Int,HP Int,Attack Int,Defense Int,Sp_Atk Int,Sp_Def Int,Speed Int) row format delimited fields terminated BY ',' lines terminated BY '\n' tblproperties("skip_header.line.count"="1");  
OK  
Time taken: 4.619 seconds  
hive>
```

### 4. to load data and type

```
load data local inpath 'Desktop/Pokemon_data_analysis/Pokemon.csv' INTO  
table pokemon;
```

```
Activities Terminal ▾ Mon 09:21
File Edit View Search Terminal Help jairam@ubuntu: ~
hive> load data local inpath 'Desktop/Pokemon_data_analysis/data/Pokemon.csv' INTO table pokemon;
Loading data to table default.pokemon
OK
Time taken: 0.689 seconds
hive> select avg(HP) from pokemon;
Query ID = jairam_20200803085752_f9d1c6e3-ff2c-4bc1-b066-31f74cd29660
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0002, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0002/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 08:58:22,825 Stage-1 map = 0%, reduce = 0%
2020-08-03 08:58:42,514 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.77 sec
2020-08-03 08:58:59,435 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.98 sec
MapReduce Total cumulative CPU time: 7 seconds 980 msec
Ended Job = job_1596468468570_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.98 sec HDFS Read: 91495 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 980 msec
OK
69.25875
Time taken: 69.087 seconds, Fetched: 1 row(s)
```

5. Find out the average HP (Hit points) of all the Pokémons, using the below query.

Command: ***select avg(HP) from pokemon;***

```

Activities Terminal Mon 09:21
File Edit View Search Terminal Help
hive> load data local inpath 'Desktop/Pokemon_data_analysis/data/Pokemon.csv' INTO table pokemon;
Loading data to table default.pokemon
OK
Time taken: 0.689 seconds
hive> select avg(HP) from pokemon;
Query ID = jairam_20200803085752_f9d1c6e3-ff2c-4bc1-b066-31f74cd29660
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0002, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0002/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 08:58:22,825 Stage-1 map = 0%, reduce = 0%
2020-08-03 08:58:42,514 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.77 sec
2020-08-03 08:58:59,435 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.98 sec
MapReduce Total cumulative CPU time: 7 seconds 980 msec
Ended Job = job_1596468468570_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.98 sec HDFS Read: 91495 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 980 msec
OK
69.25875
Time taken: 69.087 seconds, Fetched: 1 row(s)

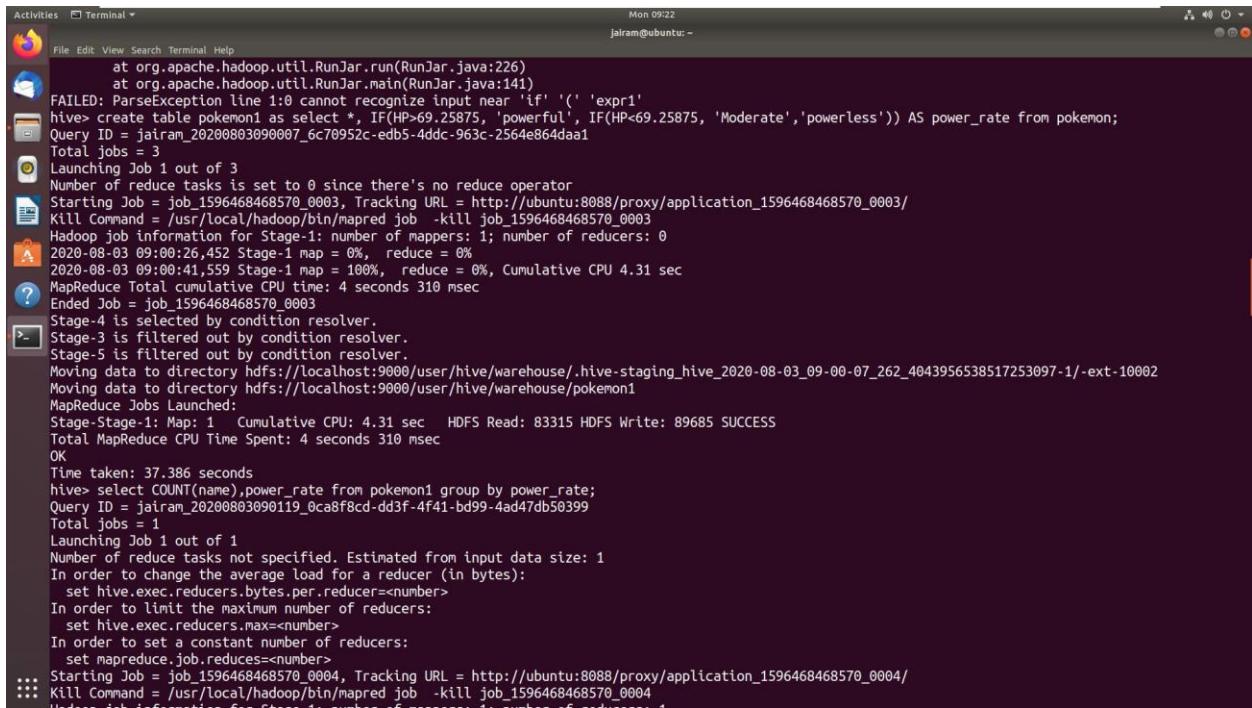
```

6. Create and insert values of existing table ‘pokemon’ into a new table ‘pokemon1’, with an additional column ‘power\_rate’ to find the count of ‘powerful’ and ‘moderate’ from the table ‘pokemon1’.Now, based on the average hit points, we will create another column called ‘power\_rate’.In order to segregate the Pokémon, we will use if condition inside the select statement, which will create one more column in our dataset. The if condition should be used in the following manner inside a Hive query.

Command: *if(expr1,expr2,expr3)*

7. Now, we will create a table based on the condition that if the HP is greater than the average HP, then it is powerful, and if the HP is less than the average, then it is Moderate and a neutral condition is considered as powerless. The same is given as a Hive query below.

Command: *create table pokemon1 as select \*, IF(HP>69.25875, 'powerful', IF(HP<69.25875, 'Moderate','powerless')) AS power\_rate from pokemon;*



```

Activities Terminal * Mon 09:22
jairam@ubuntu: ~
File Edit View Search Terminal Help
at org.apache.hadoop.util.RunJar.run(RunJar.java:226)
at org.apache.hadoop.util.RunJar.main(RunJar.java:141)
FAILED: ParseException line 1:0 cannot recognize input near 'if' '(' 'expr1'
hive> create table pokemon1 as select *, IF(HP>69.25875, 'powerful', IF(HP<69.25875, 'Moderate','powerless')) AS power_rate from pokemon;
Query ID = jairam_20200803090007_6c70952c-edb5-4ddc-963c-2564e864daa1
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1596468468570_0003, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0003/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-08-03 09:00:26,452 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:00:41,559 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.31 sec
MapReduce Total cumulative CPU time: 4 seconds 310 msec
Ended Job = job_1596468468570_0003
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/.hive-staging_hive_2020-08-03_09-00-07_262_4043956538517253097-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/pokemon1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.31 sec HDFS Read: 8315 HDFS Write: 8965 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 310 msec
OK
Time taken: 37.386 seconds
hive> select COUNT(name),power_rate from pokemon1 group by power_rate;
Query ID = jairam_20200803090119_0ca8f8cd-dd3f-4f41-bd99-4ad47db50399
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0004, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0004/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0004

```

8. With the above query, a new column power\_rate has been created. Now, we will find out the number of powerful and moderate HP Pokémons present, using the below query.

Command: *select COUNT(name),power\_rate from pokemon1 group by power\_rate;*

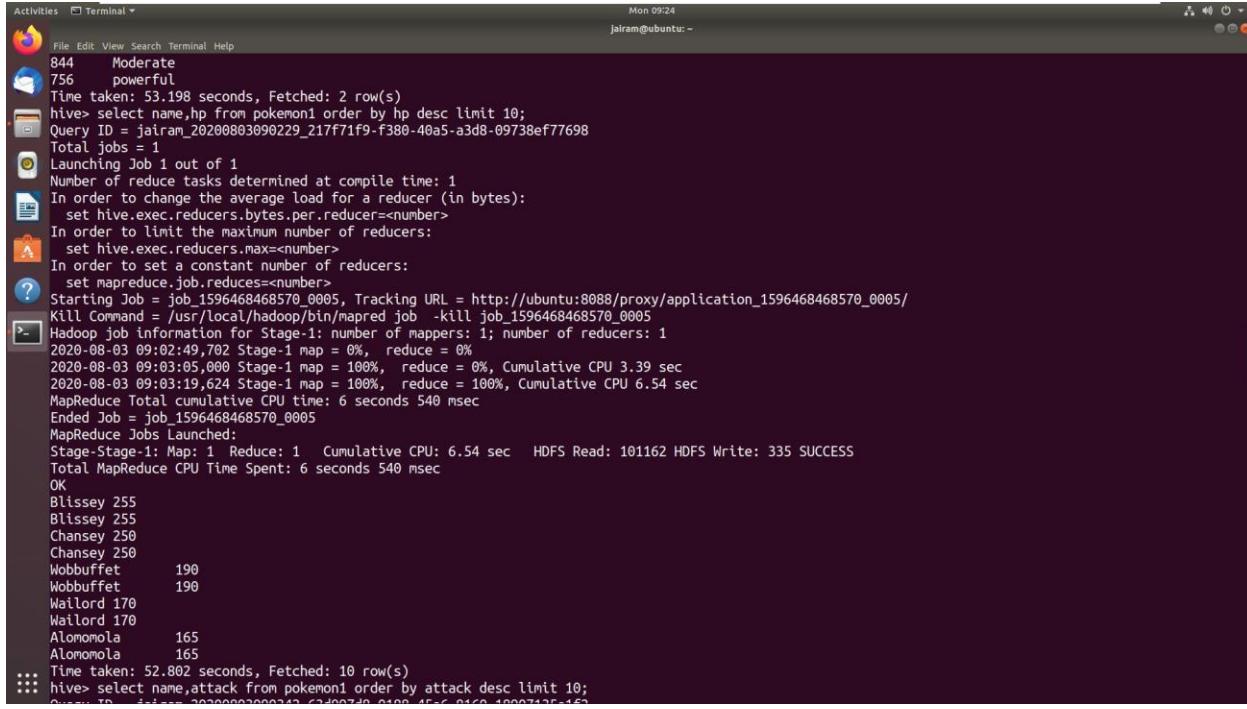
```

Activities Terminal Mon 09:22
File Edit View Search Terminal Help
at org.apache.hadoop.util.RunJar.run(RunJar.java:226)
at org.apache.hadoop.util.RunJar.main(RunJar.java:141)
FAILED: ParseException line 1:0 cannot recognize input near 'if' '(' 'expr1'
hive> create table pokemon1 as select *, IF(HP>69.25875, 'powerful', IF(HP<69.25875, 'Moderate','powerless')) AS power_rate from pokemon;
Query ID = jairam_20200803090007_6c70952c-edb5-4ddc-963c-2564e864daa1
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1596468468570_0003, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0003/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-08-03 09:00:26,452 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 4.31 sec
2020-08-03 09:00:41,559 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.31 sec
MapReduce Total cumulative CPU time: 4 seconds 310 msec
Ended Job = job_1596468468570_0003
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/.hive-staging_hive_2020-08-03_09-00-07_262_4043956538517253097-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/pokemon1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.31 sec HDFS Read: 83315 HDFS Write: 89685 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 310 msec
OK
Time taken: 37.386 seconds
hive> select COUNT(name),power_rate from pokemon1 group by power_rate;
Query ID = jairam_20200803090119_0ca8f8cd-dd5f-4f41-bd99-4ad47db50399
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0004, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0004/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

```

9. Find out the top 10 Pokémons according to their HP's using the below query.

Command: ***select name,hp from pokemon1 order by hp desc limit 10;***



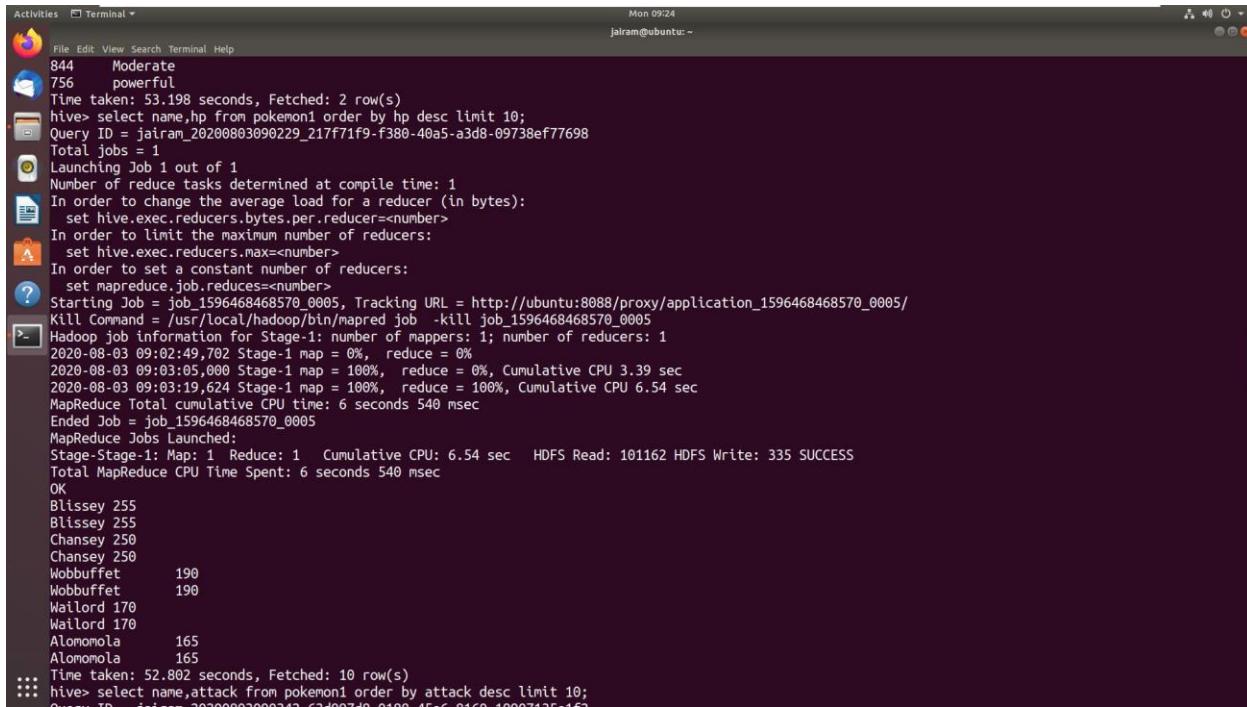
```

