LABORATORY RECORD

On

BIG DATA ANALYTICS

B.E (IT) – VII Sem

By

N.Durga Sai Lakshmi(160117737006)

Under the guidance of

Sri.K.Gangadhar Rao

Assistant Professor,

Dept. of IT,CBIT.

DEPARTMENT OF INFORMATION TECHNOLOGY

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY (A)

(Affiliated to Osmania University; Accredited by NBA(AICTE) and NAAC(UGC), ISO Certified 9001:2015)

KOKAPET(V),GANDIPET(M), RR District HYDERABAD -75

Website: www.cbit.ac.in

2020-2021

160117737006

N.Durga Sai Lakshmi

Problem Statement-1: Write a Map-reduce application to find number of occurrences of each word from the given dataset.

Description:

In the MapReduce word count example, we find out the frequency of each word. Here, the role of Mapper is to map the keys to the existing values and the role of Reducer is to aggregate the keys of common values. So, everything is represented in the form of a Key-value pair.

Procedure:

```
$cd hadoop-3.2.1/

$cd sbin/

$cd start-all.sh

$jps

$cd..

$cd wc.py/

$hadoop fs -mkdir -p /wordcount/p1

$hadoop fs -copyFromLocal word.txt /wordcount/p1

$hadoop jar /home/hduser/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -file wcmap.py -mapper wcmap.py -file wcred.py -reducer wcred.py -input /wordcount/p1 -output /wordcount/p1/output

$hadoop fs -cat /wordcount/p1/output/part-00000
```

Code:

wcmap.py

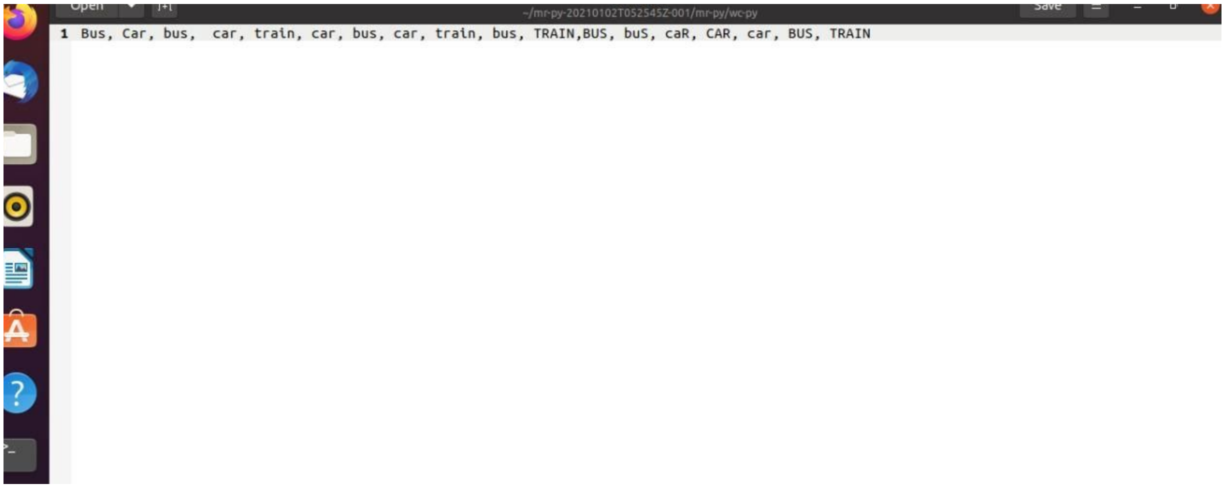
```
import sys

# input comes from STDIN (standard input)

for line in sys.stdin:

        # remove leading and trailing whitespace
        line = line.strip()

        # split the line into words

        words = line.split()

        # increase counters

        for word in words:

        # write the results to STDOUT (standard output);

        # what we output here will be the input for the

        # Reduce step, i.e. the input for reducer.py

        #

        # tab-delimited; the trivial word count is 1

                print '%s\t%s' % (word, 1)
```

wcred.py

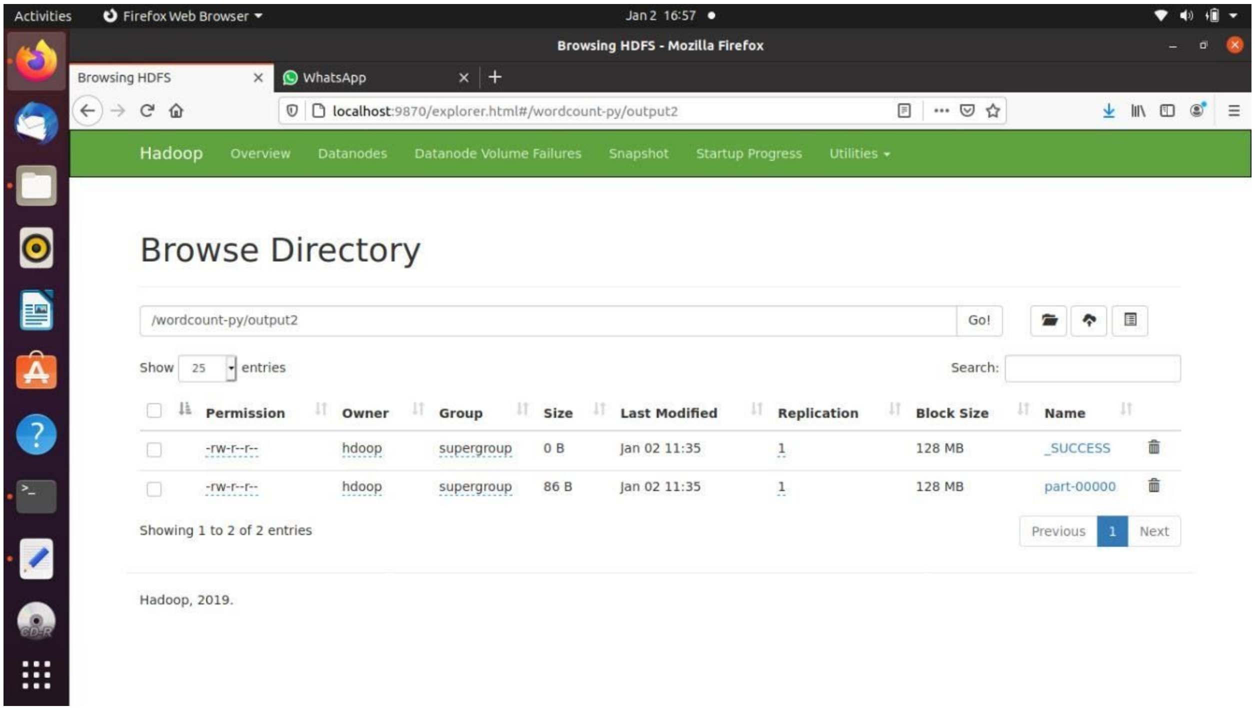
```
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
        # remove leading and trailing whitespace
        line = line.strip()

        # parse the input we got from mapper.py
        word, count = line.split('\t', 1)

        # convert count (currently a string) to int
        try:
```

```
        count = int(count)
        except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

        # this IF-switch only works because Hadoop sorts map output
        # by key (here: word) before it is passed to the reducer
        if current_word == word:
        current_count += count
        else:
        if current_word:
        # write result to STDOUT
        print '%s\t%s' % (current_word, current_count)
        current_count = count
        current_word = word
if current_word == word:
        print '%s\t%s' % (current_word, current_count)
```

INPUT :



```
Bus, Car, bus,  car, train, car, bus, car, train, bus, TRAIN,BUS, buS, caR, CAR, car, BUS, TRAIN
```



Browse Directory

/wordcount-py/output2 Go!

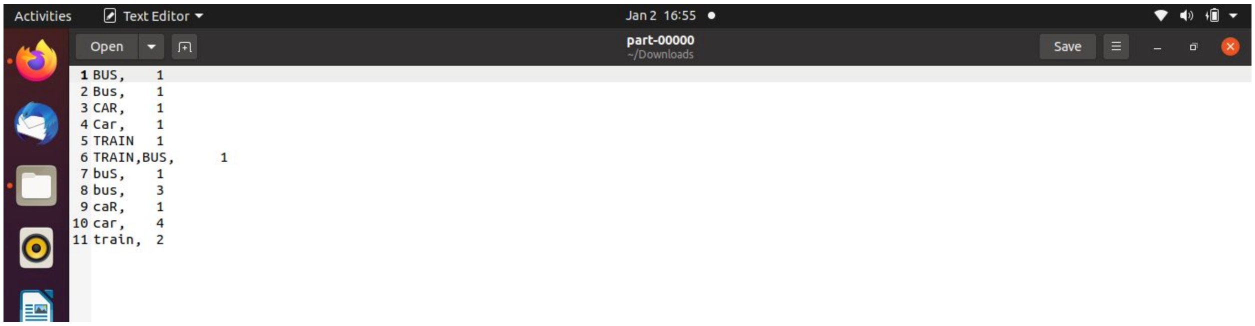
Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdoop	supergroup	0 B	Jan 02 11:35	1	128 MB	_SUCCESS
-rw-r--r--	hdoop	supergroup	86 B	Jan 02 11:35	1	128 MB	part-00000

Showing 1 to 2 of 2 entries

Hadoop, 2019.

OUTPUT:



```
BUS,    1
Bus,    1
CAR,    1
Car,    1
TRAIN   1
TRAIN,BUS,      1
buS,    1
bus,    3
caR,    1
car,    4
train,  2
```

Problem Statement-2: Write a 3map-reduce application to predict maximum stock price for given dataset.

Description:

We are trying to find out the maximum closing price of each stock symbol. This means that we have to group the records by symbol so that we can calculate the maximum closing price by symbol. So we will output Stock Symbol as the key and close price as the value for each record. We now know what is going to be the Map's input and what is going to be the maps output.