Activities Terminal Mon 09:24
jalram@ubuntu: ~
File Edit View Search Terminal Help
844 Moderate
756 powerful
Time taken: 53.198 seconds, Fetched: 2 row(s)
hive> select name,hp from pokemon1 order by hp desc limit 10;
Query ID = jairam_20200803090229_217f71f9-f380-40a5-a3d8-09738ef77698
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0005, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0005/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 09:02:49.702 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:03:05.000 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.39 sec
2020-08-03 09:03:19.624 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.54 sec
MapReduce Total cumulative CPU time: 6 seconds 540 msec
Ended Job = job_1596468468570_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.54 sec HDFS Read: 101162 HDFS Write: 335 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 540 msec
OK
Blissey 255
Blissey 255
Chansey 250
Chansey 250
Wobbuffet 190
Wobbuffet 190
Wailord 170
Wailord 170
Alomomola 165
Alomomola 165
Time taken: 52.802 seconds, Fetched: 10 row(s)
hive> select name,attack from pokemon1 order by attack desc limit 10;
Query ID = jairam_20200803090229_217f71f9-f380-40a5-a3d8-09738ef77698

```

10. Find out the top 10 Pokémons based on their Attack stat, using the below query.

Command: ***select name,attack from pokemon1 order by attack desc limit 10;***



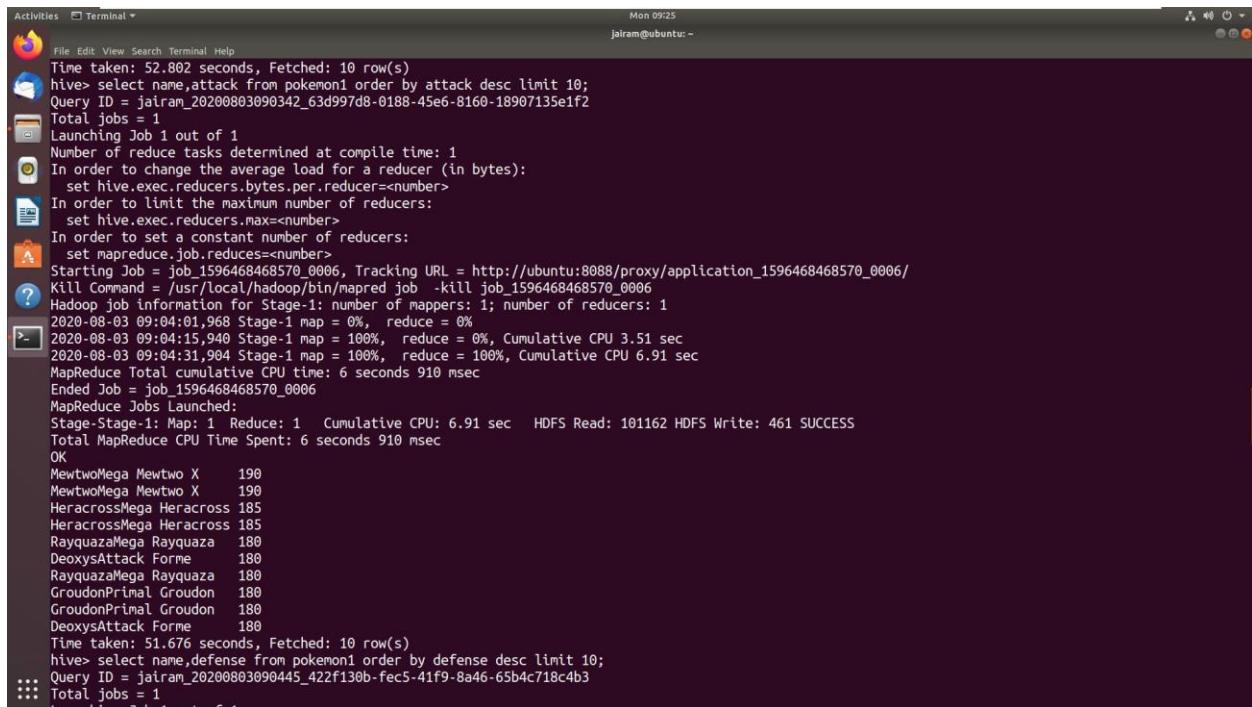
```

Activities Terminal Mon 09:24
jalram@ubuntu: ~
File Edit View Search Terminal Help
844 Moderate
756 powerful
Time taken: 53.198 seconds, Fetched: 2 row(s)
hive> select name,hp from pokemon1 order by hp desc limit 10;
Query ID = jairam_20200803090229_217f71f9-f380-40a5-a3d8-09738ef77698
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0005, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0005/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 09:02:49.702 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:03:05.000 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.39 sec
2020-08-03 09:03:19.624 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.54 sec
MapReduce Total cumulative CPU time: 6 seconds 540 msec
Ended Job = job_1596468468570_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.54 sec HDFS Read: 101162 HDFS Write: 335 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 540 msec
OK
Blissey 255
Blissey 255
Chansey 250
Chansey 250
Wobbuffet 190
Wobbuffet 190
Wailord 170
Wailord 170
Alomomola 165
Alomomola 165
Time taken: 52.802 seconds, Fetched: 10 row(s)
hive> select name,attack from pokemon1 order by attack desc limit 10;
Query ID = jairam_20200803090229_217f71f9-f380-40a5-a3d8-09738ef77698

```

11.Find out the top 10 Pokémons based on their Defense stat, using the below query. Now, we will see the list of top 10 Pokémons according to their defense, using the below query.

Command: *select name,defense from pokemon1 order by defense desc limit 10;*



```
Activities Terminal Mon 09:25
File Edit View Search Terminal Help
Time taken: 52.802 seconds, Fetched: 10 row(s)
hive> select name,attack from pokemon1 order by attack desc limit 10;
Query ID = jairam_20200803090342_63d997d8-0188-45e6-8160-18907135e1f2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0006, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0006/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 09:04:01,968 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:04:15,940 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.51 sec
2020-08-03 09:04:31,904 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.91 sec
MapReduce Total cumulative CPU time: 6 seconds 910 msec
Ended Job = job_1596468468570_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.91 sec HDFS Read: 101162 HDFS Write: 461 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 910 msec
OK
MewtwoMega Mewtwo X 190
MewtwoMega Mewtwo X 190
HeracrossMega Heracross 185
HeracrossMega Heracross 185
RayquazaMega Rayquaza 180
DeoxysAttack Forme 180
RayquazaMega Rayquaza 180
GroudonPrimal Groudon 180
GroudonPrimal Groudon 180
DeoxysAttack Forme 180
Time taken: 51.676 seconds, Fetched: 10 row(s)
hive> select name,defense from pokemon1 order by defense desc limit 10;
Query ID = jairam_20200803090445_422f130b-fec5-41f9-8a46-65b4c718c4b3
Total jobs = 1
Time taken: 51.676 seconds, Fetched: 10 row(s)
```

12.Find out the top 10 Pokémons based on their total power.

Command: *select name,total from pokemon1 order by total desc limit 10;*

13.Find out the top 10 Pokémons having a drastic change in their attack and sp.attack, using the below query.

Command: *select name,(attack-sp\_atk) as atk\_diff from pokemon1 order by atk\_diff limit 10;*

```

Activities Terminal Mon 09:26
File Edit View Search Terminal Help
GroudonPrimal Groudon    770
KyogrePrimal Kyogre     770
Time taken: 51.19 seconds, Fetched: 10 row(s)
hive> select name,(attack-sp_atk) as atk_diff from pokemon1 order by atk_diff limit 10;
Query ID = jairam_20200803090640_97bddacb-b7d0-4ac6-b8a7-4a4c00231b89
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0009, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0009/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 09:06:58,188 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:07:13,089 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.75 sec
2020-08-03 09:07:28,755 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.45 sec
MapReduce Total cumulative CPU time: 7 seconds 450 msec
Ended Job = job_1596468468570_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.45 sec HDFS Read: 102485 HDFS Write: 406 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 450 msec
OK
AlakazamMega Alakazam    -125
AlakazamMega Alakazam    -125
DarmanitanZen Mode      -110
DarmanitanZen Mode      -110
GengarMega Gengar        -105
GengarMega Gengar        -105
Chandelure       -90
Chandelure       -90
Abra            -85
Duosion          -85
Time taken: 49.656 seconds, Fetched: 10 row(s)
hive> select name,(defense-sp_defense) as def_diff from pokemon1 order by def_diff limit 10;

```

14. Find out the top 10 Pokémons having a drastic change in their defense and sp.defense, using the below query.

Command: *select name,(defense-sp\_defense) as def\_diff from pokemon1 order by def\_diff limit 10;*

```

Activities Terminal Mon 09:27
File Edit View Search Terminal Help
hive> select name,(defense-sp_defense) as def_diff from pokemon1 order by def_diff limit 10;
FAILED: SemanticException [Error 10004]: Line 1:21 Invalid table alias or column reference 'sp_defense': (possible column names are: number, name, type1,
type2, total, hp, attack, defense, sp_atk, sp_def, speed, power_rate)
hive> Select name, speed from pokemon order by speed desc limit 10;
Query ID = jairam_20200803090840_6be72595-e486-46fe-8043-e84490fd129a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0010, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0010/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 09:08:57,461 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:09:11,910 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.64 sec
2020-08-03 09:09:27,389 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.11 sec
MapReduce Total cumulative CPU time: 7 seconds 110 msec
Ended Job = job_1596468468570_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.11 sec HDFS Read: 87057 HDFS Write: 430 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 110 msec
OK
DeoxysSpeed Forme      180
DeoxysSpeed Forme      180
Ninjask 160
Ninjask 160
AerodactylMega Aerodactyl      150
AerodactylMega Aerodactyl      150
DeoxysAttack Forme     150
DeoxysNormal Forme     150
DeoxysNormal Forme     150
AlakazamMega Alakazam      150
Time taken: 49.418 seconds, Fetched: 10 row(s)
hive> select name,(defense-sp_Def) as def_diff from pokemon1 order by def_diff limit 10;
Query ID = jairam_20200803090840_6be72595-e486-46fe-8043-e84490fd129a

```

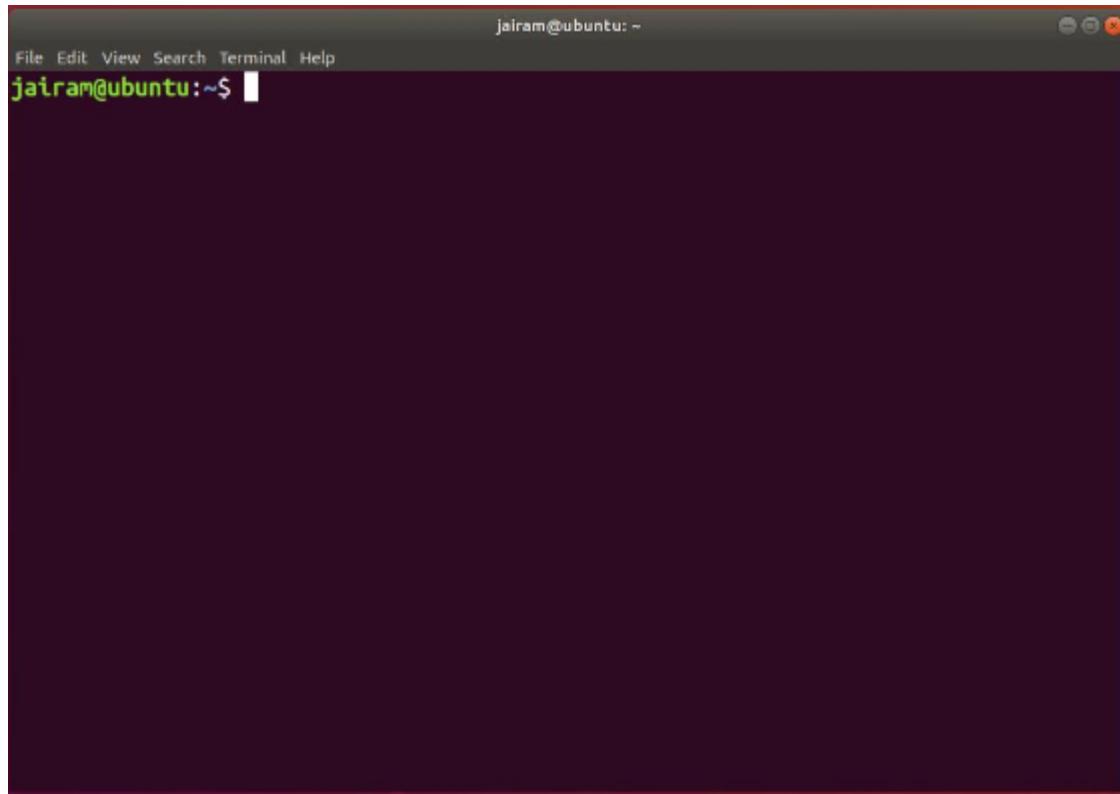
15. Find out the top 10 fastest Pokémons, using the below query.

Command: *Select name, speed from pokemon order by speed desc limit 10;*

```
Activities Terminal Mon 09:28
File Edit View Search Terminal Help jalram@ubuntu: ~
DeoxysNormal Forme 150
AlakazamMega Alakazam 150
Time taken: 49.418 seconds, Fetched: 10 row(s)
hive> select name,(defense_sp_Def) as def_diff from pokemon1 order by def_diff limit 10;
Query ID = jairam_20200803091003_bc4fcec9-6108-41ca-8320-ec30d11b007e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0011, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0011/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 09:10:20,889 Stage-1 map = 0%, reduce = 0%
2020-08-03 09:10:35,794 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.65 sec
2020-08-03 09:10:51,487 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.02 sec
MapReduce Total cumulative CPU time: 7 seconds 20 msec
Ended Job = job_1596468468570_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.02 sec HDFS Read: 102485 HDFS Write: 337 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 20 msec
OK
Blissey -125
Blissey -125
Cryogonal -105
Cryogonal -105
Regice -100
Chansey -100
Regice -100
Chansey -100
Florges -86
Florges -86
Time taken: 50.164 seconds, Fetched: 10 row(s)
hive>
```

## Time zone analysis

1. Paste the time\_zone\_analysis folder on Desktop
2. Open terminal



3. run the script file

Command :*pig -x local Desktop/time\_zone\_analysis/script/time\_zone\_analysis.pig*

```

Tue 04:47
jalram@ubuntu:~$ pig -x local Desktop/time_zone_analysis/script/time_zone_analysis.pig
2020-08-04 04:46:23 INFO org.apache.pig.ExecTypeProvider: Trying ExecType : LOCAL
2020-08-04 04:46:23 INFO org.apache.pig.ExecTypeProvider: Picked LOCAL as the ExecType
2020-08-04 04:46:23.416 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r179736) compiled Jun 02 2017, 15:41:58
2020-08-04 04:46:23.416 [main] INFO org.apache.pig.Main - Copyright 2009 The Apache Software Foundation
2020-08-04 04:46:23.416 [main] INFO org.apache.pig.Main - This software includes distributed code. Please see license information in the top-level directory of this distribution.
2020-08-04 04:46:23.424 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2020-08-04 04:46:23.424 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-08-04 04:46:23.879 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-08-04 04:46:23.881 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2020-08-04 04:46:23.996 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIC-time_zone_analysis.pig-54afea76-d008-4581-bd8e-300d9bb1df23
2020-08-04 04:46:23.910 [main] WARN org.apache.pig.PigServer - AIS is disabled since yarn.timeline-service.enabled set to false
2020-08-04 04:46:24.062 [main] INFO org.apache.pig.impl.util.SplittableMemoryManager - Selected Gen (tenured Gen) of size 69072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 4893
50752
2020-08-04 04:46:24.766 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-08-04 04:46:24.766 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 4 time(s).
2020-08-04 04:46:24.768 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTIONS 1 time(s).
2020-08-04 04:46:24.780 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalOptimizer - mapred.job.reduce.set is deprecated. Instead, use mapreduce.job.set
2020-08-04 04:46:24.780 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalOptimizer - mapred.job.reduces is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2020-08-04 04:46:24.780 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalOptimizer - (RULES_ENABLEDForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushUpForEachLimiter, PushUpFilter, SplitFilter, StreamTypeCastInsertor)
2020-08-04 04:46:24.933 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Map key required for load_tweets: $0>[id, text, user]

P-
2020-08-04 04:46:25.040 [INFO] org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.HadoopCompiler - File concatenation threshold: 100 optimistic? false
2020-08-04 04:46:25.103 [INFO] org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.HadoopCompiler - U
2020-08-04 04:46:25.168 [INFO] org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2020-08-04 04:46:25.168 [INFO] org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.QueryOptimizer - MR plan size after optimization: 3
2020-08-04 04:46:25.213 [INFO] org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.QueryOptimizer - Initializing Metrics with processName=JobTracker, sessionId=0
2020-08-04 04:46:25.247 [INFO] org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2020-08-04 04:46:25.254 [INFO] org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2020-08-04 04:46:25.254 [INFO] org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2020-08-04 04:46:25.258 [INFO] org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2020-08-04 04:46:25.258 [INFO] org.apache.hadoop.conf.Configuration.deprecation - mapred.job.output.dir is deprecated. Instead, use mapreduce.taskattempt.dir
2020-08-04 04:46:25.266 [INFO] org.apache.hadoop.data.Schematuplefrontend - Schema tuple frontend is false. It will not generate code.
2020-08-04 04:46:25.286 [INFO] org.apache.hadoop.data.Schematuplefrontend - Starting process to move generated code to distributed cache
2020-08-04 04:46:25.286 [INFO] org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1595541585282_0
2020-08-04 04:46:25.341 [INFO] org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission
2020-08-04 04:46:25.442 [INFO] org.apache.hadoop.mapreduce.Job - Configured JobName: time_zone_analysis
2020-08-04 04:46:25.442 [INFO] org.apache.hadoop.mapreduce.Job - Configuration for job: time_zone_analysis is ready initialized
2020-08-04 04:46:25.471 [INFO] [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2020-08-04 04:46:25.471 [INFO] [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2020-08-04 04:46:25.478 [INFO] [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-08-04 04:46:25.478 [INFO] [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
2020-08-04 04:46:25.511 [INFO] [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2020-08-04 04:46:25.511 [INFO] [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local390335987_0001
2020-08-04 04:46:25.571 [INFO] [JobControl] INFO org.apache.hadoop.mapred.LocalJobRunner - The url to track the job: http://localhost:8080/
2020-08-04 04:46:25.584 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2020-08-04 04:46:25.584 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.LocalJobRunner - Configuration for job: time_zone_analysis is ready initialized
2020-08-04 04:46:25.639 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.JobConf - Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.job.address
2020-08-04 04:46:25.639 [INFO] [Thread-5] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter Algorithm version is 1
2020-08-04 04:46:25.641 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter is org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigOutputCommitter
2020-08-04 04:46:25.641 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.LocalJobRunner - Job completed successfully
2020-08-04 04:46:25.641 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.LocalJobRunner - User classes may not be found. See Job or Job#setJar(String).
2020-08-04 04:46:25.641 [INFO] [Thread-5] INFO org.apache.hadoop.mapred.LocalJobRunner - Job completed successfully
Tue 04:48
jalram@ubuntu:~$ 

```

```

Tue 04:48
jalram@ubuntu:~$ pig part-r-00000
File Edit View Search Terminal Help
          Open...       Save...
part-r-00000   /desktop/time_zone_analysis/output
Success!
job_stats (time in seconds):
  JobId  Maps Reduces MaxMapTime MinMapTime
  job_local390335987.0001 1      1        2.0
  OIN_GROUP_BY
  job_local390335987.0001 1      0        n/a
  job_local390335987.0002 1      1        n/a
  job_local390335987.0003 1      1        n/a

Input(s):
Successfully read 2477 records from: "file:///home/jalram/time_zone_analysis/part-r-00000"
Successfully read 4846 records from: "file:///home/jalram/time_zone_analysis/part-r-00000"

Output(s):
Successfully stored 16 records to: "file:///home/jalram/time_zone_analysis/part-r-00000"

Counters:
Total records written : 16
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local390335987.0001 ->      job_local21337
job_local21337 ->      job_local390335987.0002
job_local390335987.0002 ->      job_local390335987.0003
job_local390335987.0003 ->      job_local21337

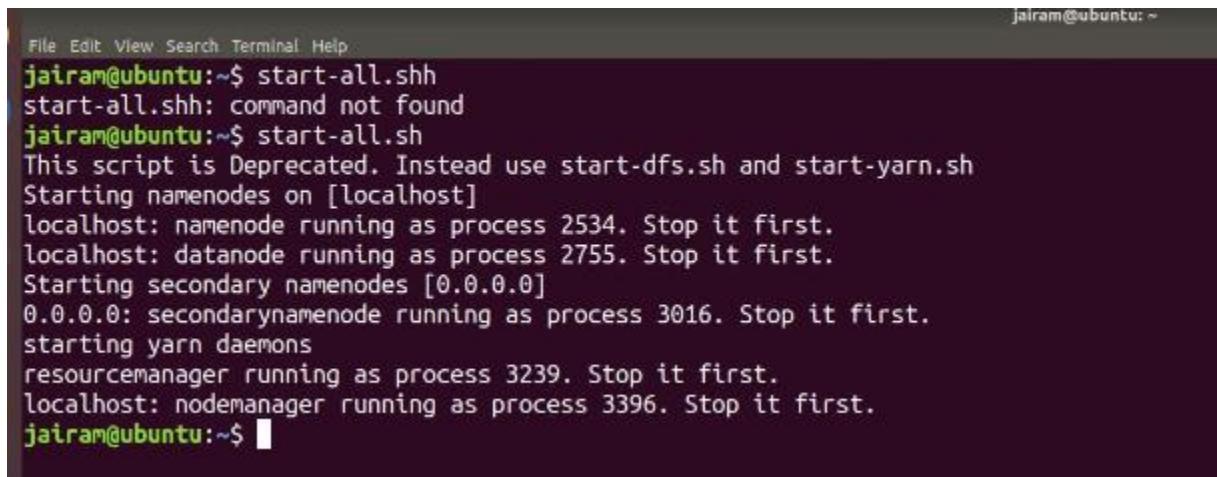
Plain Text Tab Width: 8 □ Ln1, Col1 □ INS
2020-08-04 04:46:32.300 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.300 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.300 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.300 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.314 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.316 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.325 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.326 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.331 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-04 04:46:32.363 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
jalram@ubuntu:~$ 

```

## WordCount using java

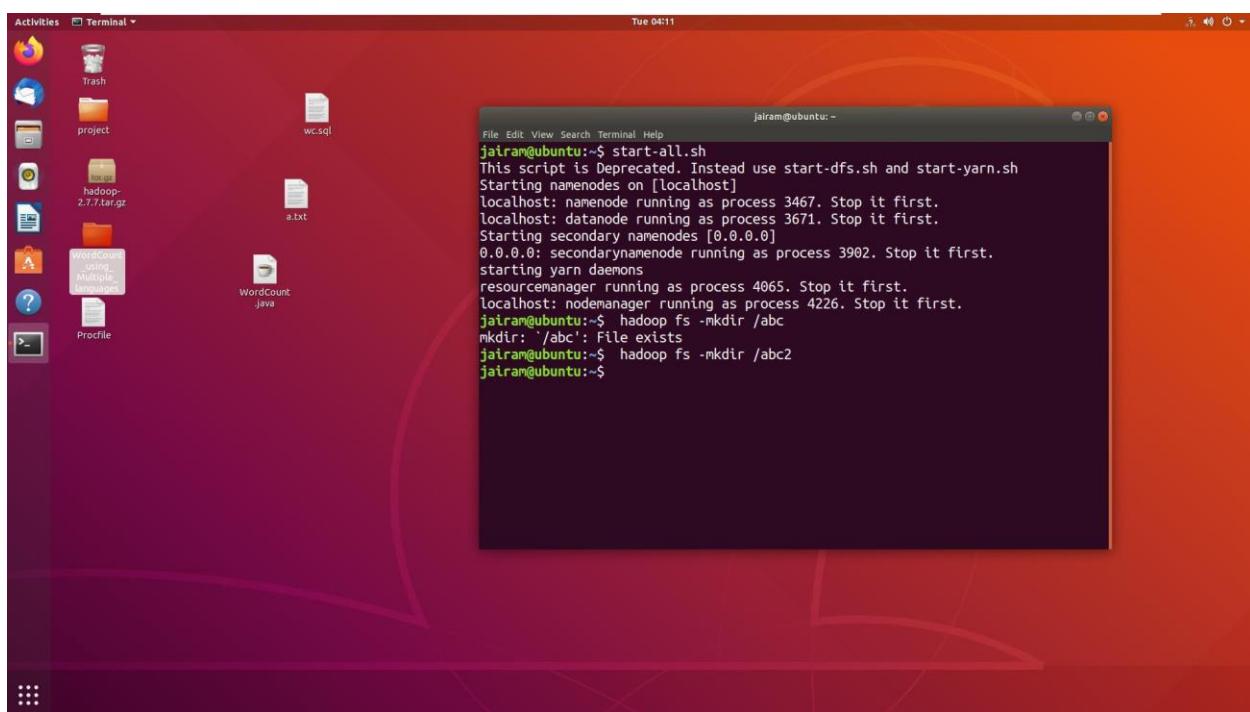
1. Create a text file containing a paragraph i.e a.txt

2. Open terminal and type Command: *start-all.sh*

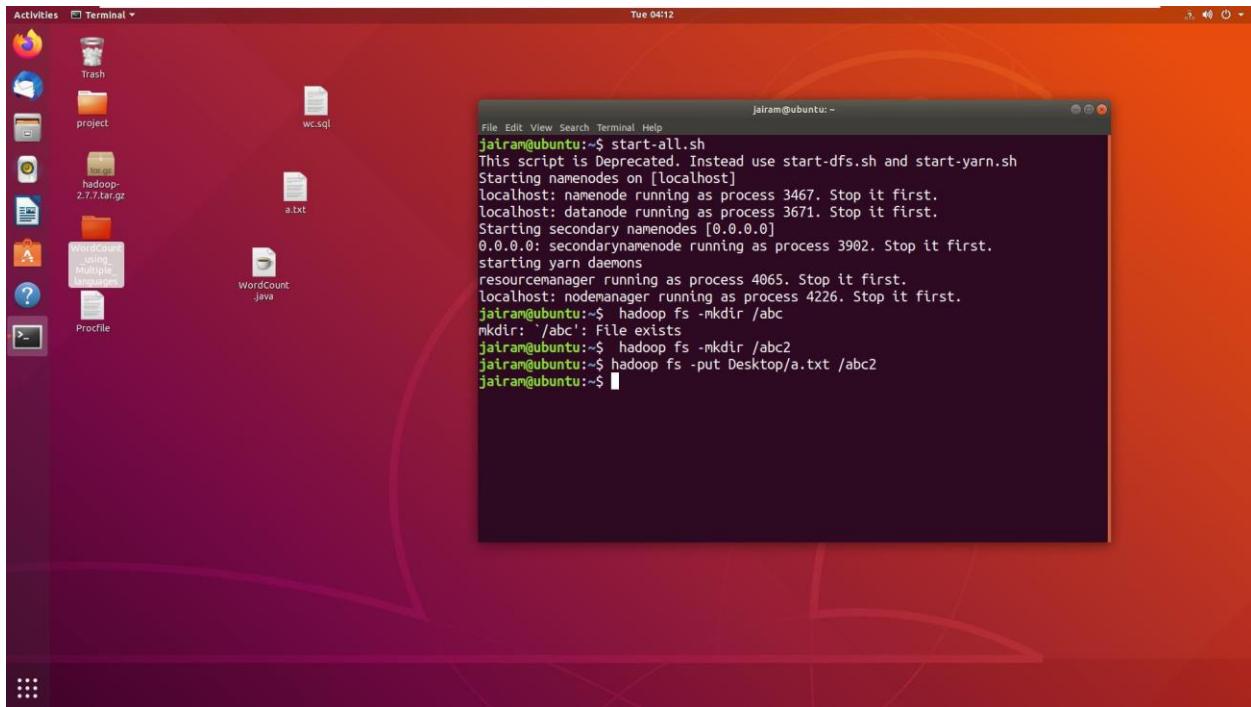