Procedure:

$cd hadoop-3.2.1/

```
$cd sbin/

$cd start-all.sh

$jps

$cd..

$cd stock.py/

$hadoop fs -mkdir /stocksip

$hadoop fs -copyFromLocal stocks.txt  /stocksip

$hadoop jar /home/hduser/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -file
stock-mapper.py  -mapper stock-mapper.py  -file stock-reducer.py -reducer stock-reducer.py
-input /stocksip/stocks.txt -output /stocksout/output

$hadoop fs –cat stocksout/output/part-00000
```

Code:

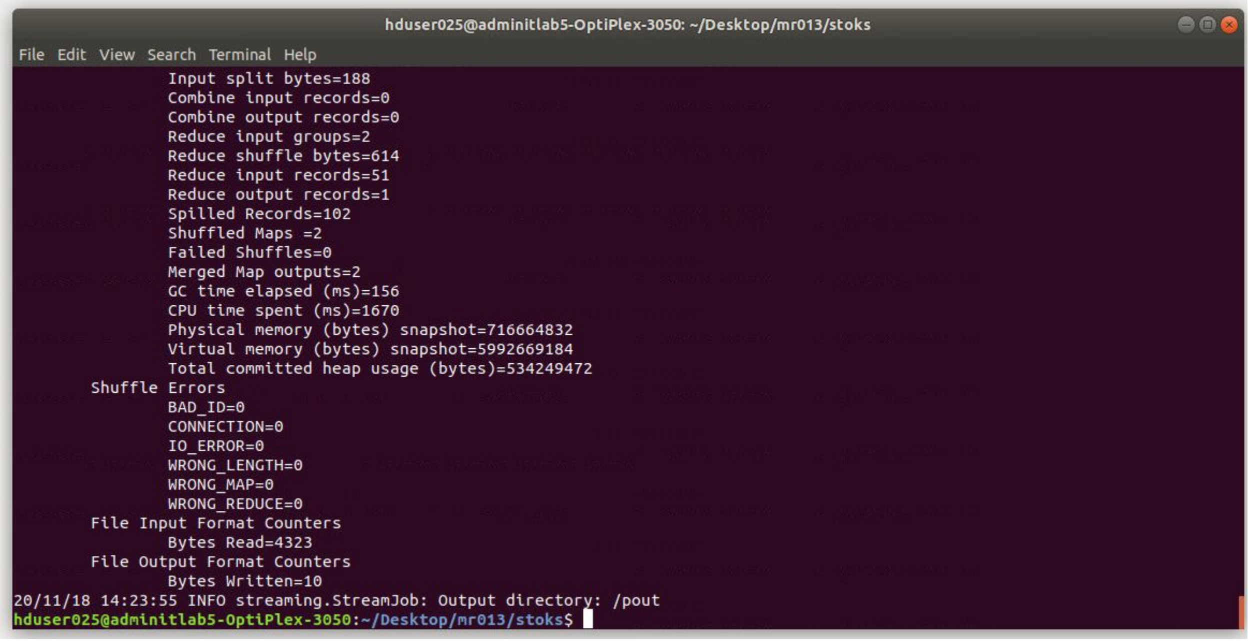
Stock-mapper.py

```
import sys

for line in sys.stdin:
        line = line.strip()
        data = line.split(",")
        stock, price = data[1], data[6]
        print("%s\t%s"%(stock,price))
```

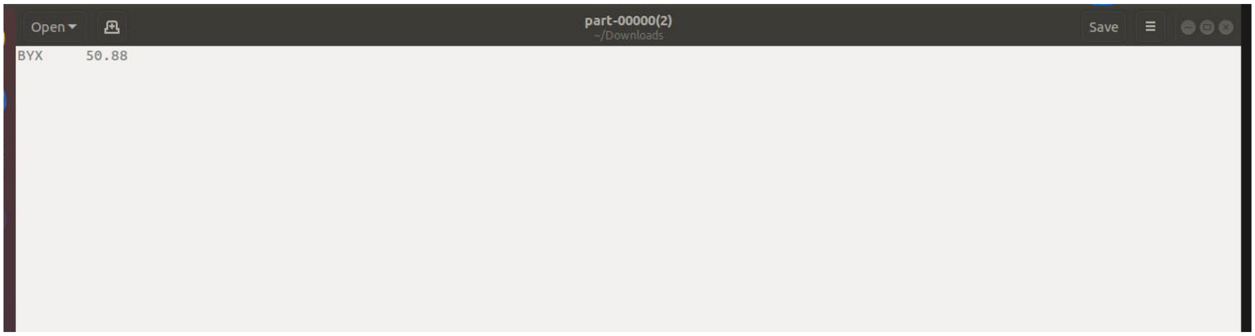
stock-reducer.py

```
import sys

max_price = 9999
max_stock = None

for line in sys.stdin:
        line = line.strip()
        stock,price = line.split("\t",1)
        if max_stock and max_stock!=stock:
                if max_price > price:
                        max_price = price
                        max_stock = stock

        else:
                max_stock, max_price = stock,max(max_price,price)

if max_stock:
        if max_price > price:
                        max_price = price
                        max_stock = stock

print("%s\t%s"%(max_stock,max_price))
```

OUTPUT:


hduser025@adminitlab5-OptiPlex-3050: ~/Desktop/mr013/stoks
File Edit View Search Terminal Help

```
                Input split bytes=188
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=614
                Reduce input records=51
                Reduce output records=1
                Spilled Records=102
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=156
                CPU time spent (ms)=1670
                Physical memory (bytes) snapshot=716664832
                Virtual memory (bytes) snapshot=5992669184
                Total committed heap usage (bytes)=534249472
        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=4323
        File Output Format Counters
                Bytes Written=10
20/11/18 14:23:55 INFO streaming.StreamJob: Output directory: /pout
hduser025@adminitlab5-OptiPlex-3050:~/Desktop/mr013/stoks$
```



```
BYX	50.88
```

160117737006
N.Durga Sai Lakshmi

Problem Statement-3: Write a map-reduce application to find maximum temperature for a given year from the NCDC weather dataset.

Description:

MapReduce is based on set of key value pairs. So first we have to decide on the types for the key/value pairs for the input.

Map Phase: The input for Map phase is set of weather data files as shown in snap shot. The types of input key value pairs are *LongWritable* and Text and the types of output key value pairs are *Text* and *IntWritable*. Each Map task extracts the temperature data from the given year file. The output of the map phase is set of key value pairs. Set of keys are the years. Values are the temperature of each year.

Reduce Phase: Reduce phase takes all the values associated with a particular key. That is all the temperature values belong to a particular year is fed to a same reducer. Then each reducer finds the highest recorded temperature for each year. The types of output key value pairs in Map phase is same for the types of input key value pairs in reduce phase (*Text* and *IntWritable*). The types of output key value pairs in reduce phase is too *Text* and *IntWritable*.

Procedure:

```
$cd hadoop-3.2.1/

$cd sbin/

$cd start-all.sh

$jps

$cd..

$cd temp.py/

$ hadoop fs -mkdir /NCDCWeatherData

$hadoop fs -copyFromLocal NCDCWeatherData/*  /NCDCWeatherData

$hadoop jar /home/hduser/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -file
temp-mapper.py  -mapper temp-mapper.py  -file temp-reducer.py -reducer temp-reducer.py
-input /NCDCWeatherData -output /weatherout/output
```

```
$ hadoop fs –cat weatherout /output/part-00000
```

Code:

temp-mapper.py

```
import re
import sys
for line in sys.stdin:
val = line.strip()
(year, temp, q) = (val[15:19], val[87:92], val[92:93])
if (temp != "+9999" and re.match("[01459]", q)):
print "%st%s" % (year, temp)
```

temp-reducer.py

```
import sys
(last_key, max_val) = (None, 0)
for line in sys.stdin:
(key, val) = line.strip().split("t")
if last_key and last_key != key:
print "%st%s" % (last_key, max_val)
(last_key, max_val) = (key, int(val))
else:
(last_key, max_val) = (key, max(max_val, int(val)))
if last_key:
print "%st%s" % (last_key, max_val)
```

Expected Output:

1901 year has maximum temperature.

Problem Statement-4: Write a map-reduce application to find how many times a particular page has been accessed (use from the Apache Web Server log data).

Description:

In today’s world the usage of internet has become very high and using all the logs from web server we can actually predict the customer moods in buying the product or can analyze the interests of Customer.

Procedure:

```
$cd hadoop-3.2.1/
```

```
$cd sbin/

$cd start-all.sh

$jps

$cd..

$cd web-log.py/

$hadoop fs -mkdir -p /weblogip

$hadoop fs -copyFromLocal test_access_log  /weblogip
$hadoop jar /home/hduser/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -file
my_mapper_by_page.py  -mapper my_mapper_by_page.py  -file my_reducer.py -reducer
my_reducer.py -input /weblogip/test_access_log -output /weblogout/output

$ hadoop fs –cat weblogout /output/part-00000
```

Code:

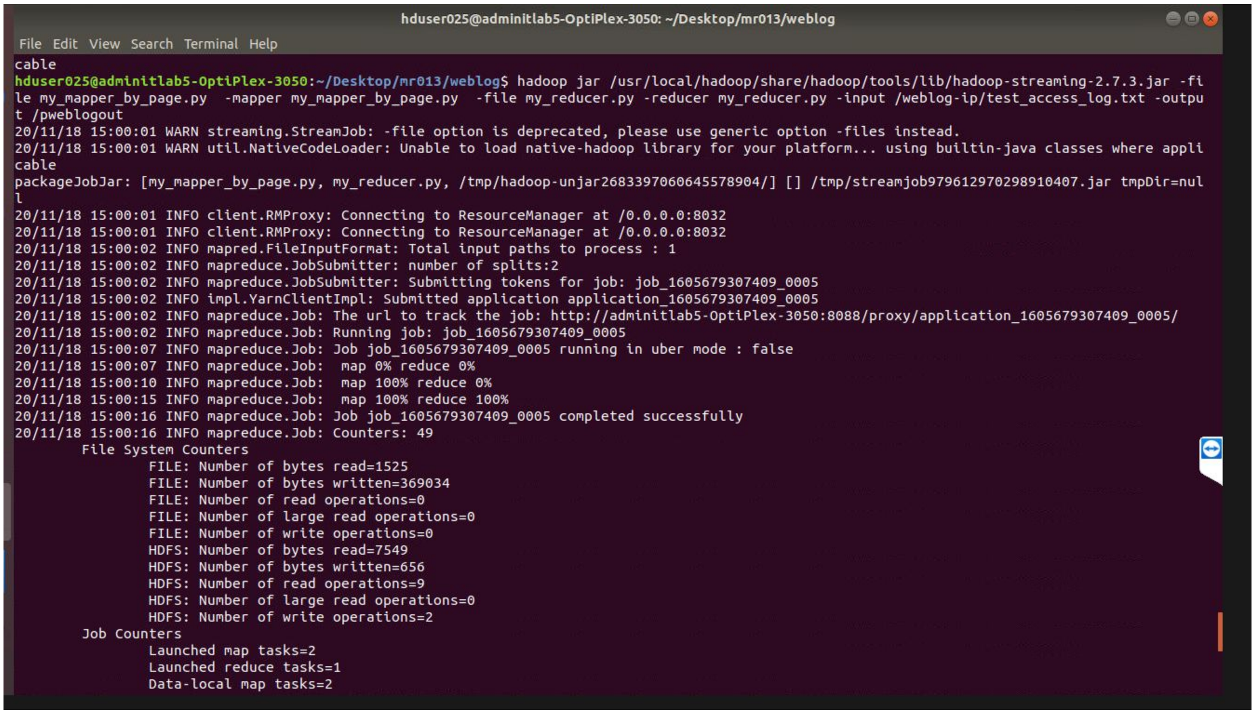
my_mapper_by_page.py

```
import sys

for line in sys.stdin:
        data = line.strip().split(" ")
        if len(data) == 10:
        page = data[6]
        print page
```

my_reducer.py

```
import sys

number = 0
oldKey = None

for line in sys.stdin:
        thisKey = line

   if oldKey and oldKey != thisKey:
        print oldKey, "\t", number
        oldKey = thisKey;
        number = 0

        oldKey = thisKey
        number += 1

if oldKey != None:
        print oldKey, "\t", number
```