```
File Edit View Search Terminal Help
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$
```

3. Command :**hadoop fs -mkdir /abc2**

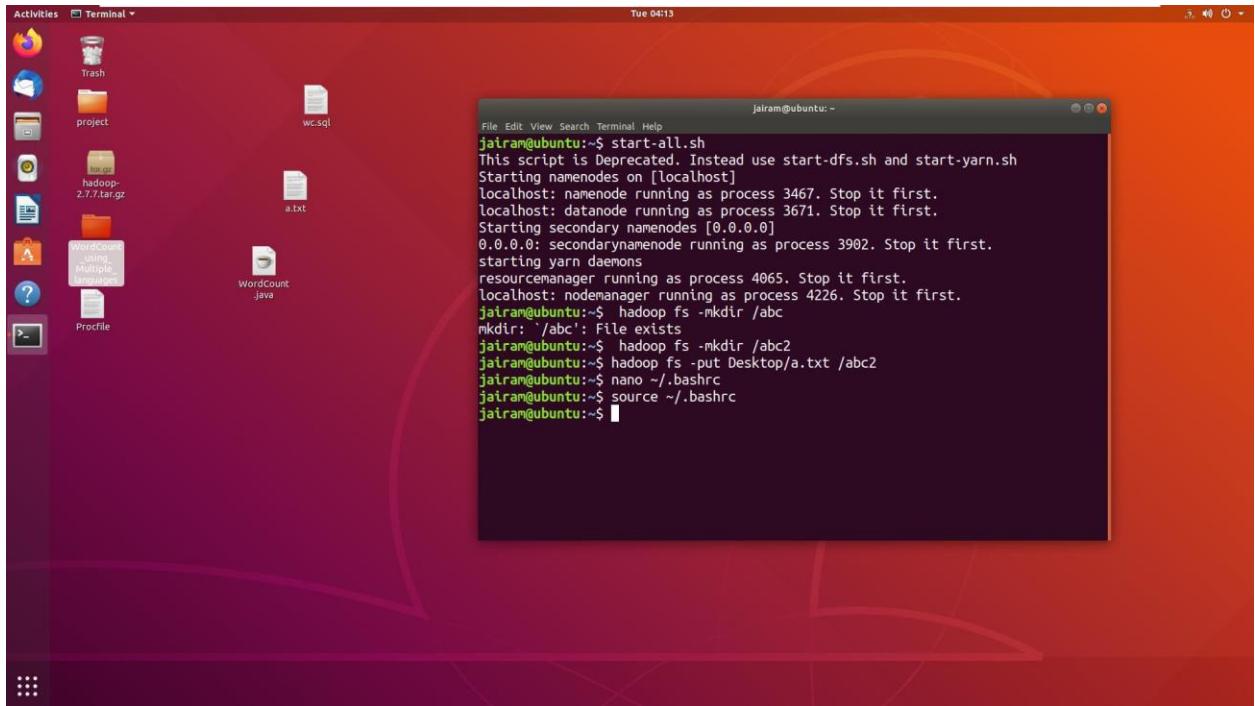


4. Command :*hadoop fs -put Desktop/a.txt /abc2*



5. Open bashrc file by *nano ~/.bashrc* and type

*export HADOOP\_CLASSPATH=\$JAVA\_HOME/lib/tools.jar*

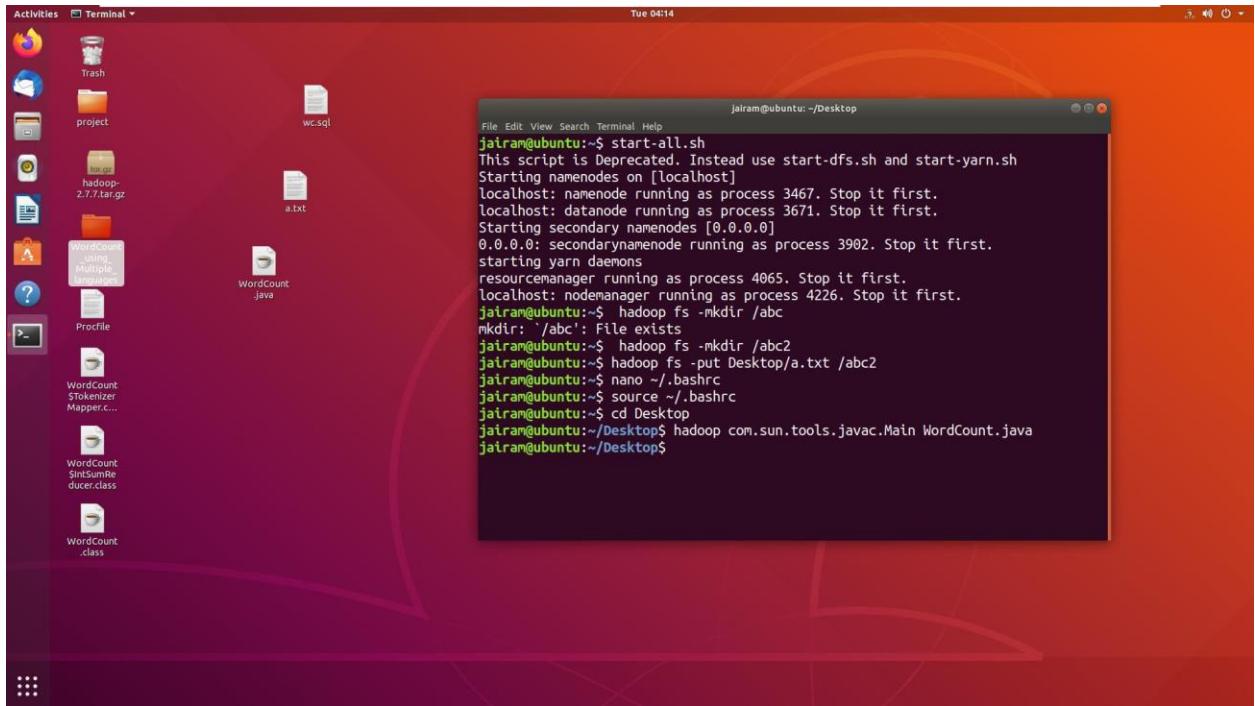


6. Then command :*source ~/.bashrc*

7. command :*cd Desktop*

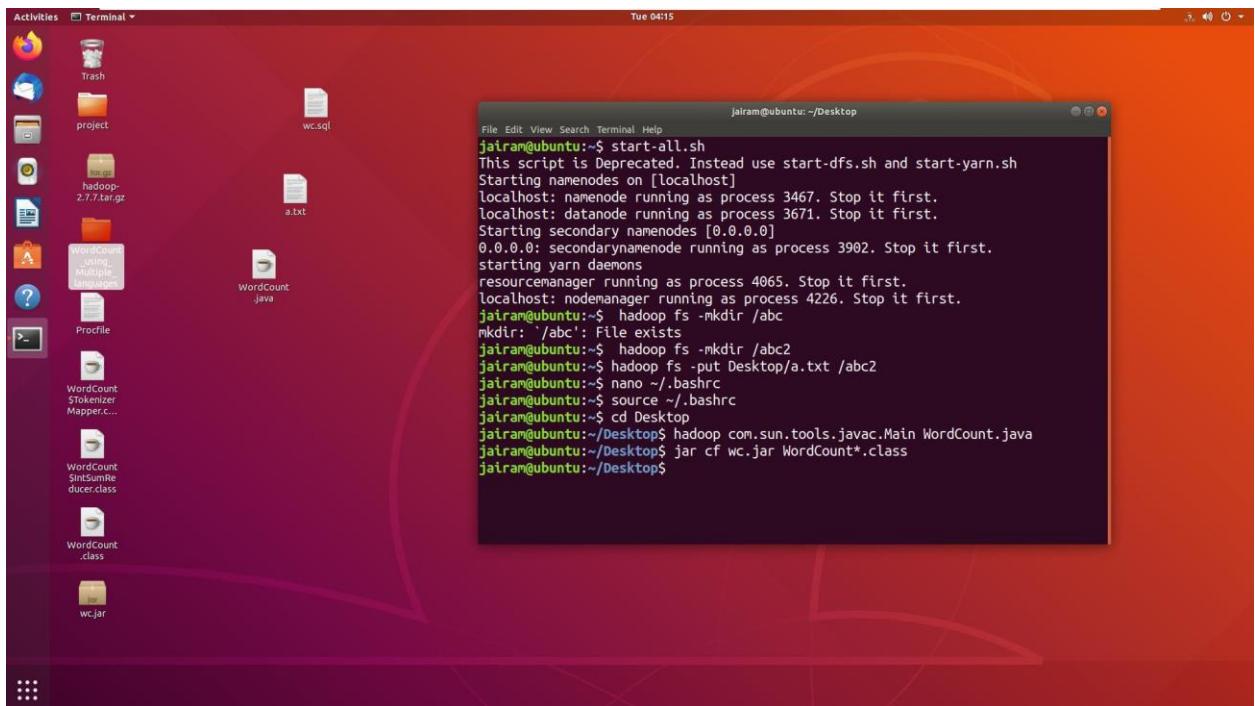
8. Now compile Wordcount.java file suppose that file is on Desktop

Command: *hadoop com.sun.tools.javac.Main WordCount.java*



## 9. To combine all class files

Command:*jar cf wc.jar WordCount\*.class*



## 10. To execute

Command: *hadoop jar wc.jar WordCount /abc2/a.txt /output*

```
jairam@ubuntu:~/Desktop$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 3467. Stop it first.
localhost: datanode running as process 3671. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3902. Stop it first.
starting yarn daemons
resourcemanager running as process 4065. Stop it first.
localhost: nodemanager running as process 4226. Stop it first.
jairam@ubuntu:~$ hadoop fs -mkdir /abc
mkdir: '/abc': File exists
jairam@ubuntu:~$ hadoop fs -mkdir /abc2
jairam@ubuntu:~$ hadoop fs -put Desktop/a.txt /abc2
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ cd Desktop
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main WordCount.java
jairam@ubuntu:~/Desktop$ jar cf wc.jar WordCount*.class
jairam@ubuntu:~/Desktop$ hadoop jar wc.jar WordCount /abc2/a.txt /output
20/08/04 04:15:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
```

## 11. To check output folder is there or not

Command: *hadoop fs -ls /*

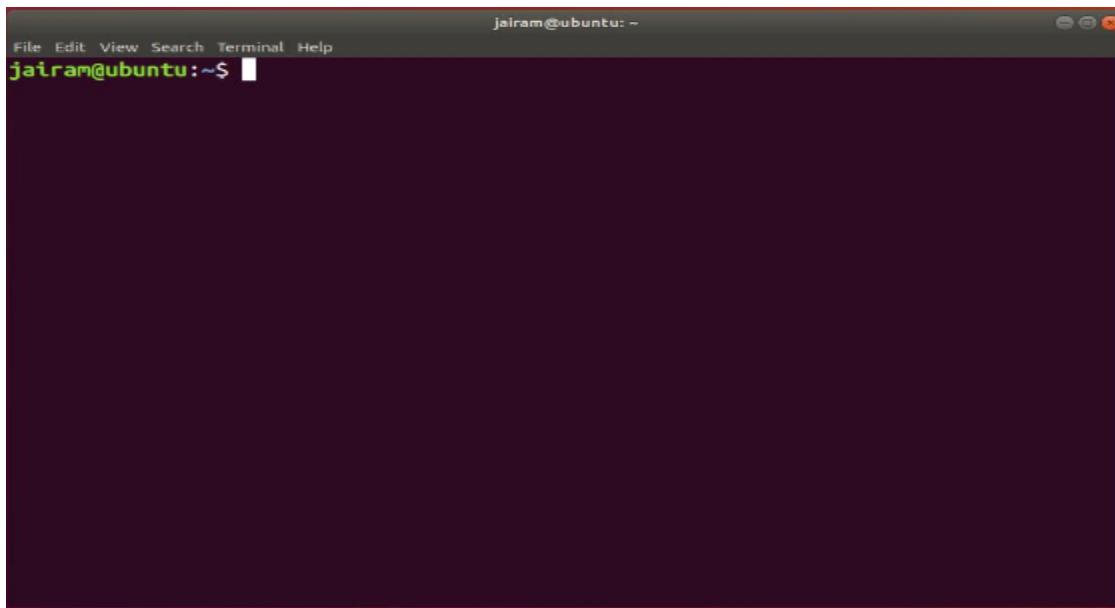
```
jairam@ubuntu:~/Desktop$ hadoop fs -ls /
Found 4 items
drwxr-xr-x  - jairam supergroup      0 2020-08-03 10:32 /abc
drwxr-xr-x  - jairam supergroup      0 2020-08-04 04:12 /abc2
drwxr-xr-x  - jairam supergroup      0 2020-08-03 10:34 /output
drwx----- - jairam supergroup      0 2020-08-02 22:52 /tmp
```

## 12.hadoop fs -cat /output/part-r-00000

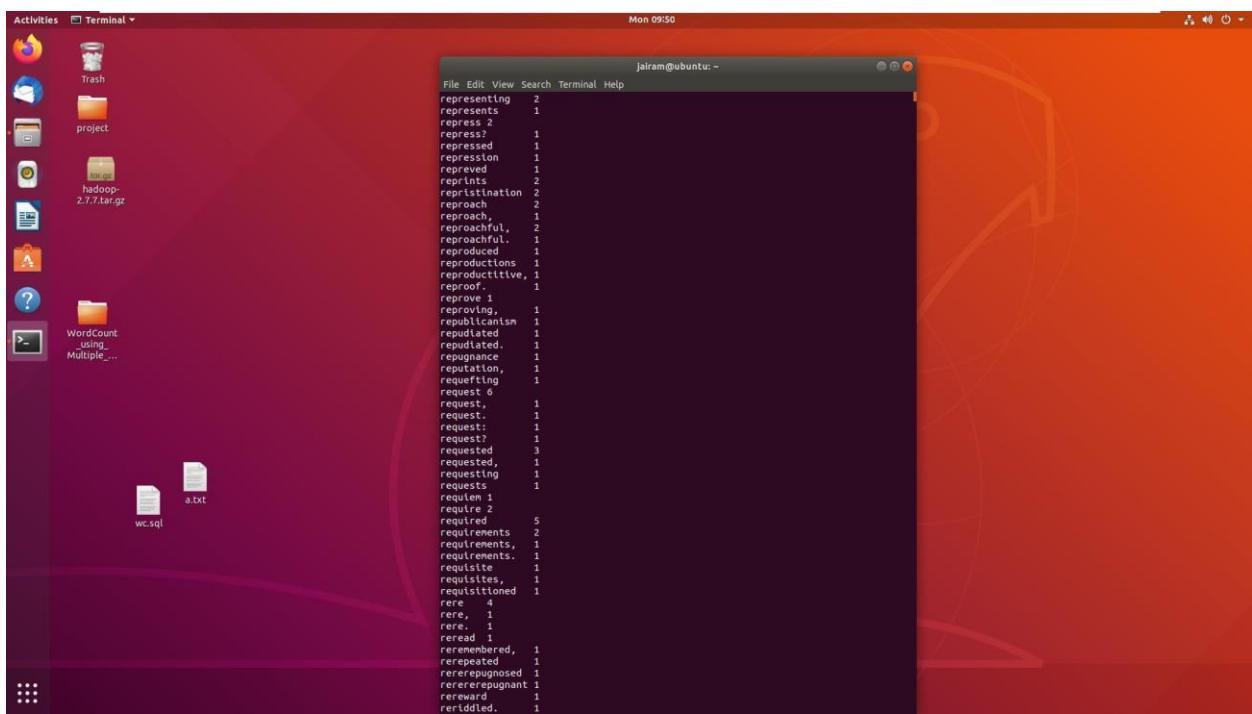
```
Activities Terminal Mon 10:36
File Edit View Search Terminal Help
Jairam@ubuntu: ~/Desktop
represented. 1
representing 2
represents 1
repress 2
repress? 1
repressed 1
repression 1
repreved 1
reprints 2
repristination 2
reproach 2
reproach, 1
reproachful, 2
reproachful. 1
reproduced 1
reproductions 1
reproducttive, 1
reprove 1
reproving, 1
republicanism 1
reputed 1
reputated 1
reputalid. 1
repugnace 1
reputation, 1
requesting 1
request? 1
request? 1
requested 3
requested, 1
requesting 1
request? 1
requiem 1
require 2
required 5
requirements 2
requirements, 1
requirements. 1
requisite 1
requisites, 1
requisitioned 1
rere 4
rere, 1
rere 1
reread 1
reremembered, 1
rerepeated 1
rerepugnose 1
rerepugnant 1
rereared 1
residited, 1
res... 1
res... 2
```

## WordCount using hive

1. . Open terminal



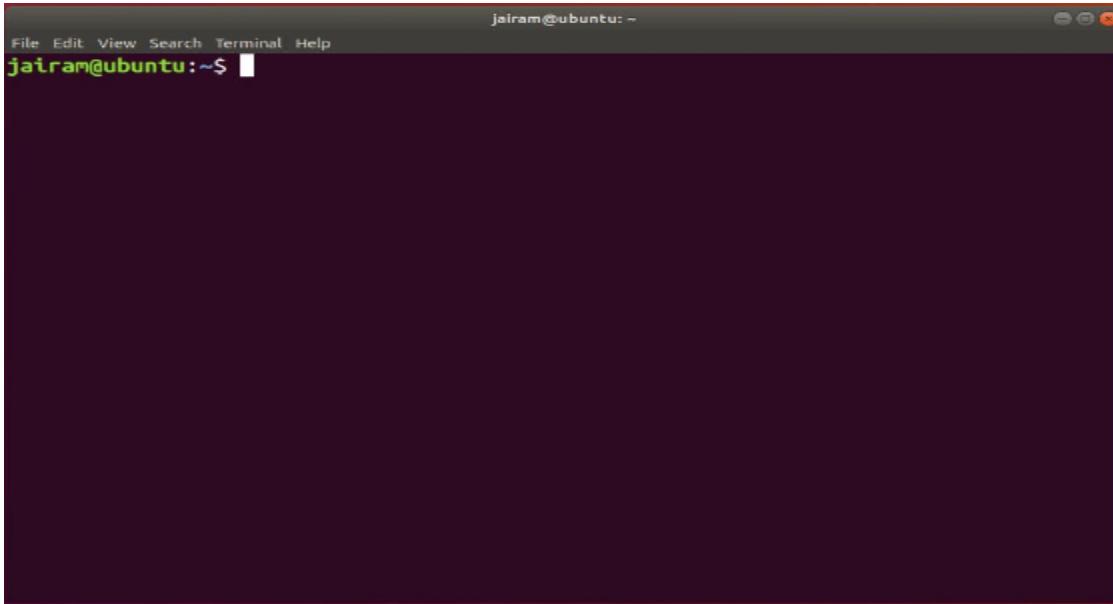
2. Command: *hive -f Desktop/wc.sql*





## WordCount using pig

1. . Open terminal



2. Command: *pig -x local* to lauch the pig

```
jairam@ubuntu:~$ pig -x local
20/08/04 04:19:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
20/08/04 04:19:49 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2020-08-04 04:19:49,621 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2020-08-04 04:19:49,621 [main] INFO org.apache.pig.Main - Logging error messages to: /home/jairam/pig_1596539989619.log
2020-08-04 04:19:49,649 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/jairam/.pigbootup not found
2020-08-04 04:19:49,855 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-08-04 04:19:49,856 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2020-08-04 04:19:50,051 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-08-04 04:19:50,067 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-50cb85ff-ef2f-4b9d-958c-c8b6b876d244
2020-08-04 04:19:50,067 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grun> ■
```

3. Command: `grunt > A = load 'Desktop/a.txt' using TextLoader() as (word:chararray);`