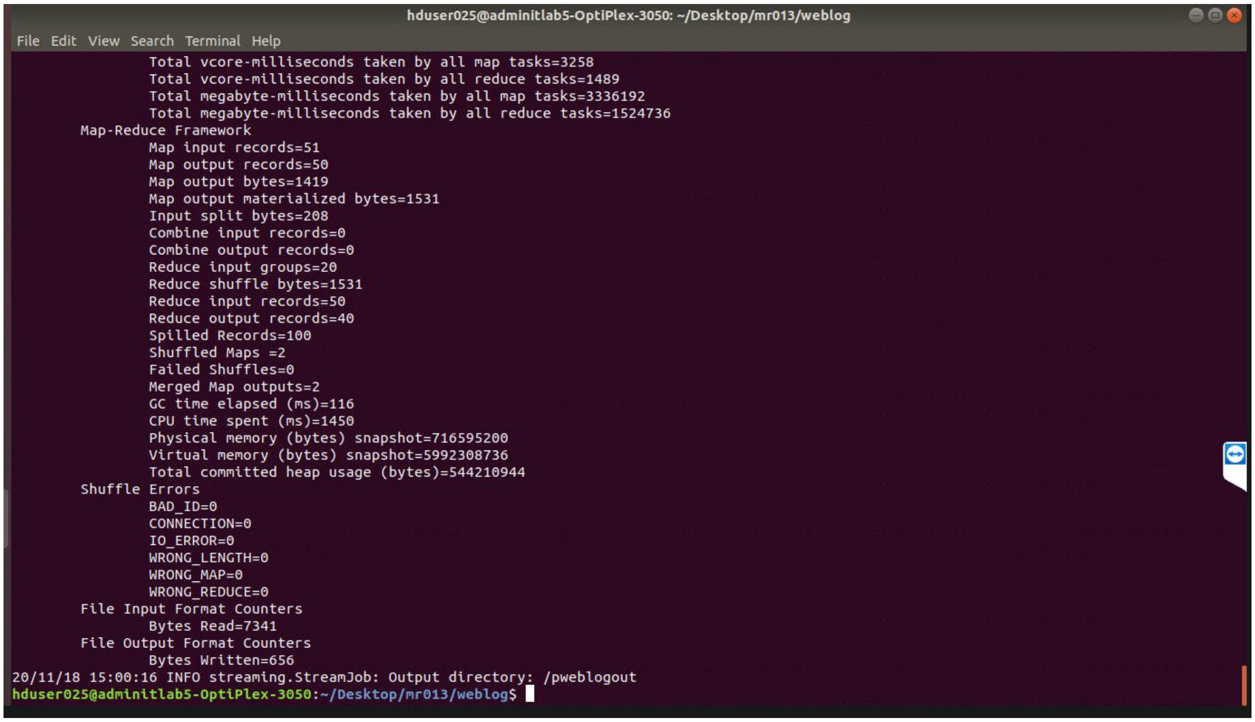
OUTPUT:



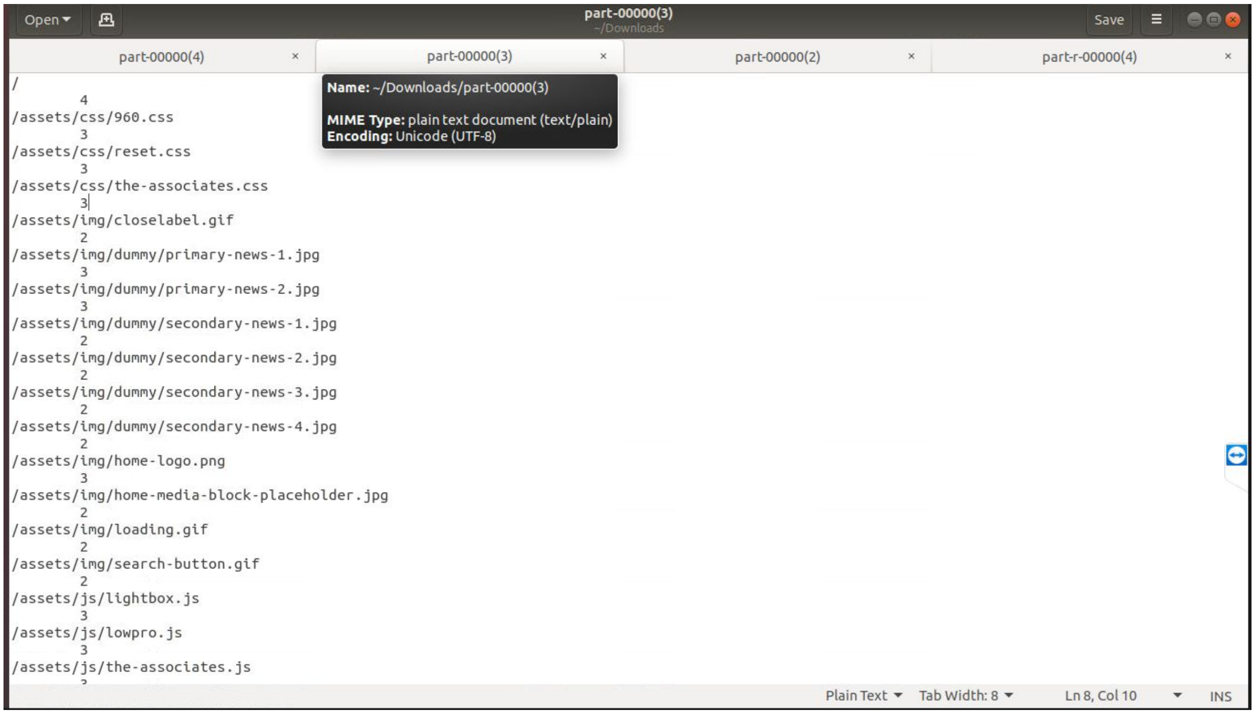
```
cable
hduser025@adminitlab5-OptiPlex-3050:~/Desktop/mr013/weblog$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -fi
le my_mapper_by_page.py  -mapper my_mapper_by_page.py  -file my_reducer.py -reducer my_reducer.py -input /weblog-ip/test_access_log.txt -outpu
t /pweblogout
20/11/18 15:00:01 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
20/11/18 15:00:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appli
cable
packageJobJar: [my_mapper_by_page.py, my_reducer.py, /tmp/hadoop-unjar2683397060645578904/] [] /tmp/streamjob979612970298910407.jar tmpDir=nul
l
20/11/18 15:00:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/11/18 15:00:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/11/18 15:00:02 INFO mapred.FileInputFormat: Total input paths to process : 1
20/11/18 15:00:02 INFO mapreduce.JobSubmitter: number of splits:2
20/11/18 15:00:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1605679307409_0005
20/11/18 15:00:02 INFO impl.YarnClientImpl: Submitted application application_1605679307409_0005
20/11/18 15:00:02 INFO mapreduce.Job: The url to track the job: http://adminitlab5-OptiPlex-3050:8088/proxy/application_1605679307409_0005/
20/11/18 15:00:02 INFO mapreduce.Job: Running job: job_1605679307409_0005
20/11/18 15:00:07 INFO mapreduce.Job: Job job_1605679307409_0005 running in uber mode : false
20/11/18 15:00:07 INFO mapreduce.Job:  map 0% reduce 0%
20/11/18 15:00:10 INFO mapreduce.Job:  map 100% reduce 0%
20/11/18 15:00:15 INFO mapreduce.Job:  map 100% reduce 100%
20/11/18 15:00:16 INFO mapreduce.Job: Job job_1605679307409_0005 completed successfully
20/11/18 15:00:16 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1525
                FILE: Number of bytes written=369034
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7549
                HDFS: Number of bytes written=656
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
```

hduser025@adminitlab5-OptiPlex-3050: ~/Desktop/mr013/weblog
File Edit View Search Terminal Help

160117737006
N.Durga Sai Lakshmi


hduser025@adminitlab5-OptiPlex-3050: ~/Desktop/mr013/weblog
File Edit View Search Terminal Help

```
                Total vcore-milliseconds taken by all map tasks=3258
                Total vcore-milliseconds taken by all reduce tasks=1489
                Total megabyte-milliseconds taken by all map tasks=3336192
                Total megabyte-milliseconds taken by all reduce tasks=1524736
        Map-Reduce Framework
                Map input records=51
                Map output records=50
                Map output bytes=1419
                Map output materialized bytes=1531
                Input split bytes=208
                Combine input records=0
                Combine output records=0
                Reduce input groups=20
                Reduce shuffle bytes=1531
                Reduce input records=50
                Reduce output records=40
                Spilled Records=100
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=116
                CPU time spent (ms)=1450
                Physical memory (bytes) snapshot=716595200
                Virtual memory (bytes) snapshot=5992308736
                Total committed heap usage (bytes)=544210944
        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=7341
        File Output Format Counters
                Bytes Written=656
20/11/18 15:00:16 INFO streaming.StreamJob: Output directory: /pweblogout
hduser025@adminitlab5-OptiPlex-3050:~/Desktop/mr013/weblog$
```



Name: ~/Downloads/part-00000(3)

MIME Type: plain text document (text/plain)
Encoding: Unicode (UTF-8)

```
/
        4
/assets/css/960.css
        3
/assets/css/reset.css
        3
/assets/css/the-associates.css
        3
/assets/img/closelabel.gif
        2
/assets/img/dummy/primary-news-1.jpg
        3
/assets/img/dummy/primary-news-2.jpg
        3
/assets/img/dummy/secondary-news-1.jpg
        2
/assets/img/dummy/secondary-news-2.jpg
        2
/assets/img/dummy/secondary-news-3.jpg
        2
/assets/img/dummy/secondary-news-4.jpg
        2
/assets/img/home-logo.png
        3
/assets/img/home-media-block-placeholder.jpg
        2
/assets/img/loading.gif
        2
/assets/img/search-button.gif
        2
/assets/js/lightbox.js
        3
/assets/js/lowpro.js
        3
/assets/js/the-associates.js
```

Problem Statement-5: Write a pig script to find max temp from the given dataset.

Description:

160117737006
N.Durga Sai Lakshmi

The Apache Pig MAX function is used to find out the maximum of the numeric values or chararrays in a single-column bag. It requires a preceding GROUP ALL statement for global maximums and a GROUP BY statement for group maximums. However, it ignores the NULL values.