```

Activities Terminal * Tue 04/23
jairam@ubuntu:~$ pig -x local
jairam@ubuntu:~$ pig -x local
20/08/04 04:19:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
20/08/04 04:19:49 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2020-08-04 04:19:49,621 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2020-08-04 04:19:49,621 [main] INFO org.apache.pig.Main - Logging error messages to: /home/jairam/pig_159653989619.log
2020-08-04 04:19:49,649 [main] INFO org.apache.pig.impl.Utils - Default bootup file /home/jairam/.pigbootup not found
2020-08-04 04:19:49,855 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-08-04 04:19:49,856 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///home/jairam/Desktop/a.txt
2020-08-04 04:19:50,051 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-08-04 04:19:50,067 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-50cb85ff-ef2f-4b9d-958c-c8b6b876d244
2020-08-04 04:19:50,067 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load 'Desktop/a.txt' using TextLoader() as (words:chararray);
2020-08-04 04:22:55,260 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> 

```

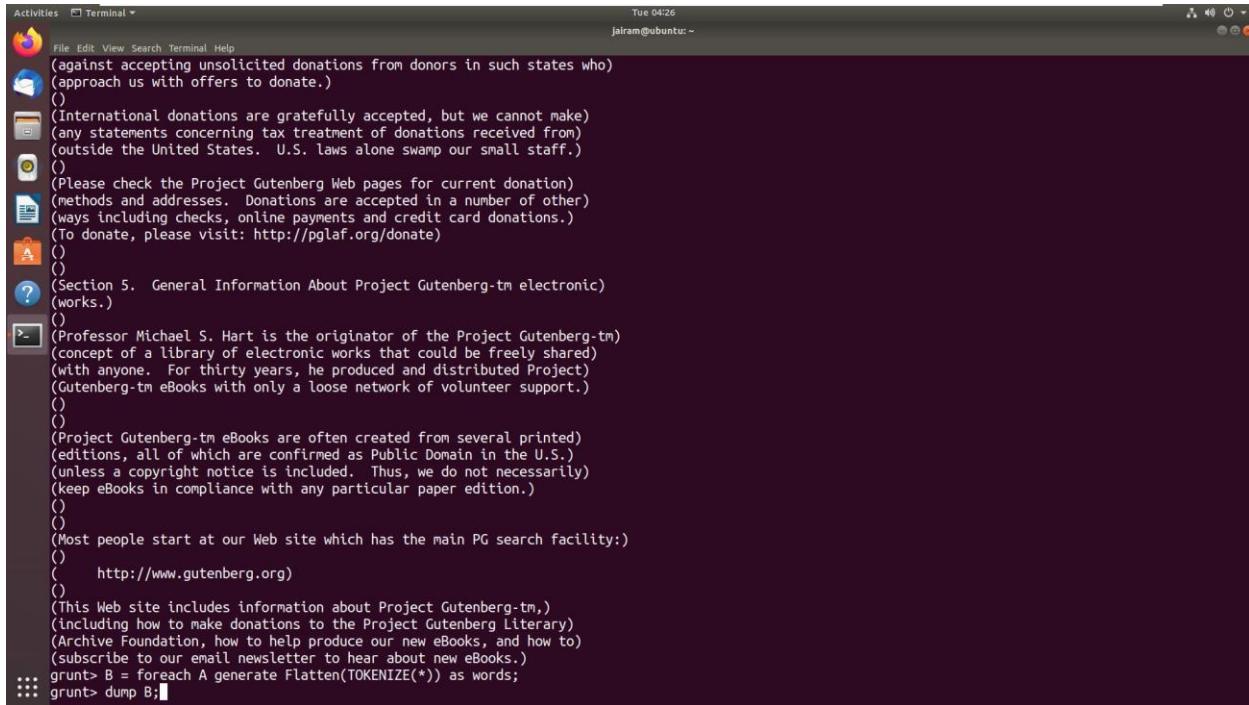
```

Activities Terminal * Tue 04/24
jairam@ubuntu:~$ pig -x local
jairam@ubuntu:~$ pig -x local
( have not met the solicitation requirements, we know of no prohibition)
(against accepting unsolicited donations from donors in such states who
(approach us with offers to donate.)
()
(International donations are gratefully accepted, but we cannot make)
(any statements concerning tax treatment of donations received from)
(outside the United States. U.S. laws alone swamp our small staff.)
()
(Please check the Project Gutenberg Web pages for current donation)
(methods and addresses. Donations are accepted in a number of other)
(ways including checks, online payments and credit card donations.)
(To donate, please visit: http://pglaf.org/donate)
()
()
(Section 5. General Information About Project Gutenberg-tm electronic)
(works.)
()
(Professor Michael S. Hart is the originator of the Project Gutenberg-tm)
(concept of a library of electronic works that could be freely shared)
(with anyone. For thirty years, he produced and distributed Project)
(Gutenberg-tm eBooks with only a loose network of volunteer support.)
()
()
(Project Gutenberg-tm eBooks are often created from several printed)
(editions, all of which are confirmed as Public Domain in the U.S.)
(unless a copyright notice is included. Thus, we do not necessarily)
(keep eBooks in compliance with any particular paper edition.)
()
()
( Most people start at our Web site which has the main PG search facility:)
()
()
(http://www.gutenberg.org)
()
(This Web site includes information about Project Gutenberg-tm,
(including how to make donations to the Project Gutenberg Literary)
(Archive Foundation, how to help produce our new eBooks, and how to)
(subscribe to our email newsletter to hear about new eBooks.)
...
grunt> 

```

4. Command: **grunt > B = foreach A generate flatten(TOKENIZE(\*)) as words;**

**grunt> dump B;**



```
Tue 04/26
jairam@ubuntu: ~

File Edit View Search Terminal Help
(against accepting unsolicited donations from donors in such states who)
(approach us with offers to donate.)
()
(International donations are gratefully accepted, but we cannot make)
(any statements concerning tax treatment of donations received from)
(outside the United States. U.S. laws alone swamp our small staff.)
()
(Please check the Project Gutenberg Web pages for current donation)
(methods and addresses. Donations are accepted in a number of other)
(ways including checks, online payments and credit card donations.)
(To donate, please visit: http://pglaf.org/donate)
()
()
()
(Section 5. General Information About Project Gutenberg-tm electronic)
(works.)
()
(Professor Michael S. Hart is the originator of the Project Gutenberg-tm)
(concept of a library of electronic works that could be freely shared)
(with anyone. For thirty years, he produced and distributed Project)
(Gutenberg-tm eBooks with only a loose network of volunteer support.)
()
()
(Project Gutenberg-tm eBooks are often created from several printed)
(editions, all of which are confirmed as Public Domain in the U.S.)
(unless a copyright notice is included. Thus, we do not necessarily)
(keep eBooks in compliance with any particular paper edition.)
()
()
(Most people start at our Web site which has the main PG search facility:)
()
(
    http://www.gutenberg.org
)
()
(This Web site includes information about Project Gutenberg-tm,)
(including how to make donations to the Project Gutenberg Literary)
(Archive Foundation, how to help produce our new eBooks, and how to)
(subscribe to our email newsletter to hear about new eBooks.)
grunt> B = foreach A generate Flatten(TOKENIZE(*)) as words;
grunt> dump B;
```

5. Command: **grunt >C = group B by words**

**grunt > dump C;**



```
set ...
"illustrate" ...
"\\" ...
"run" ...
"exec" ...
"\$default" ...
"\$declare" ...
"scriptDone" ...
"" ...
"" ...
<EOL> ...
";" ...

Details at logfile: /home/jairam/pig_1596539989619.log
grunt> C = group B by words;
grunt> dump C;
```

6. Command: `grunt >forceach c generate group , COUNT(B);`

*grunt > dump D;*

The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "jelram@ubuntu:~". The terminal content displays the output of a Hadoop WordCount job. It includes metrics like total records written (44110), total bytes written (0), and spill counts (0). It also shows the DAG for the job and log entries from 2020-08-04 at 04:37:04.992, indicating issues with initializing JVM Metrics. The final log entry shows success for the MapReduceLauncher.

```
Counters:
Total records written : 44110
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1587939684_0005

2020-08-04 04:37:04.992 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2020-08-04 04:37:04.992 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2020-08-04 04:37:04.993 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2020-08-04 04:37:05.004 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.MapReduceLayer.MapReduceLauncher - Success!
grunt>
```

A screenshot of a Linux desktop environment, likely Ubuntu, showing a terminal window and a file manager window.

The terminal window (Activities Text Editor) displays the following output:

```
I 3
G 5
H 5
I 2455
J 13
K 2
L 1
M 5
O 240
P 3
R 3
S 6
T 2
U 1
W 4
X 4
Y 1
a 2384
b 5972
c 3
d 1
e 10
h 1
l 10
m 8
n 2
o 3
p 2
r 1
s 4
w 1
x 7
y 6
z 1
S 1
$S 1
&c 2
```

The file manager window shows the following files in the current directory:

- WordCount.class
- wc.jar
- part-r-00000 (selected)

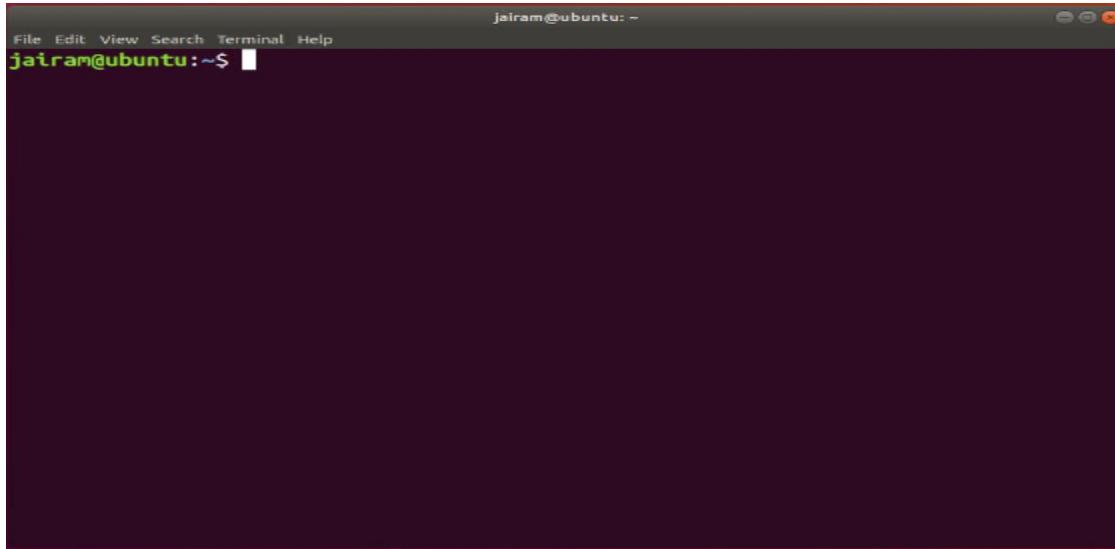
The status bar at the bottom of the terminal window indicates:

- Matlab
- Tab Width: 8
- Ln 1, Col 1
- INS

A tooltip "part-r-00000 selected (461.2 kB)" is visible near the selected file in the file manager.

## Data compression using Hive

### 1. Open terminal



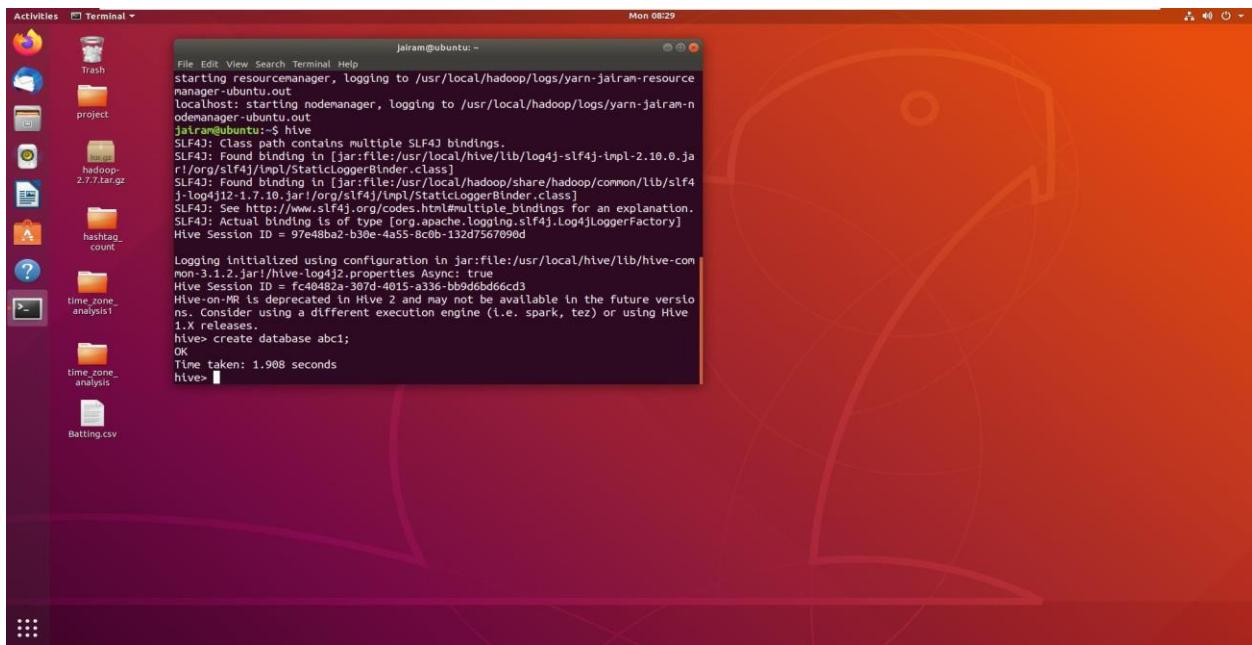
### 2. Command: *hive* for launching hive

```
jairam@ubuntu: ~
File Edit View Search Terminal Help
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-jairam-secondarynamenode-ubuntu.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-jairam-resource
manager-ubuntu.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-jairam-n
odemanager-ubuntu.out
jairam@ubuntu:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.ja
r!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4
j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 97e48ba2-b38e-4a55-8c0b-132d7567090d

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-com
mon-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fc40482a-307d-4015-a336-bb9d6bd66cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive> 
```

### 3. To create database...

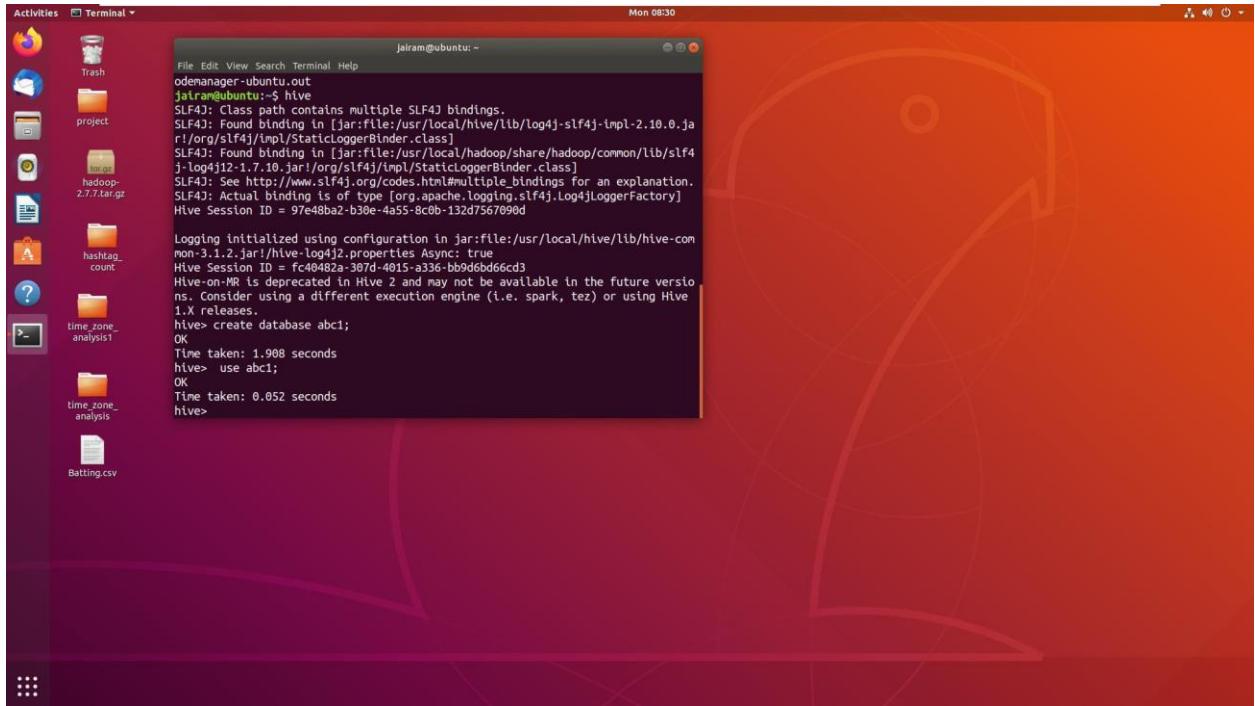
Command: *create database abc;*