Procedure:

$cd hadoop-3.2.1/

$cd sbin/

$cd start-all.sh

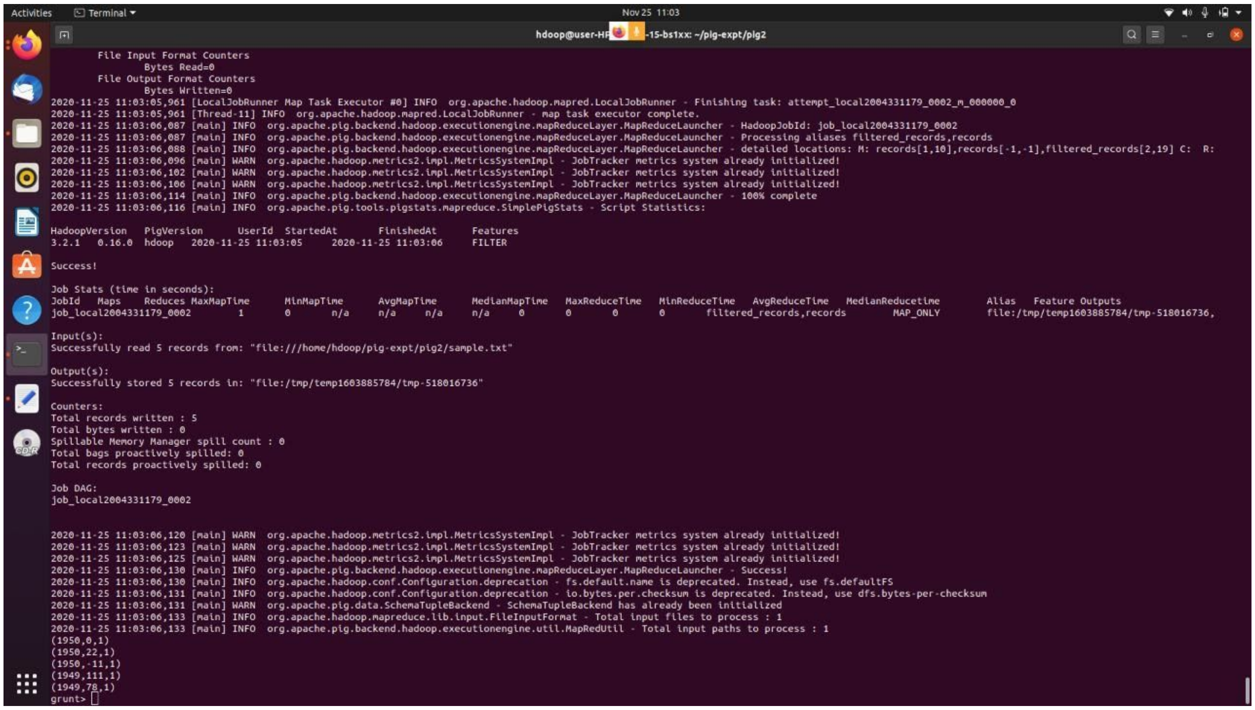
$jps

$cd..

$pig –x local

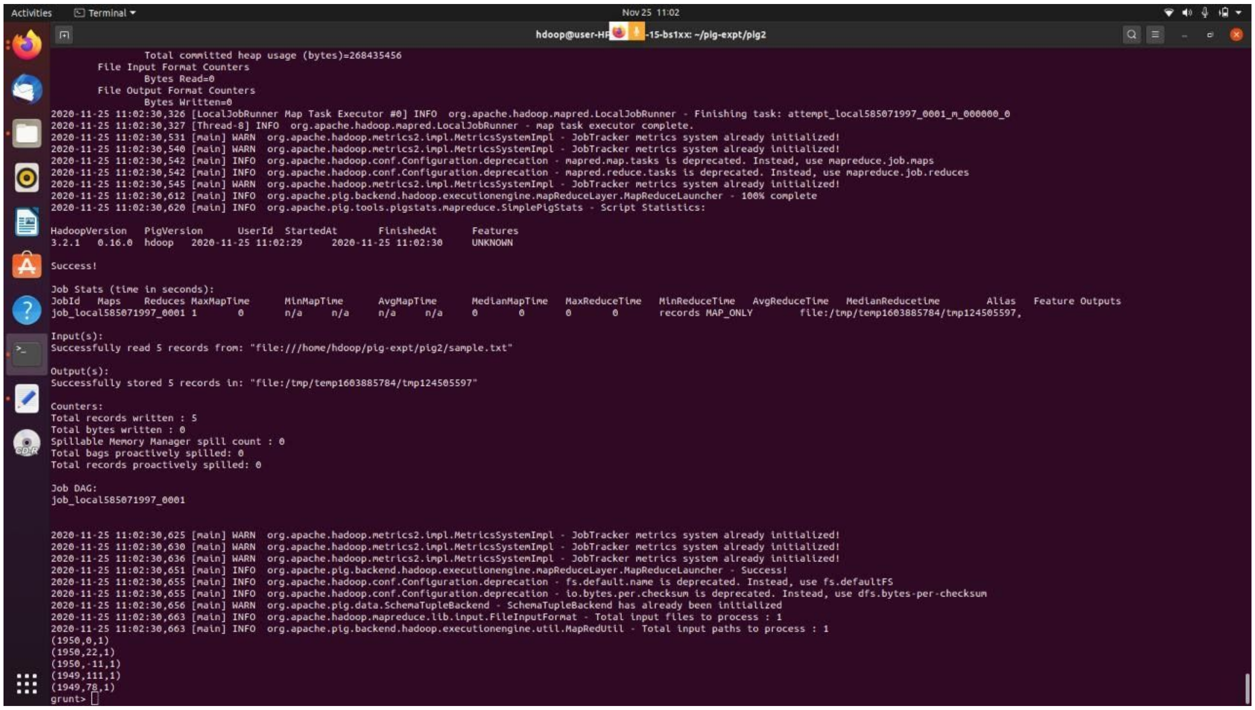
Code:

grunt>records = LOAD 'sample.txt' AS (year:chararray, temperature:int, quality:int);



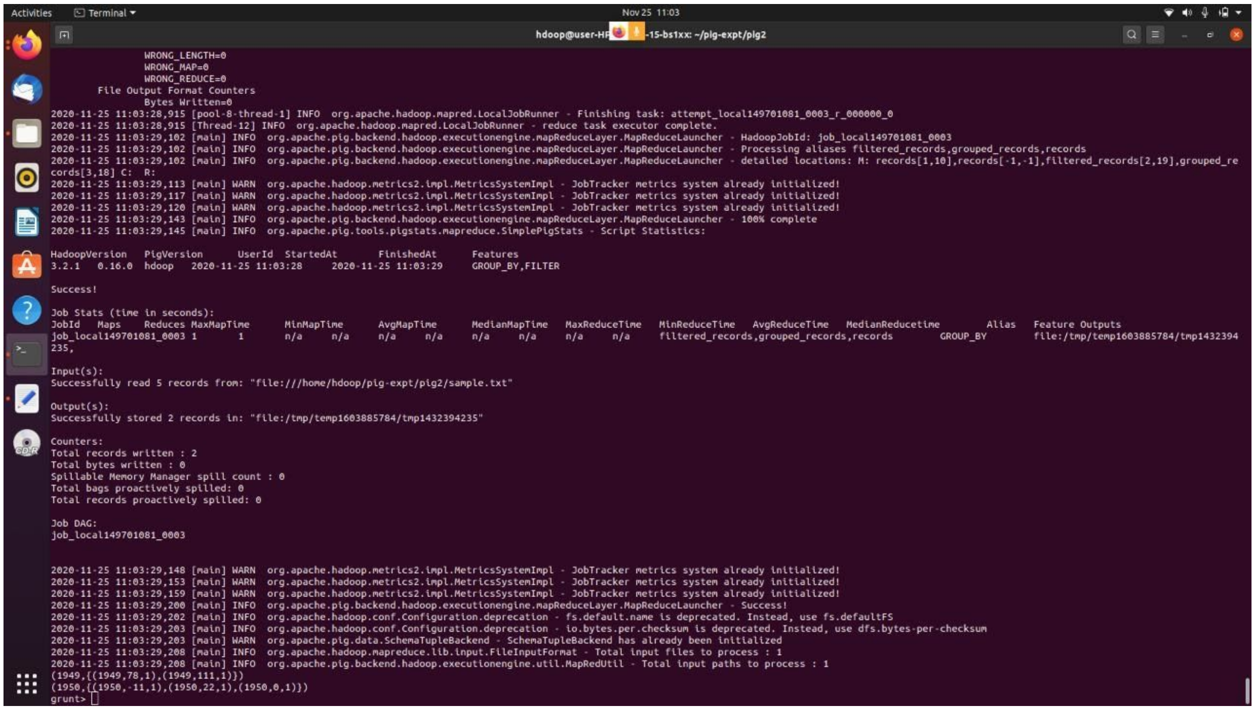
```
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2020-11-25 11:03:05,961 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local2004331179_0002_m_000000_0
2020-11-25 11:03:05,961 [Thread-11] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2020-11-25 11:03:06,087 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local2004331179_0002
2020-11-25 11:03:06,087 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases filtered_records,records
2020-11-25 11:03:06,088 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: records[1,10],records[-1,-1],filtered_records[2,19] C:  R:
2020-11-25 11:03:06,096 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:06,102 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:06,106 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:06,114 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-11-25 11:03:06,116 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 11:03:05     2020-11-25 11:03:06     FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local2004331179_0002        1       0       n/a     n/a     n/a     n/a     0       0       0       0       filtered_records,records        MAP_ONLY        file:/tmp/temp1603885784/tmp-518016736,

Input(s):
Successfully read 5 records from: "file:///home/hdoop/pig-expt/pig2/sample.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp1603885784/tmp-518016736"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2004331179_0002


2020-11-25 11:03:06,120 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:06,123 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:06,125 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:06,130 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 11:03:06,130 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 11:03:06,131 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 11:03:06,131 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 11:03:06,133 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 11:03:06,133 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)
grunt> 
```

grunt>filtered_records = FILTER records BY temperature != 9999 AND (quality == 0 OR quality == 1 OR quality == 4 OR quality == 5 OR quality == 9);



```
                Total committed heap usage (bytes)=268435456
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2020-11-25 11:02:30,326 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local585071997_0001_m_000000_0
2020-11-25 11:02:30,327 [Thread-8] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2020-11-25 11:02:30,531 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:02:30,540 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:02:30,542 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2020-11-25 11:02:30,542 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2020-11-25 11:02:30,545 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:02:30,612 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-11-25 11:02:30,620 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 11:02:29     2020-11-25 11:02:30     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local585071997_0001 1       0       n/a     n/a     n/a     n/a     0       0       0       0       records MAP_ONLY        file:/tmp/temp1603885784/tmp124505597,

Input(s):
Successfully read 5 records from: "file:///home/hdoop/pig-expt/pig2/sample.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp1603885784/tmp124505597"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local585071997_0001


2020-11-25 11:02:30,625 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:02:30,630 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:02:30,636 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:02:30,651 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 11:02:30,655 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 11:02:30,655 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 11:02:30,656 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 11:02:30,663 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 11:02:30,663 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)
grunt>
```

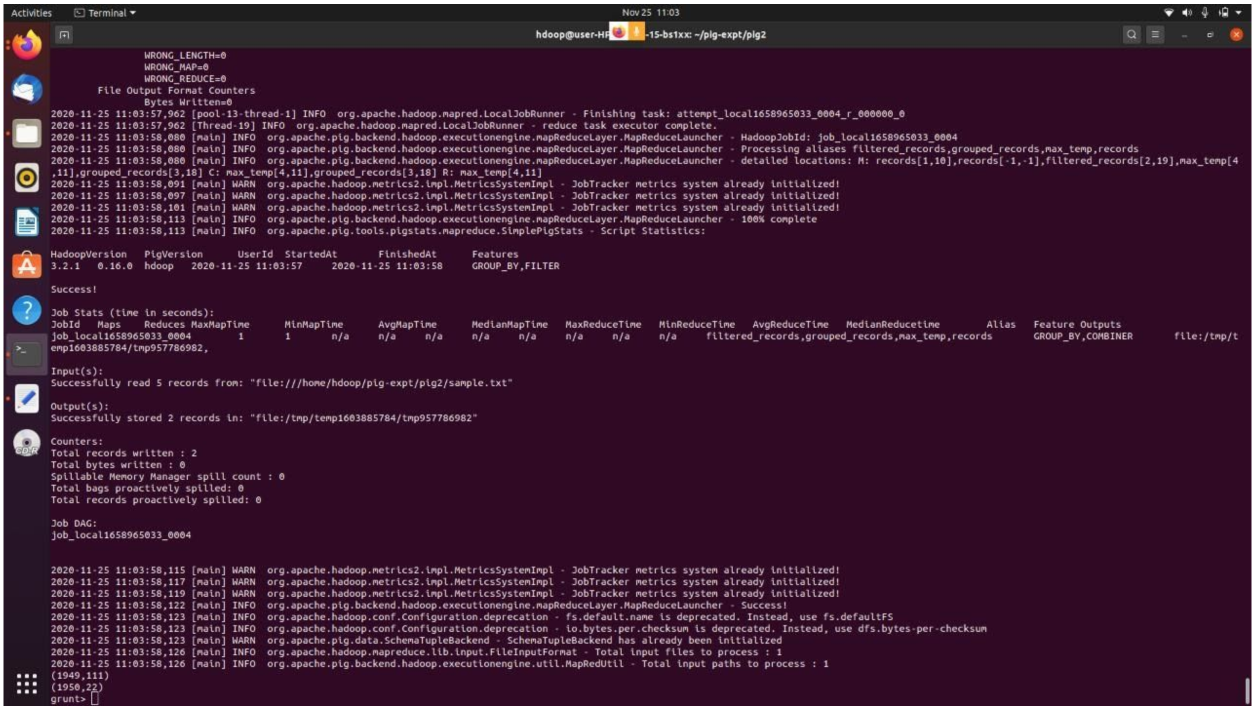
grunt>grouped_records = GROUP filtered_records BY year;



```
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=0
2020-11-25 11:03:28,915 [pool-8-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local149701081_0003_r_000000_0
2020-11-25 11:03:28,915 [Thread-12] INFO  org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2020-11-25 11:03:29,102 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local149701081_0003
2020-11-25 11:03:29,102 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases filtered_records,grouped_records,records
2020-11-25 11:03:29,102 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: records[1,10],records[-1,-1],filtered_records[2,19],grouped_re
cords[3,18] C:  R:
2020-11-25 11:03:29,113 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:29,117 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:29,120 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:29,143 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-11-25 11:03:29,145 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 11:03:28     2020-11-25 11:03:29     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local149701081_0003 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     filtered_records,grouped_records,records        GROUP_BY        file:/tmp/temp1603885784/tmp1432394
235,

Input(s):
Successfully read 5 records from: "file:///home/hdoop/pig-expt/pig2/sample.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp1603885784/tmp1432394235"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local149701081_0003


2020-11-25 11:03:29,148 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:29,153 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:29,159 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:29,200 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 11:03:29,202 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 11:03:29,203 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 11:03:29,203 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 11:03:29,208 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 11:03:29,208 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1949,{(1949,78,1),(1949,111,1)})
(1950,{(1950,-11,1),(1950,22,1),(1950,0,1)})
grunt>
```

grunt>max_temp = FOREACH grouped_records GENERATE group, MAX(filtered_records.temperature);

grunt>DUMP max_temp;



```
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=0
2020-11-25 11:03:57,962 [pool-13-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1658965033_0004_r_000000_0
2020-11-25 11:03:57,962 [Thread-19] INFO  org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2020-11-25 11:03:58,080 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local1658965033_0004
2020-11-25 11:03:58,080 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases filtered_records,grouped_records,max_temp,records
2020-11-25 11:03:58,080 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: records[1,10],records[-1,-1],filtered_records[2,19],max_temp[4
,11],grouped_records[3,18] C: max_temp[4,11],grouped_records[3,18] R: max_temp[4,11]
2020-11-25 11:03:58,091 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:58,097 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:58,101 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:58,113 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-11-25 11:03:58,113 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 11:03:57     2020-11-25 11:03:58     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1658965033_0004        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     filtered_records,grouped_records,max_temp,records       GROUP_BY,COMBINER       file:/tmp/t
emp1603885784/tmp957786982,

Input(s):
Successfully read 5 records from: "file:///home/hdoop/pig-expt/pig2/sample.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp1603885784/tmp957786982"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1658965033_0004


2020-11-25 11:03:58,115 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:58,117 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:58,119 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:03:58,122 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 11:03:58,123 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 11:03:58,123 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 11:03:58,123 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 11:03:58,126 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 11:03:58,126 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1949,111)
(1950,22)
grunt> 
```

Sample.txt

```
1950    0       1
1950    22      1
1950    -11     1
1949    111     1
1949    78      1
```

Problem Statement-6: Write Pig script to implement Word Count Job.

Description: Pig is a high-level programming language useful for analyzing large data sets. Pig was a result of development effort at Yahoo!

In a MapReduce framework, programs need to be translated into a series of Map and Reduce stages. However, this is not a programming model which data analysts are familiar with. So, in order to bridge this gap, an abstraction called Pig was built on top of Hadoop.

Procedure:

$cd hadoop-3.2.1/

$cd sbin/

$cd start-all.sh

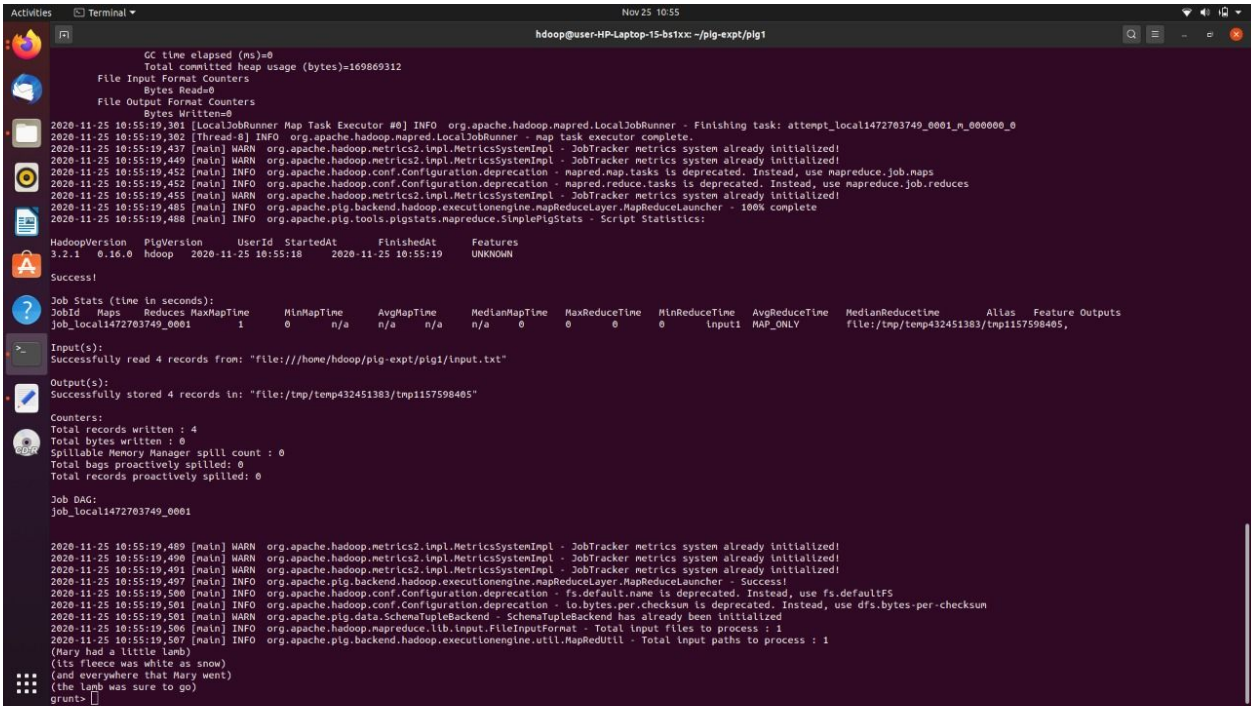
$jps

$cd..

$pig –x local

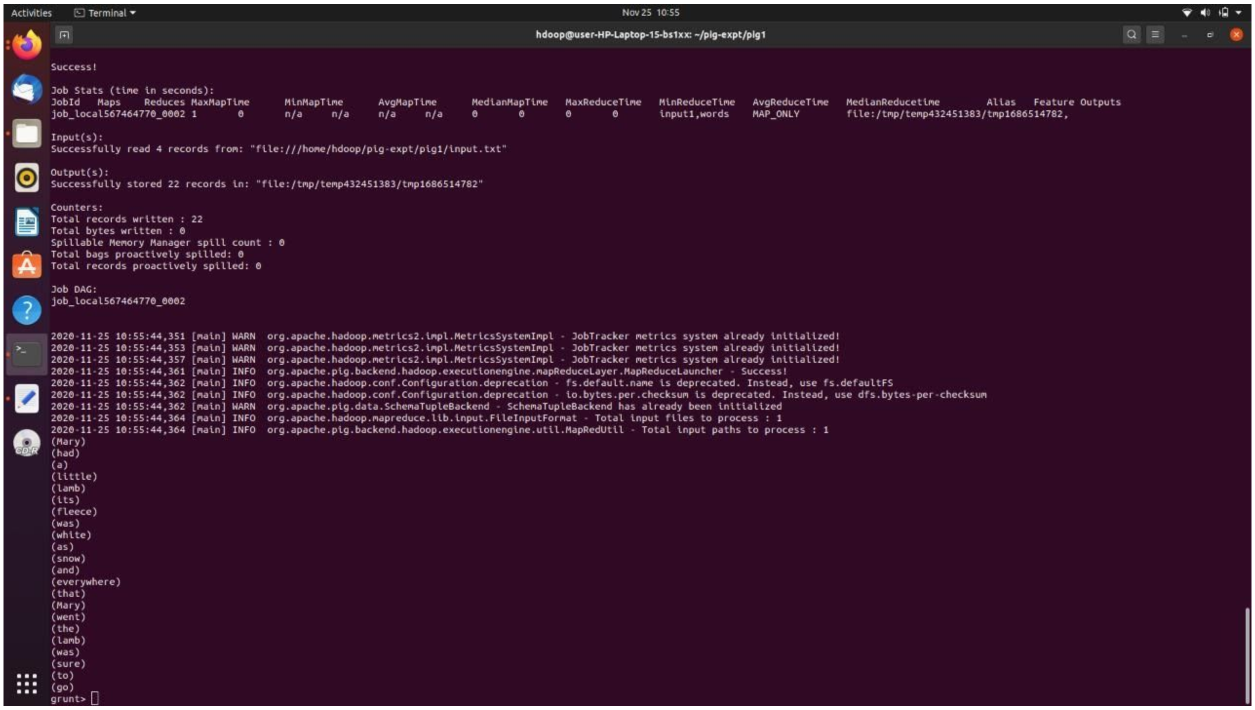
Code:

grunt>input1 = load 'sample.txt' as (line);
Output



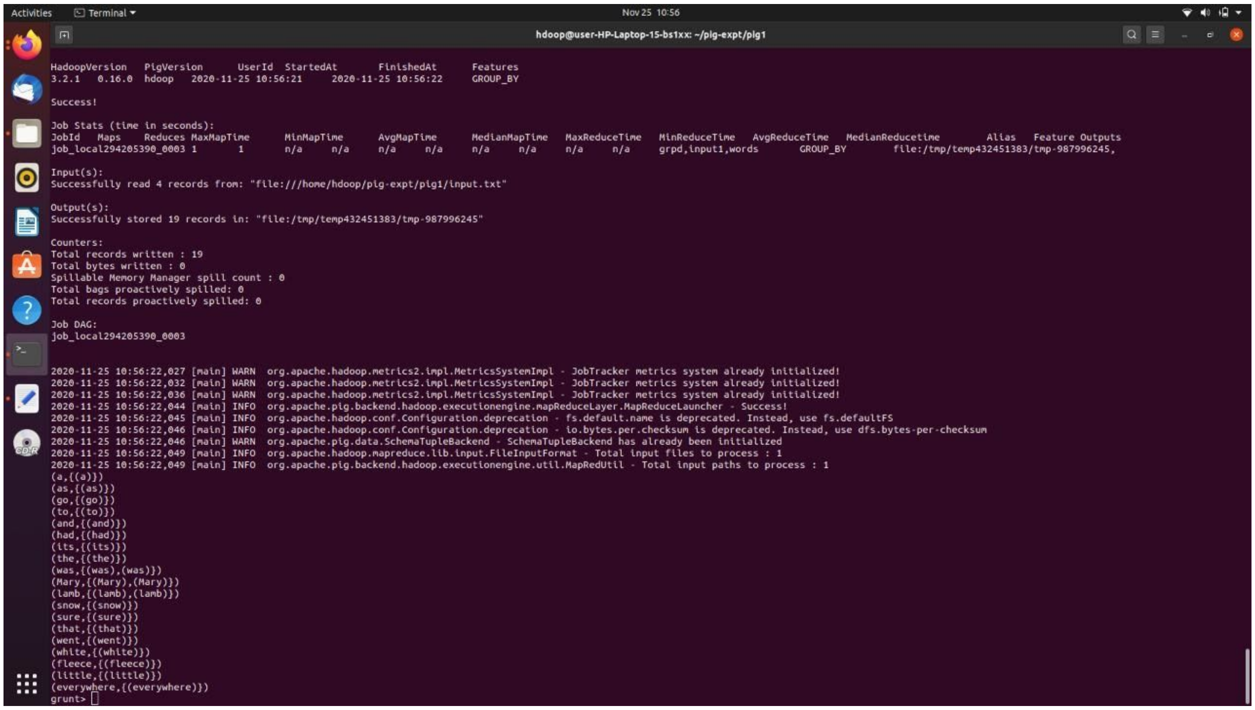
```
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=169869312
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2020-11-25 10:55:19,301 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1472703749_0001_m_000000_0
2020-11-25 10:55:19,302 [Thread-8] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2020-11-25 10:55:19,437 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:19,449 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:19,452 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2020-11-25 10:55:19,452 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2020-11-25 10:55:19,455 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:19,485 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-11-25 10:55:19,488 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 10:55:18     2020-11-25 10:55:19     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1472703749_0001        1       0       n/a     n/a     n/a     n/a     0       0       0       0       input1  MAP_ONLY        file:/tmp/temp432451383/tmp1157598405,

Input(s):
Successfully read 4 records from: "file:///home/hdoop/pig-expt/pig1/input.txt"

Output(s):
Successfully stored 4 records in: "file:/tmp/temp432451383/tmp1157598405"

Counters:
Total records written : 4
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1472703749_0001


2020-11-25 10:55:19,489 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:19,490 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:19,491 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:19,497 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 10:55:19,500 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 10:55:19,501 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 10:55:19,501 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 10:55:19,506 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 10:55:19,507 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Mary had a little lamb)
(its fleece was white as snow)
(and everywhere that Mary went)
(the lamb was sure to go)
grunt> 
```

grunt>words = foreach input generate flatten(TOKENIZE(line)) as word;



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local567464770_0002 1       0       n/a     n/a     n/a     n/a     0       0       0       0       input1,words    MAP_ONLY        file:/tmp/temp432451383/tmp1686514782,

Input(s):
Successfully read 4 records from: "file:///home/hdoop/pig-expt/pig1/input.txt"

Output(s):
Successfully stored 22 records in: "file:/tmp/temp432451383/tmp1686514782"

Counters:
Total records written : 22
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local567464770_0002


2020-11-25 10:55:44,351 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:44,353 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:44,357 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:55:44,361 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 10:55:44,362 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 10:55:44,362 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 10:55:44,362 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 10:55:44,364 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 10:55:44,364 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Mary)
(had)
(a)
(little)
(lamb)
(its)
(fleece)
(was)
(white)
(as)
(snow)
(and)
(everywhere)
(that)
(Mary)
(went)
(the)
(lamb)
(was)
(sure)
(to)
(go)
grunt> 
```

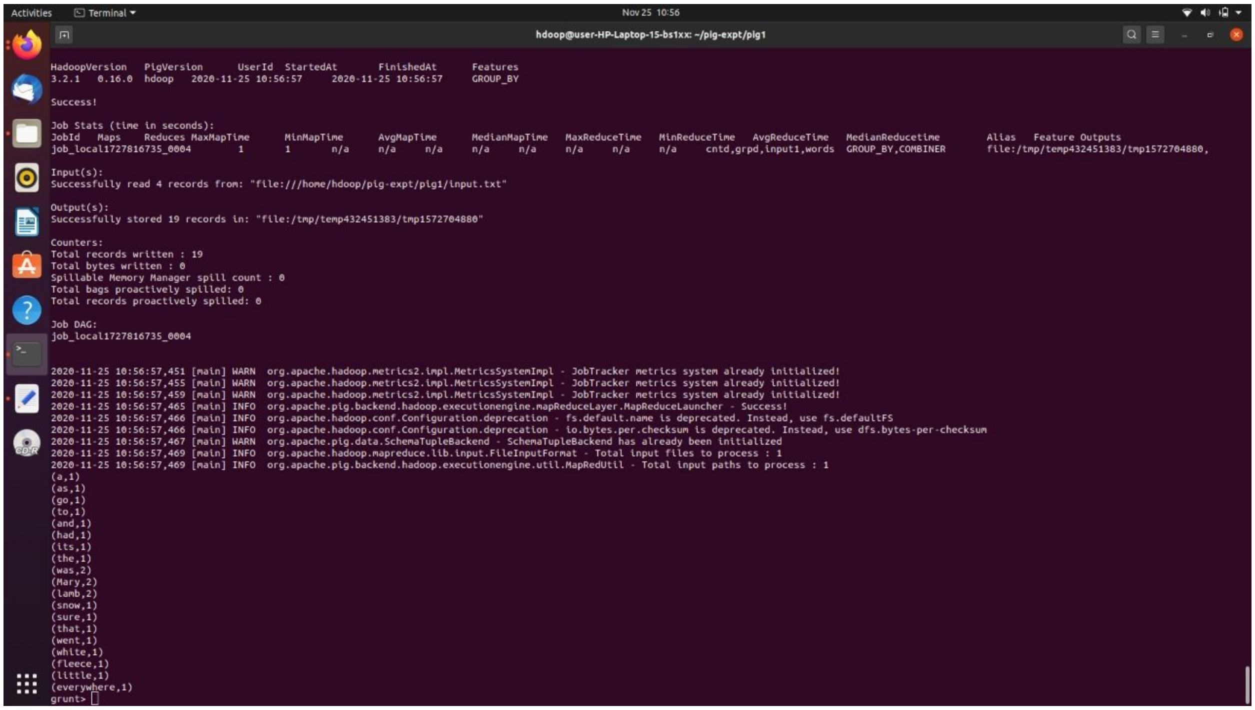
grunt>grpd = group words by word;



```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 10:56:21     2020-11-25 10:56:22     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local294205390_0003 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     grpd,input1,words       GROUP_BY        file:/tmp/temp432451383/tmp-987996245,

Input(s):
Successfully read 4 records from: "file:///home/hdoop/pig-expt/pig1/input.txt"

Output(s):
Successfully stored 19 records in: "file:/tmp/temp432451383/tmp-987996245"

Counters:
Total records written : 19
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local294205390_0003


2020-11-25 10:56:22,027 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:56:22,032 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:56:22,036 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:56:22,044 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 10:56:22,045 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 10:56:22,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 10:56:22,046 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 10:56:22,049 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 10:56:22,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(a,{(a)})
(as,{(as)})
(go,{(go)})
(to,{(to)})
(and,{(and)})
(had,{(had)})
(its,{(its)})
(the,{(the)})
(was,{(was),(was)})
(Mary,{(Mary),(Mary)})
(lamb,{(lamb),(lamb)})
(snow,{(snow)})
(sure,{(sure)})
(that,{(that)})
(went,{(went)})
(white,{(white)})
(fleece,{(fleece)})
(little,{(little)})
(everywhere,{(everywhere)})
grunt> 
```

grunt>cntd = foreach grpd generate group, COUNT(words);

grunt>dump cntd;



```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.1   0.16.0  hdoop   2020-11-25 10:56:57     2020-11-25 10:56:57     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1727816735_0004        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     cntd,grpd,input1,words  GROUP_BY,COMBINER       file:/tmp/temp432451383/tmp1572704880,

Input(s):
Successfully read 4 records from: "file:///home/hdoop/pig-expt/pig1/input.txt"

Output(s):
Successfully stored 19 records in: "file:/tmp/temp432451383/tmp1572704880"

Counters:
Total records written : 19
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1727816735_0004


2020-11-25 10:56:57,451 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:56:57,455 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:56:57,459 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 10:56:57,465 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-25 10:56:57,466 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-25 10:56:57,466 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2020-11-25 10:56:57,467 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 10:56:57,469 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 10:56:57,469 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(a,1)
(as,1)
(go,1)
(to,1)
(and,1)
(had,1)
(its,1)
(the,1)
(was,2)
(Mary,2)
(lamb,2)
(snow,1)
(sure,1)
(that,1)
(went,1)
(white,1)
(fleece,1)
(little,1)
(everywhere,1)
grunt> 
```

Sample.txt

Mary had a little lamb
its fleece was white as snow
and everywhere that Mary went
the lamb was sure to go

Problem Statement-7: Find the Number of Products Sold in Each Country for the given dataset using pig framework.

Description:
Apache Pig enables people to focus more on analyzing bulk data sets and to spend less time writing Map-Reduce programs. Similar to Pigs, who eat anything, the Apache Pig programming language is designed to work upon any kind of data.

Procedure:

```
$cd hadoop-3.2.1/

$cd sbin/

$cd start-all.sh
```

160117737006

N.Durga Sai Lakshmi

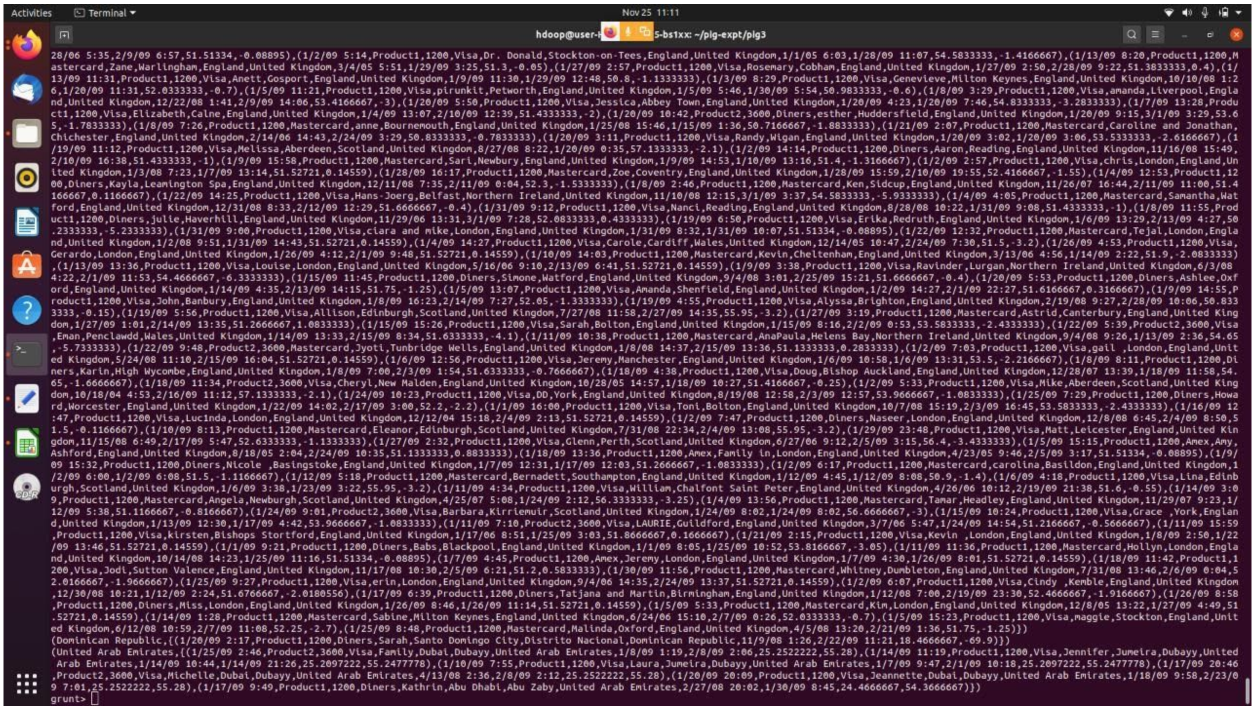
$jps

$cd..

$pig –x local

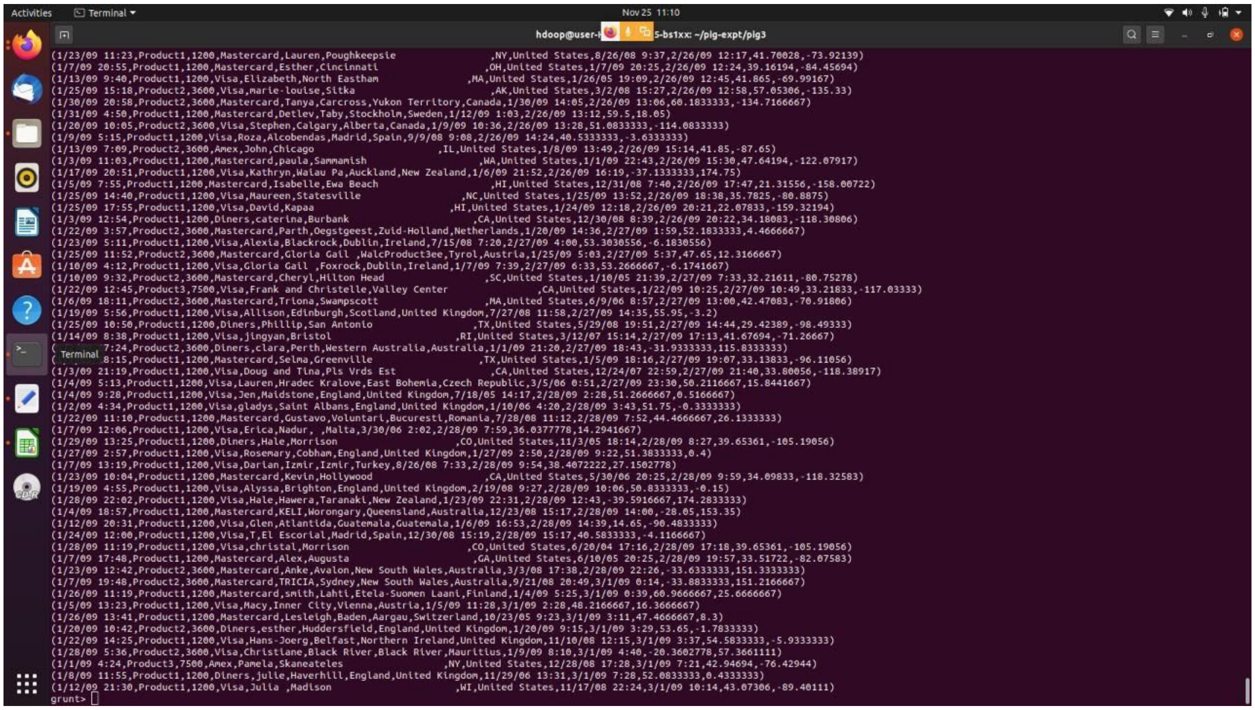
Code:

grunt>salesTable = LOAD '/SalesJan2009.csv' USING PigStorage(',') AS (Transaction_date:chararray,Product:chararray,Price:chararray,Payment_Type:chararray,Name:chararray,City:chararray,State:chararray,Country:chararray,Account_Created:chararray,Last_Login:chararray,Latitude:chararray,Longitude:chararray);



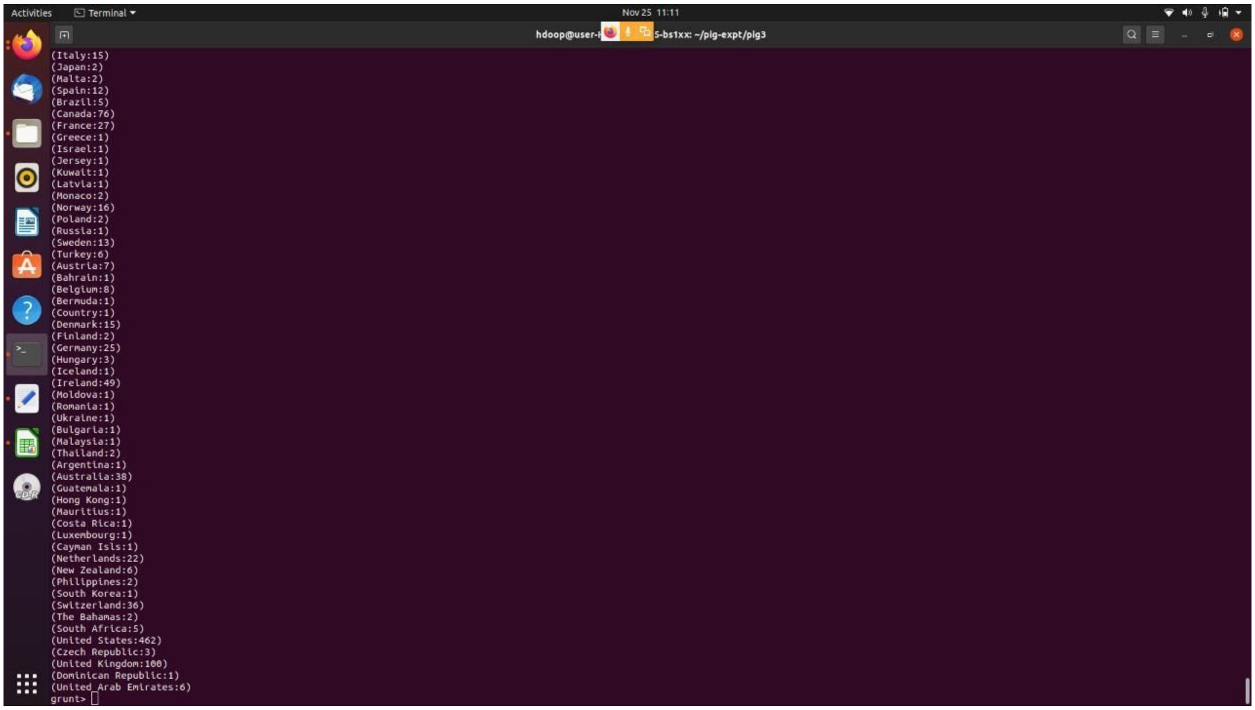
```
28/06 5:35,2/9/09 6:57,51.51334,-0.08895),(1/2/09 5:14,Product1,1200,Visa,Dr. Donald,Stockton-on-Tees,England,United Kingdom,1/1/05 6:03,1/28/09 11:07,54.5833333,-1.4166667),(1/13/09 8:20,Product1,1200,M
astercard,Zane,Warlingham,England,United Kingdom,3/4/05 5:51,1/29/09 3:25,51.3,-0.05),(1/27/09 2:57,Product1,1200,Visa,Rosemary,Cobham,England,United Kingdom,1/27/09 2:50,2/28/09 9:22,51.3833333,0.4),(1/
13/09 11:31,Product1,1200,Visa,Anett,Gosport,England,United Kingdom,1/9/09 11:30,1/29/09 12:48,50.8,-1.1333333),(1/3/09 8:29,Product1,1200,Visa,Genevieve,Milton Keynes,England,United Kingdom,10/10/08 1:2
6,1/20/09 11:31,52.0333333,-0.7),(1/5/09 11:21,Product1,1200,Visa,pirunkit,Petworth,England,United Kingdom,1/5/09 5:46,1/30/09 5:54,50.9833333,-0.6),(1/8/09 3:29,Product1,1200,Visa,amanda,Liverpool,Engla
nd,United Kingdom,12/22/08 1:41,2/9/09 14:06,53.4166667,-3),(1/20/09 5:50,Product1,1200,Visa,Jessica,Abbey Town,England,United Kingdom,1/20/09 4:23,1/20/09 7:46,54.8333333,-3.2833333),(1/7/09 13:28,Produ
ct1,1200,Visa,Elizabeth,Calne,England,United Kingdom,1/4/09 13:07,2/10/09 12:39,51.4333333,-2),(1/20/09 10:42,Product2,3600,Diners,esther,Huddersfield,England,United Kingdom,1/20/09 9:15,3/1/09 3:29,53.6
5,-1.7833333),(1/8/09 7:26,Product1,1200,Mastercard,anne,Bournemouth,England,United Kingdom,1/25/08 15:46,1/15/09 1:36,50.7166667,-1.8833333),(1/21/09 2:07,Product1,1200,Mastercard,Caroline and Jonathan,
Chichester,England,United Kingdom,2/14/06 14:43,2/24/09 3:29,50.8333333,-0.7833333),(1/20/09 3:11,Product1,1200,Visa,Randy,Wigan,England,United Kingdom,1/20/09 3:02,1/20/09 3:06,53.5333333,-2.6166667),(1
/19/09 11:12,Product1,1200,Visa,Melissa,Aberdeen,Scotland,United Kingdom,8/27/08 8:22,1/20/09 0:35,57.1333333,-2.1),(1/2/09 14:14,Product1,1200,Diners,Aaron,Reading,England,United Kingdom,11/16/08 15:49,
2/10/09 16:38,51.4333333,-1),(1/9/09 15:58,Product1,1200,Mastercard,Sari,Newbury,England,United Kingdom,1/9/09 14:53,1/10/09 13:16,51.4,-1.3166667),(1/2/09 2:57,Product1,1200,Visa,chris,London,England,Un
ited Kingdom,1/3/08 7:23,1/7/09 13:14,51.52721,0.14559),(1/28/09 16:17,Product1,1200,Mastercard,Zoe,Coventry,England,United Kingdom,1/28/09 15:59,2/10/09 19:55,52.4166667,-1.55),(1/4/09 12:53,Product1,12
00,Diners,Kayla,Leamington Spa,England,United Kingdom,12/11/08 7:35,2/11/09 0:04,52.3,-1.5333333),(1/8/09 2:46,Product1,1200,Mastercard,Ken,Sidcup,England,United Kingdom,11/26/07 16:44,2/11/09 11:00,51.4
166667,0.1166667),(1/22/09 14:25,Product1,1200,Visa,Hans-Joerg,Belfast,Northern Ireland,United Kingdom,11/10/08 12:15,3/1/09 3:37,54.5833333,-5.9333333),(1/4/09 4:05,Product1,1200,Mastercard,Samantha,Wat
ford,England,United Kingdom,12/31/08 8:33,2/12/09 12:29,51.6666667,-0.4),(1/31/09 9:12,Product1,1200,Visa,Nanci,Reading,England,United Kingdom,8/28/08 10:22,1/31/09 9:08,51.4333333,-1),(1/8/09 11:55,Prod
uct1,1200,Diners,julie,Haverhill,England,United Kingdom,11/29/06 13:31,3/1/09 7:28,52.0833333,0.4333333),(1/19/09 6:50,Product1,1200,Visa,Erika,Redruth,England,United Kingdom,1/6/09 13:29,2/13/09 4:27,50
.2333333,-5.2333333),(1/31/09 9:00,Product1,1200,Visa,ciara and mike,London,England,United Kingdom,1/31/09 8:32,1/31/09 10:07,51.51334,-0.08895),(1/22/09 12:32,Product1,1200,Mastercard,Tejal,London,Engla
nd,United Kingdom,1/2/08 9:51,1/31/09 14:43,51.52721,0.14559),(1/4/09 14:27,Product1,1200,Visa,Carole,Cardiff,Wales,United Kingdom,12/14/05 10:47,2/24/09 7:30,51.5,-3.2),(1/26/09 4:53,Product1,1200,Visa,
Gerardo,London,England,United Kingdom,1/26/09 4:12,2/1/09 9:48,51.52721,0.14559),(1/10/09 14:03,Product1,1200,Mastercard,Kevin,Cheltenham,England,United Kingdom,3/13/06 4:56,1/14/09 2:22,51.9,-2.0833333)
,(1/13/09 13:36,Product1,1200,Visa,Louise,London,England,United Kingdom,5/16/06 9:10,2/13/09 6:41,51.52721,0.14559),(1/9/09 3:38,Product1,1200,Visa,Ravinder,Lurgan,Northern Ireland,United Kingdom,6/3/08
4:22,2/1/09 11:53,54.4666667,-6.3333333),(1/15/09 11:45,Product1,1200,Diners,Simone,Watford,England,United Kingdom,9/4/08 3:01,2/25/09 15:21,51.6666667,-0.4),(1/20/09 5:53,Product1,1200,Diners,Ashlee,Oxf
ord,England,United Kingdom,1/14/09 4:35,2/13/09 14:15,51.75,-1.25),(1/5/09 13:07,Product1,1200,Visa,Amanda,Shenfield,England,United Kingdom,1/2/09 14:27,2/1/09 22:27,51.6166667,0.3166667),(1/9/09 14:55,P