```
jalram@ubuntu:~$ start-stop-daemon --start --background --pidfile /var/run/yarn-jalram-resource-manager-ubuntu.out --exec /usr/local/hadoop/bin/yarn --resource-manager
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-jalram-node-ubuntu.out
jalram@ubuntu:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4JLoggerFactory]
Hive Session ID = 97e48ba2-b30e-4a55-8c0b-132d7567090d

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.1-hadoop2-properties Async: true
Hive Session ID = fc49498f-07d4-4115-a336-bb9dd6d6cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> create database abc;
OK
Time taken: 1.908 seconds
hive>
```

### 4. Command: use abc:

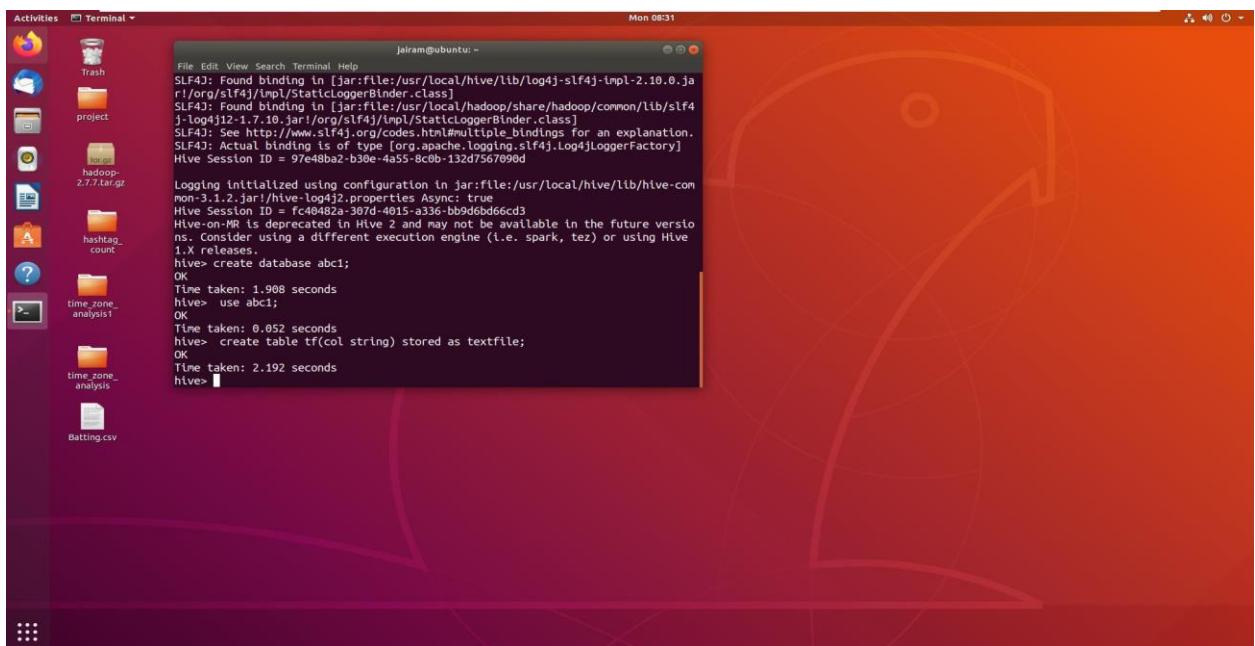


A screenshot of an Ubuntu desktop environment. On the left is a vertical dock with icons for the Dash, Home, Applications, and Activities. A terminal window titled 'Terminal' is open in the Activities dock, showing the following command-line session:

```
File Edit View Search Terminal Help
jairam@ubuntu:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4JLoggerFactory]
Hive Session ID = 97e48ba2-b30e-4a55-8c0b-132d7567090d

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-2.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fc40482a-307d-4015-a336-bb9d6bd6cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> create database abc1;
OK
Time taken: 1.908 seconds
hive> use abc1;
OK
Time taken: 0.052 seconds
hive>
```

## 5. Command: *create table tf(col string) stored as textfile;*



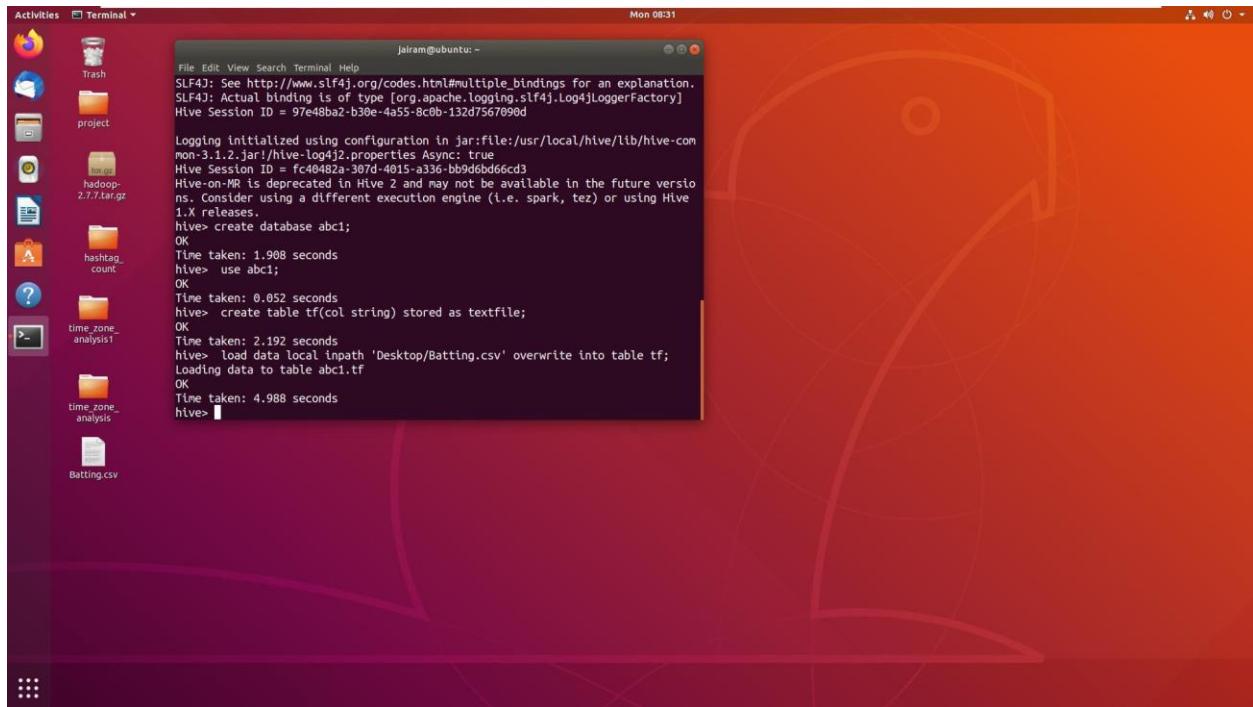
A screenshot of an Ubuntu desktop environment. On the left is a vertical dock with icons for the Dash, Home, Applications, and Activities. A terminal window titled 'Terminal' is open in the Activities dock, showing the following command-line session:

```
File Edit View Search Terminal Help
jairam@ubuntu:~$ hive
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4JLoggerFactory]
Hive Session ID = 97e48ba2-b30e-4a55-8c0b-132d7567090d

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-2.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fc40482a-307d-4015-a336-bb9d6bd6cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> create database abc1;
OK
Time taken: 1.908 seconds
hive> use abc1;
OK
Time taken: 0.052 seconds
hive> create table tf(col string) stored as textfile;
OK
Time taken: 2.192 seconds
hive> |
```

## 6. To create optimized row column file...

Command: *load data local inpath 'Desktop/Batting.csv' overwrite into table tf;*



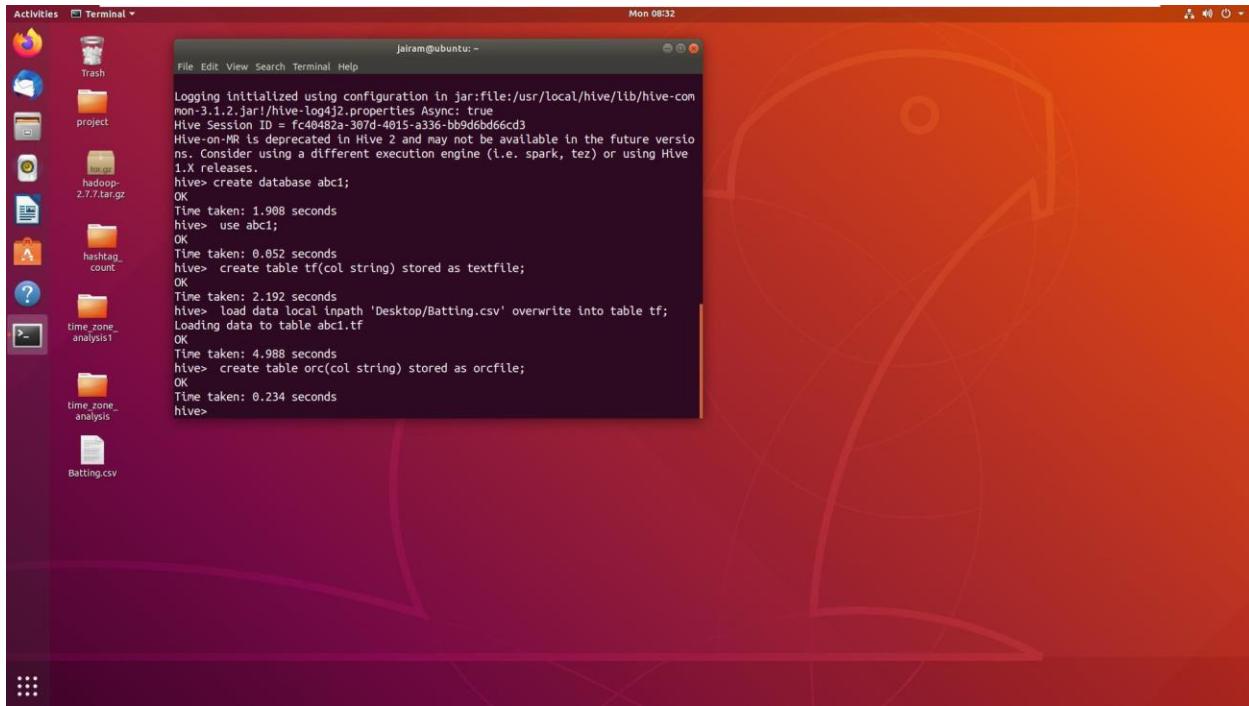
A screenshot of a Linux desktop environment, likely Ubuntu, showing a terminal window titled "Terminal". The terminal window displays the following Hive session:

```
jairam@ubuntu: ~
File Edit View Search Terminal Help
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4JLoggerFactory]
Hive Session ID = 97e48ba2-b30e-4a55-8c0b-132d7567090d

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-com
mon-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fc40482a-307d-4015-a336-bb9d6bd66cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive> create database abc1;
OK
Time taken: 1.908 seconds
hive> use abc1;
OK
Time taken: 0.052 seconds
hive> create table tf(col string) stored as textfile;
OK
Time taken: 2.192 seconds
hive> load data local inpath 'Desktop/Batting.csv' overwrite into table tf;
Loading data to table abc1.tf
OK
Time taken: 4.988 seconds
hive> |
```

7. you can't load direct data in orc table so by using this command you can insert data in orc table

command: *create table orc(col string) stored as orcfile;*



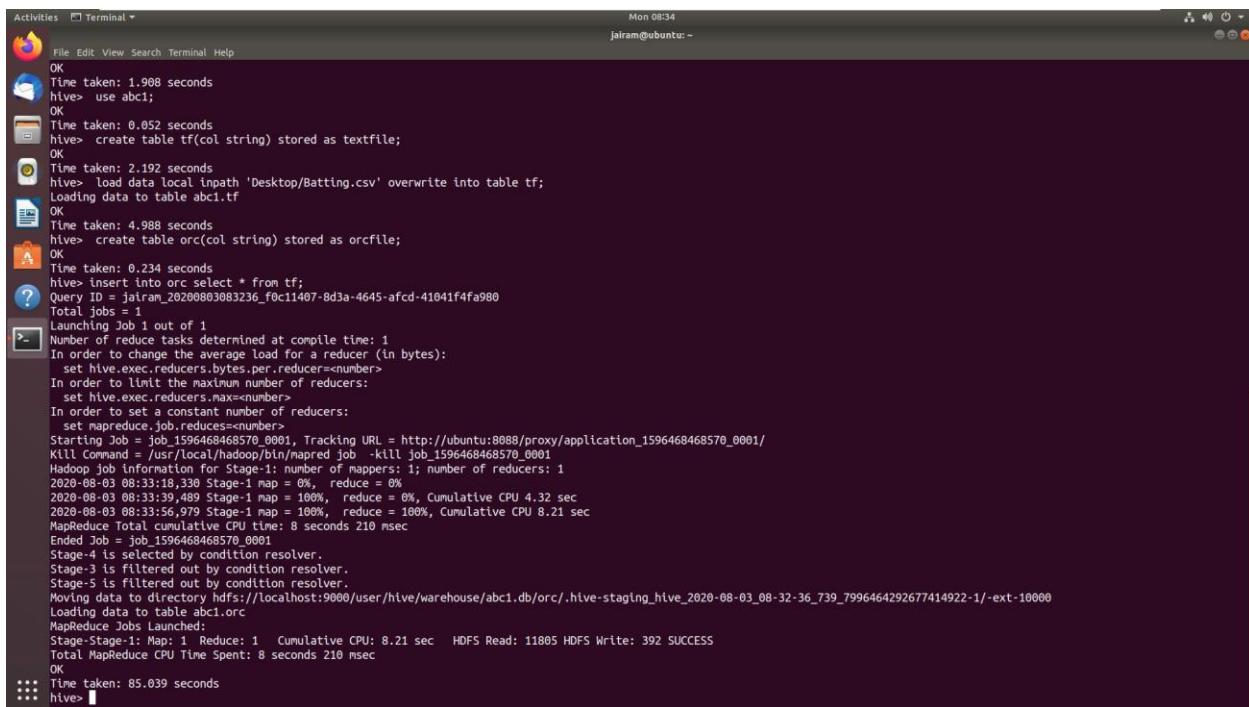
A screenshot of an Ubuntu desktop environment. On the left, there's a dock with icons for Dash, Home, Trash, project, hadoop-2.7.7.tar.gz, hashtag\_count, time\_zone\_analysis1, time\_zone\_analysis2, and Batting.csv. A terminal window titled 'Terminal' is open in the center, showing the following Hive session:

```

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fc40482a-307d-4015-a336-bb9d6bd6cd3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> create database abc1;
OK
Time taken: 1.908 seconds
hive> use abc1;
OK
Time taken: 0.052 seconds
hive> create table tf(col string) stored as textfile;
OK
Time taken: 2.192 seconds
hive> load data local inpath 'Desktop/Batting.csv' overwrite into table tf;
Loading data to table abc1.tf
OK
Time taken: 4.988 seconds
hive> create table orc(col string) stored as orcfile;
OK
Time taken: 0.234 seconds
hive>

```

## 8. command: *insert into orc select \* from tf;*



A screenshot of an Ubuntu desktop environment. On the left, there's a dock with icons for Dash, Home, Trash, project, hadoop-2.7.7.tar.gz, hashtag\_count, time\_zone\_analysis1, time\_zone\_analysis2, and Batting.csv. A terminal window titled 'Terminal' is open in the center, showing the following Hive session and MapReduce job output:

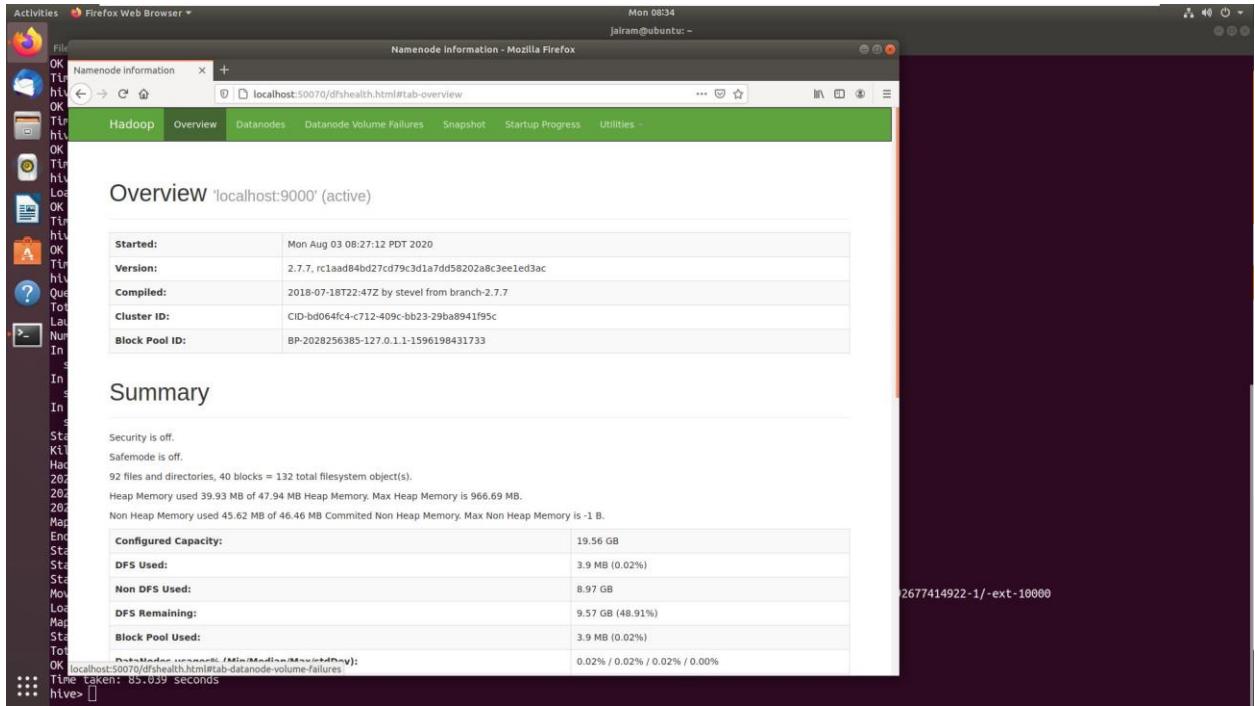
```

Mon 08:34
jairam@ubuntu: ~

File Edit View Search Terminal Help
OK
Time taken: 1.908 seconds
hive> use abc1;
OK
Time taken: 0.052 seconds
hive> create table tf(col string) stored as textfile;
OK
Time taken: 2.192 seconds
hive> load data local inpath 'Desktop/Batting.csv' overwrite into table tf;
Loading data to table abc1.tf
OK
Time taken: 4.988 seconds
hive> create table orc(col string) stored as orcfile;
OK
Time taken: 0.234 seconds
hive> Insert into orc select * from tf;
Query ID = jairam_20200803083236_f0c11407-8d3a-4645-afcd-41041f4fa980
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1596468468570_0001, Tracking URL = http://ubuntu:8088/proxy/application_1596468468570_0001/
Kill Command = /usr/local/hadoop/bin/mapred job -kill job_1596468468570_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-03 08:33:18,330 Stage-1 map = 0%, reduce = 0%
2020-08-03 08:33:39,489 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.32 sec
2020-08-03 08:33:56,979 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.21 sec
MapReduce Total cumulative CPU time: 8 seconds 210 msec
Ended Job = job_1596468468570_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/abc1.db/orc/.hive-staging_hive_2020-08-03_08-32-36_739_7996464292677414922-1/-ext-10000
Loading data to table abc1.orc
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.21 sec HDFS Read: 11805 HDFS Write: 392 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 210 msec
OK
Time taken: 85.039 seconds
hive>

```

## 9. now open browser then type localhost:50070



## 10.then utilities

## 11.then browse the file system

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwx-----	jairam	supergroup	0 B	8/2/2020, 10:52:18 PM	0	0 B	tmp
drwxr-xr-x	jairam	supergroup	0 B	8/2/2020, 1:31:41 AM	0	0 B	top5rating
drwxr-xr-x	jairam	supergroup	0 B	8/1/2020, 10:51:49 PM	0	0 B	uber2
drwxr-xr-x	jairam	supergroup	0 B	8/1/2020, 11:18:51 PM	0	0 B	uber3
drwxr-xr-x	jairam	supergroup	0 B	8/1/2020, 11:25:22 PM	0	0 B	uberOP
drwxr-xr-x	jairam	supergroup	0 B	8/1/2020, 10:56:09 PM	0	0 B	uberoutput
drwxr-xr-x	jairam	supergroup	0 B	8/3/2020, 4:15:32 AM	0	0 B	user
drwxr-xr-x	jairam	supergroup	0 B	8/2/2020, 1:28:01 AM	0	0 B	youtube
drwxr-xr-x	jairam	supergroup	0 B	8/2/2020, 1:35:38 AM	0	0 B	youtube3

12.then

**/user/hive/warehouse/abc.db/tf** //// here check the size of file

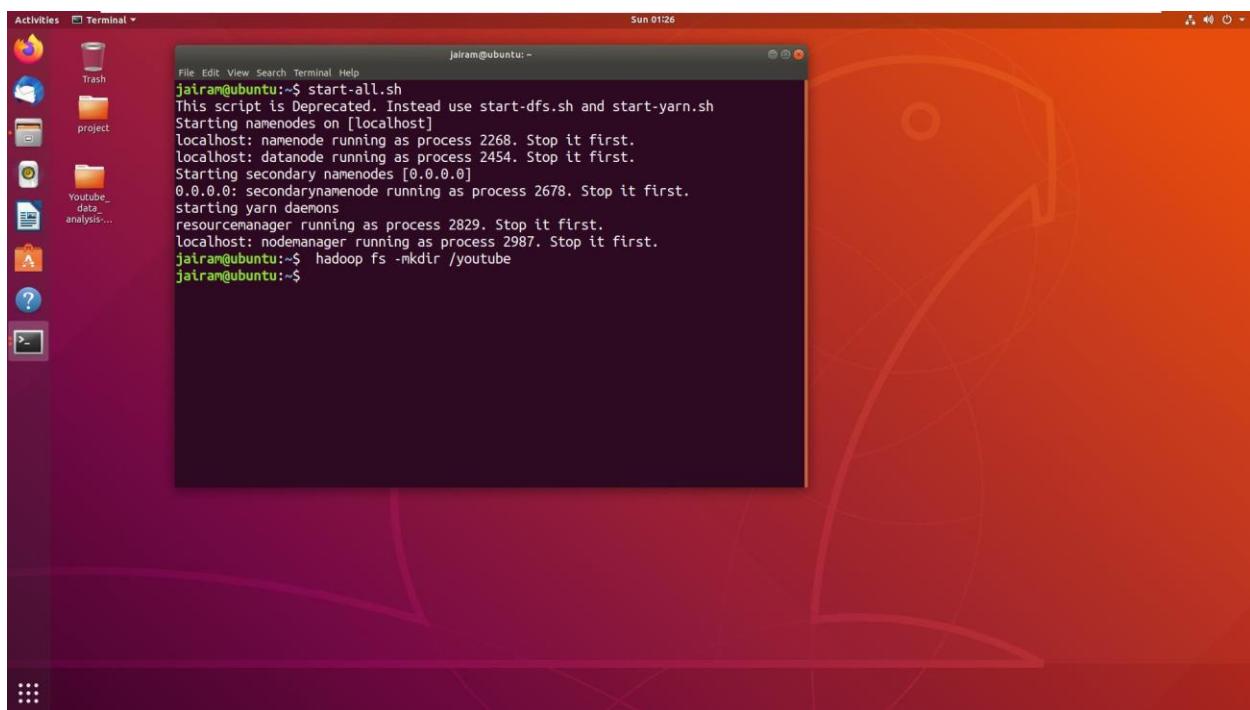
**/user/hive/warehouse/abc.db/orc** //// now check the compressed form of data the data is now converted into binary format and also its size also decreases.

## Youtube data analysis

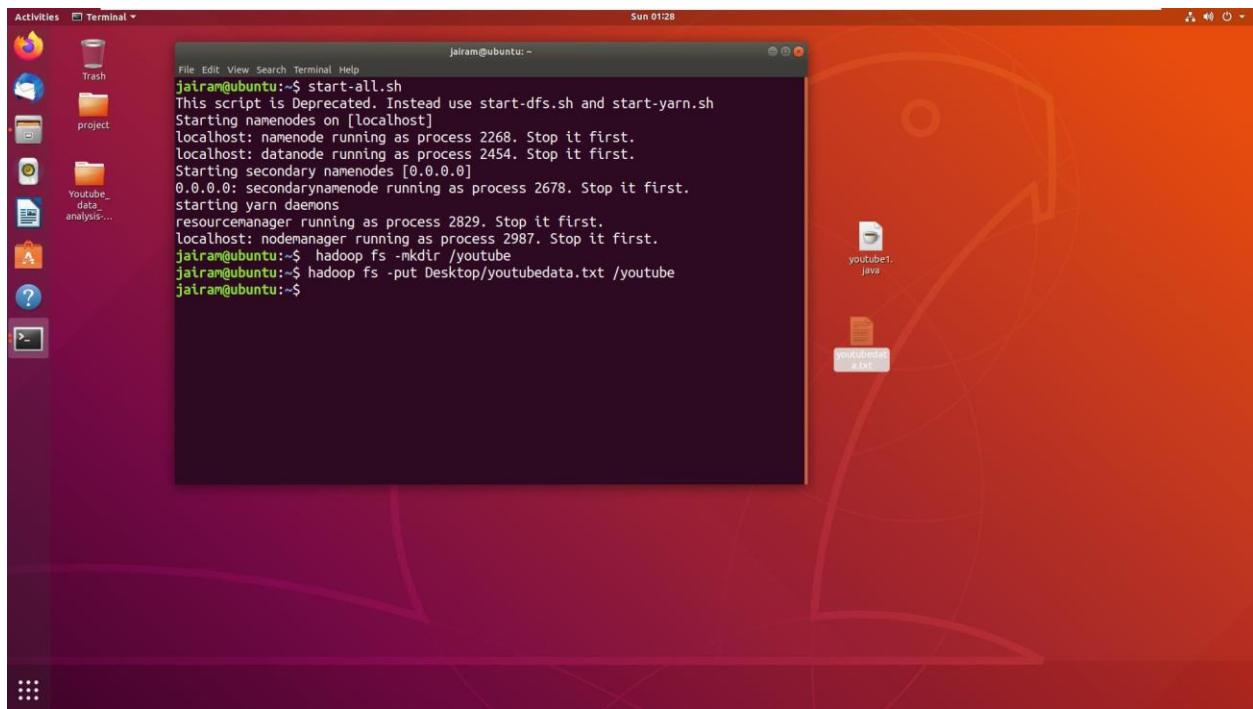
1. Open terminal and type Command: *start-all.sh*

```
jairam@ubuntu:~$ start-all.shh
start-all.shh: command not found
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2534. Stop it first.
localhost: datanode running as process 2755. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3016. Stop it first.
starting yarn daemons
resourcemanager running as process 3239. Stop it first.
localhost: nodemanager running as process 3396. Stop it first.
jairam@ubuntu:~$
```

2. command: *hadoop fs -mkdir /youtube*



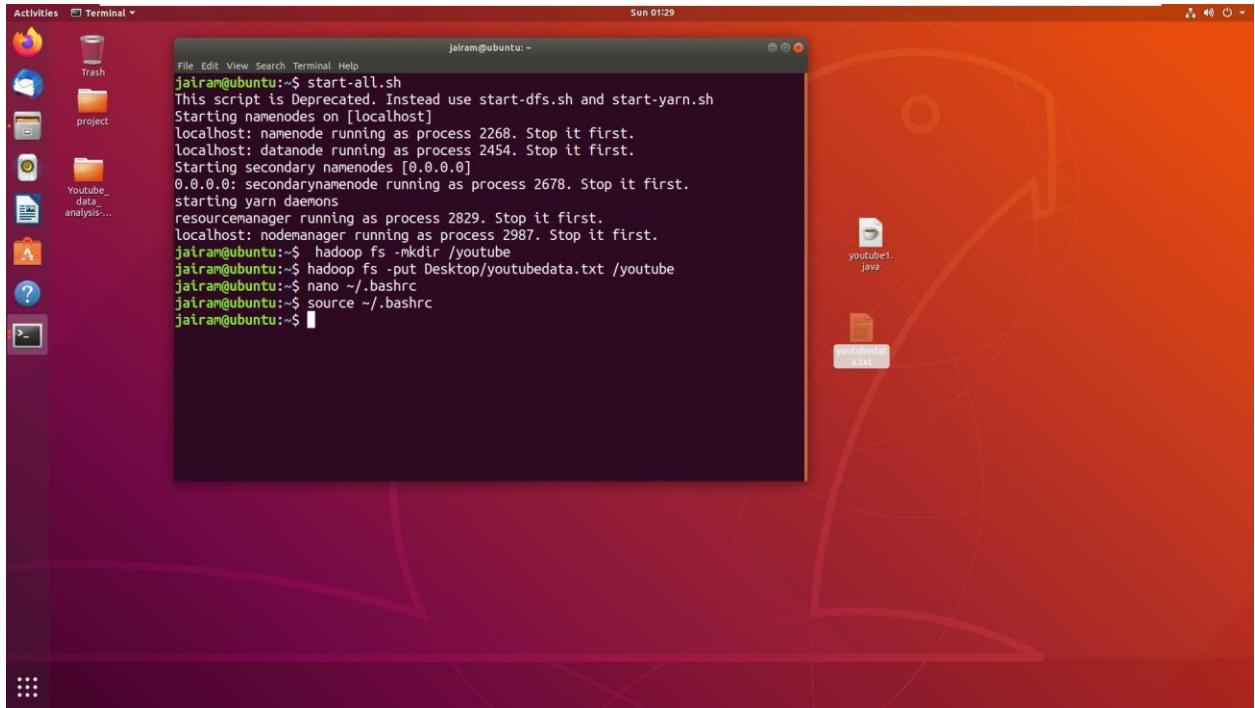
3. command: ***hadoop fs -put Desktop/youtubedata.txt /youtube***



4. Open bashrc file by ***nano ~/.bashrc*** and type

```
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
```

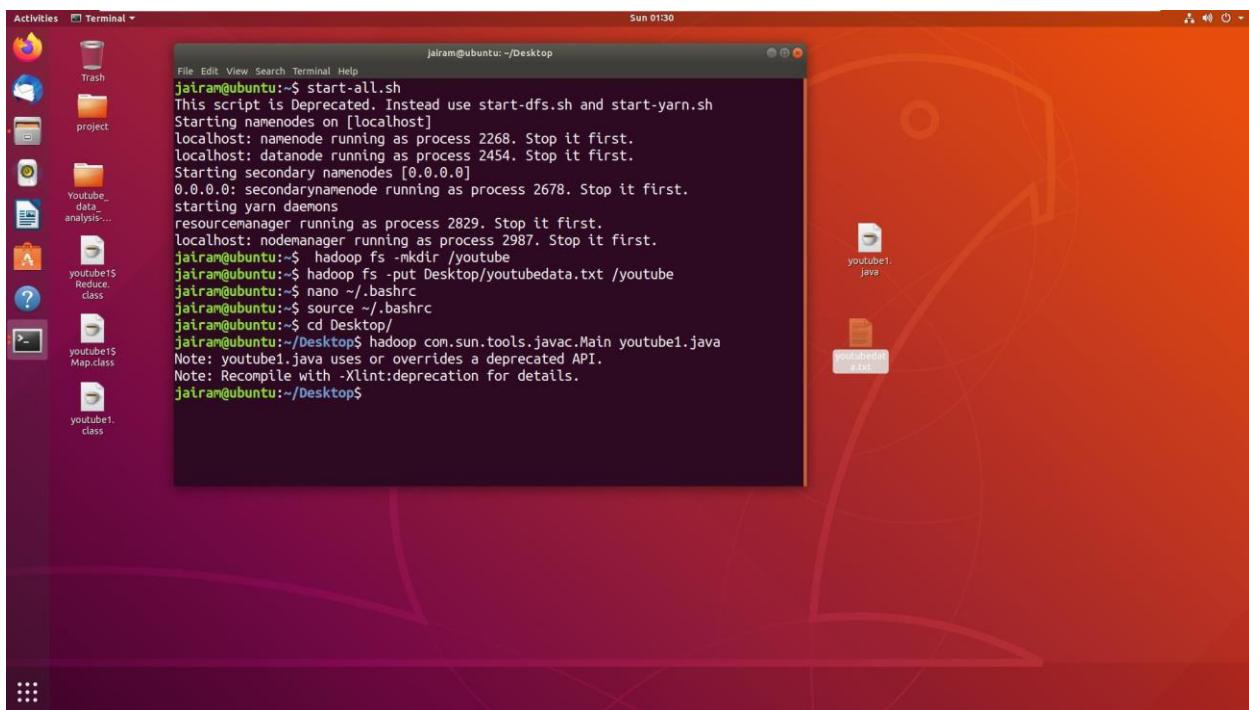
5. Then commnd: ***source ~/.bashrc***



6. Command: ***cd Desktop***

7. Now compile youtube1.java file suppose that file is on Desktop

Command: ***hadoop com.sun.tools.javac.Main youtube1.java***

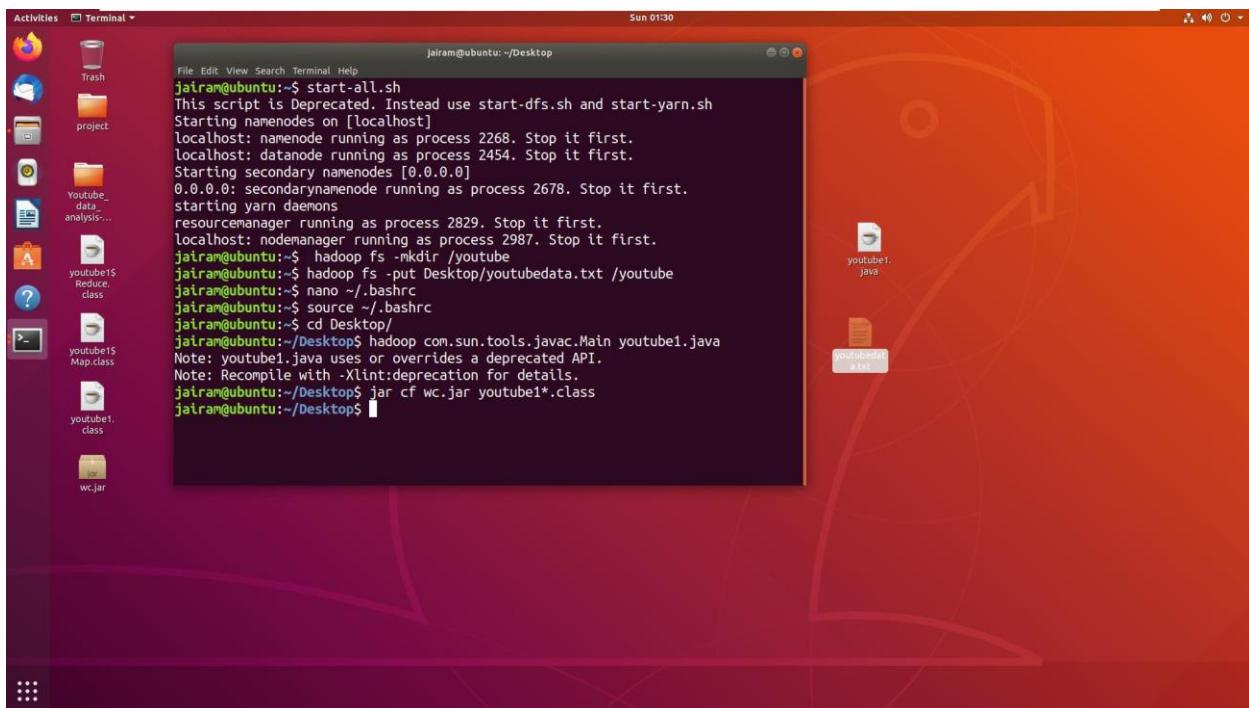


A screenshot of an Ubuntu desktop environment. On the left, there's a dock with icons for Dash, Home, Applications, and Activities. A terminal window titled "Terminal" is open in the Activities dock, showing the following command-line session:

```
jairam@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: namenode running as process 2268. Stop it first.
localhost: datanode running as process 2454. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 2678. Stop it first.
starting yarn daemons
resourcemanager running as process 2829. Stop it first.
localhost: nodemanager running as process 2987. Stop it first.
jairam@ubuntu:~$ hadoop fs -mkdir /youtube
jairam@ubuntu:~$ nano ~/.bashrc
jairam@ubuntu:~$ source ~/.bashrc
jairam@ubuntu:~$ cd Desktop/
jairam@ubuntu:~/Desktop$ hadoop com.sun.tools.javac.Main youtube1.java
Note: youtube1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
jairam@ubuntu:~/Desktop$
```

## 8. To combine all class files

Command; ***jar cf wc.jar youtube1\*.class***

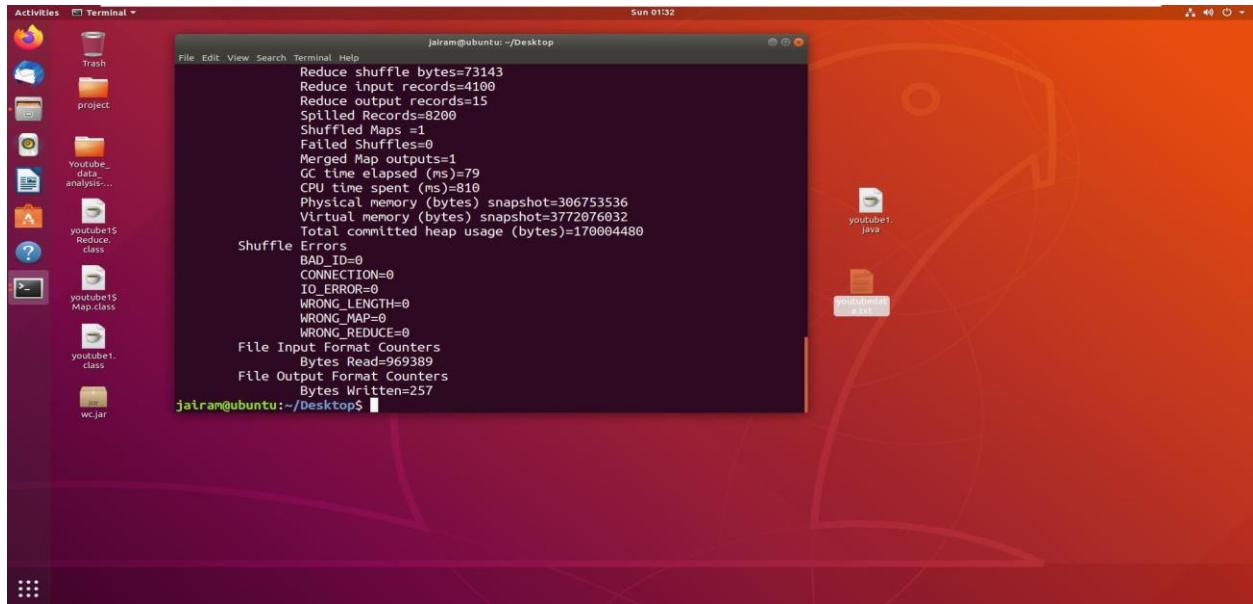


A screenshot of an Ubuntu desktop environment, similar to the one above. The terminal window shows the same initial setup and Hadoop startup logs. However, at the end of the session, the user runs the command:

```
jairam@ubuntu:~/Desktop$ jar cf wc.jar youtube1*.class
```

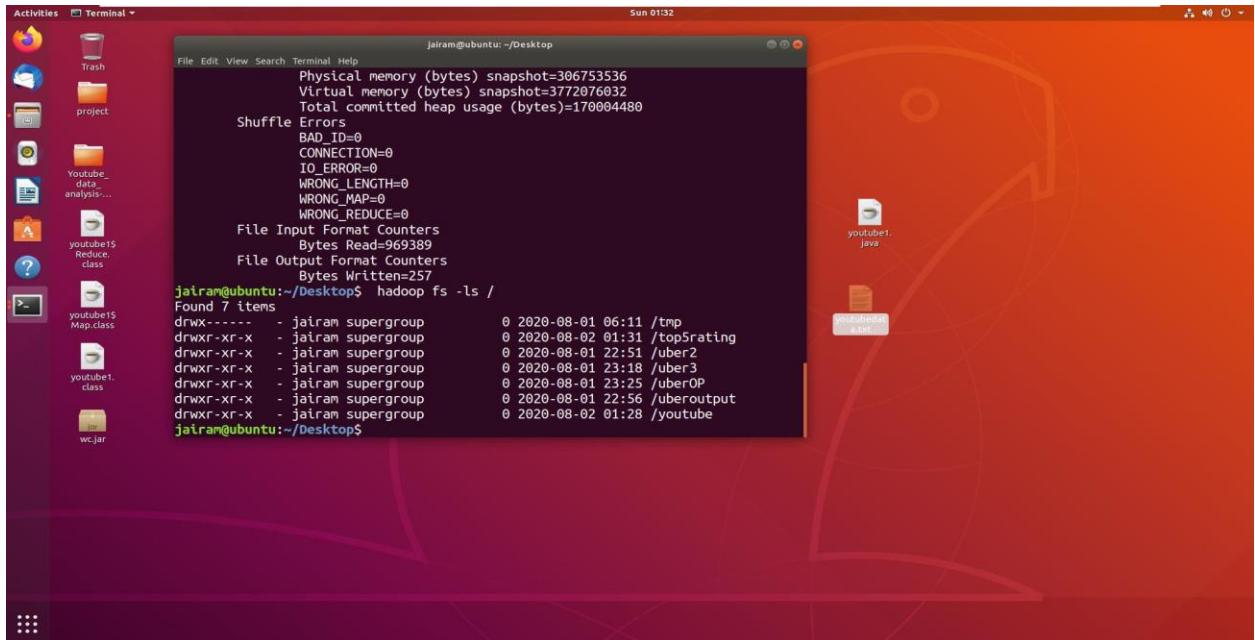
## 9. To execute

Command: ***hadoop jar wc.jar youtube1 /youtube/youtubedata.txt /top5rating***

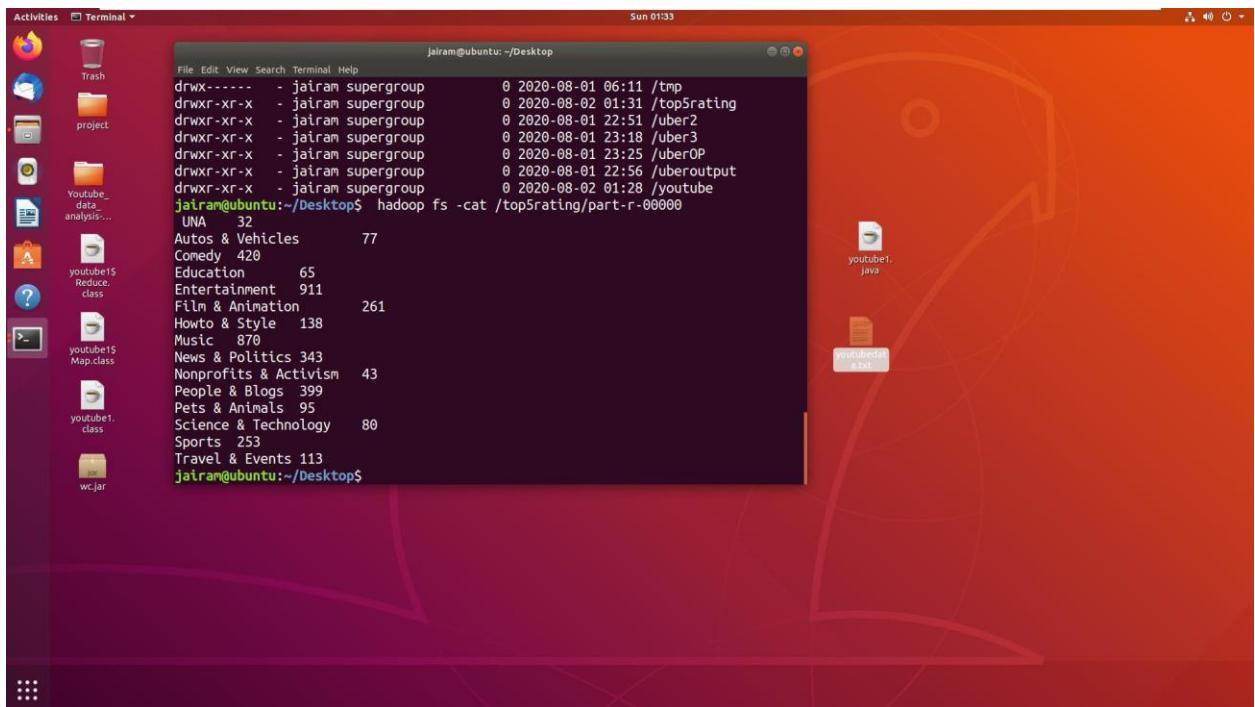


## 10. To check output folder is there or not

Command: ***hadoop fs -ls /***



11.command: hadoop fs -cat /top5rating/part-r-00000



## **Bibliography And Reference Taken.**

**Jairam J S(2020)** - Beginner's Guide to Big data

Project Workbook Of Big data by cosmic skill.