roduct1,1200,Visa,John,Banbury,England,United Kingdom,1/8/09 16:23,2/14/09 7:27,52.05,-1.3333333),(1/19/09 4:55,Product1,1200,Visa,Alyssa,Brighton,England,United Kingdom,2/19/08 9:27,2/28/09 10:06,50.833
3333,-0.15),(1/19/09 5:56,Product1,1200,Visa,Allison,Edinburgh,Scotland,United Kingdom,7/27/08 11:58,2/27/09 14:35,55.95,-3.2),(1/27/09 3:19,Product1,1200,Mastercard,Astrid,Canterbury,England,United King
dom,1/27/09 1:01,2/14/09 13:35,51.2666667,1.0833333),(1/15/09 15:26,Product1,1200,Visa,Sarah,Bolton,England,United Kingdom,1/15/09 8:16,2/2/09 0:53,53.5833333,-2.4333333),(1/22/09 5:39,Product2,3600,Visa
,Eman,Penclawdd,Wales,United Kingdom,1/14/09 13:33,2/15/09 8:34,51.6333333,-4.1),(1/11/09 10:38,Product1,1200,Mastercard,AnaPaula,Helens Bay,Northern Ireland,United Kingdom,9/4/08 9:26,1/13/09 2:36,54.65
,-5.7333333),(1/22/09 9:48,Product2,3600,Mastercard,Jyoti,Tunbridge Wells,England,United Kingdom,1/8/08 14:37,2/15/09 13:36,51.1333333,0.2833333),(1/2/09 7:03,Product1,1200,Visa,gail ,London,England,Unit
ed Kingdom,5/24/08 11:10,2/15/09 16:04,51.52721,0.14559),(1/6/09 12:56,Product1,1200,Visa,Jeremy,Manchester,England,United Kingdom,1/6/09 10:58,1/6/09 13:31,53.5,-2.2166667),(1/8/09 8:11,Product1,1200,Di
ners,Karin,High Wycombe,England,United Kingdom,1/8/09 7:00,2/3/09 1:54,51.6333333,-0.7666667),(1/18/09 4:38,Product1,1200,Visa,Doug,Bishop Auckland,England,United Kingdom,12/28/07 13:39,1/18/09 11:58,54.
65,-1.6666667),(1/18/09 11:34,Product2,3600,Visa,Cheryl,New Malden,England,United Kingdom,10/28/05 14:57,1/18/09 10:27,51.4166667,-0.25),(1/2/09 5:33,Product1,1200,Visa,Mike,Aberdeen,Scotland,United King
dom,10/18/04 4:53,2/16/09 11:12,57.1333333,-2.1),(1/24/09 10:23,Product1,1200,Visa,DD,York,England,United Kingdom,8/19/08 12:58,2/3/09 12:57,53.9666667,-1.0833333),(1/25/09 7:29,Product1,1200,Diners,Howa
rd,Worcester,England,United Kingdom,1/22/09 14:02,2/17/09 3:00,52.2,-2.2),(1/1/09 16:00,Product1,1200,Visa,Toni,Bolton,England,United Kingdom,10/7/08 15:19,2/3/09 16:45,53.5833333,-2.4333333),(1/16/09 12
:47,Product1,1200,Visa,Lucinda,London,England,United Kingdom,12/12/04 15:18,2/4/09 2:13,51.52721,0.14559),(1/2/09 7:47,Product1,1200,Diners,Naseer,London,England,United Kingdom,12/8/08 6:45,2/4/09 8:50,5
1.5,-0.1166667),(1/10/09 8:13,Product1,1200,Mastercard,Eleanor,Edinburgh,Scotland,United Kingdom,7/31/08 22:34,2/4/09 13:08,55.95,-3.2),(1/29/09 23:48,Product1,1200,Visa,Matt,Leicester,England,United Kin
gdom,11/15/08 6:49,2/17/09 5:47,52.6333333,-1.1333333),(1/27/09 2:32,Product1,1200,Visa,Glenn,Perth,Scotland,United Kingdom,6/27/06 9:12,2/5/09 3:15,56.4,-3.4333333),(1/5/09 15:15,Product1,1200,Amex,Amy,
Ashford,England,United Kingdom,8/18/05 2:04,2/24/09 10:35,51.1333333,0.8833333),(1/18/09 13:36,Product1,1200,Amex,Family in,London,England,United Kingdom,4/23/05 9:46,2/5/09 3:17,51.51334,-0.08895),(1/9/
09 15:32,Product1,1200,Diners,Nicole ,Basingstoke,England,United Kingdom,1/7/09 12:31,1/17/09 12:03,51.2666667,-1.0833333),(1/2/09 6:17,Product1,1200,Mastercard,carolina,Basildon,England,United Kingdom,1
/2/09 6:00,1/2/09 6:08,51.5,-1.1166667),(1/12/09 5:18,Product1,1200,Mastercard,Bernadett,Southampton,England,United Kingdom,1/12/09 4:45,1/12/09 8:08,50.9,-1.4),(1/6/09 4:18,Product1,1200,Visa,Lina,Edinb
urgh,Scotland,United Kingdom,1/6/09 3:38,1/23/09 3:22,55.95,-3.2),(1/11/09 4:34,Product1,1200,Visa,William,Chalfont Saint Peter,England,United Kingdom,4/26/06 10:12,2/19/09 21:38,51.6,-0.55),(1/14/09 3:0
9,Product1,1200,Mastercard,Angela,Newburgh,Scotland,United Kingdom,4/25/07 5:08,1/24/09 2:12,56.3333333,-3.25),(1/4/09 13:56,Product1,1200,Mastercard,Tamar,Headley,England,United Kingdom,11/29/07 9:23,1/
12/09 5:38,51.1166667,-0.8166667),(1/24/09 9:01,Product2,3600,Visa,Barbara,Kirriemuir,Scotland,United Kingdom,1/24/09 8:02,1/24/09 8:02,56.6666667,-3),(1/15/09 10:24,Product1,1200,Visa,Grace ,York,Englan
d,United Kingdom,1/13/09 12:30,1/17/09 4:42,53.9666667,-1.0833333),(1/11/09 7:10,Product2,3600,Visa,LAURIE,Guildford,England,United Kingdom,3/7/06 5:47,1/24/09 14:54,51.2166667,-0.5666667),(1/11/09 15:59
,Product1,1200,Visa,kirsten,Bishops Stortford,England,United Kingdom,1/17/06 8:51,1/25/09 3:03,51.8666667,0.1666667),(1/21/09 2:15,Product1,1200,Visa,Kevin ,London,England,United Kingdom,1/8/09 2:50,1/22
/09 13:46,51.52721,0.14559),(1/1/09 9:21,Product1,1200,Diners,Babs,Blackpool,England,United Kingdom,1/1/09 8:05,1/25/09 10:52,53.8166667,-3.05),(1/11/09 11:36,Product1,1200,Mastercard,Hollyn,London,Engla
nd,United Kingdom,10/14/08 14:23,1/25/09 11:16,51.51334,-0.08895),(1/7/09 4:45,Product1,1200,Amex,Jeremy,London,England,United Kingdom,1/7/09 4:30,1/26/09 8:01,51.52721,0.14559),(1/18/09 11:42,Product1,1
200,Visa,Jodi,Sutton Valence,England,United Kingdom,11/17/08 10:30,2/5/09 6:21,51.2,0.5833333),(1/30/09 11:56,Product1,1200,Mastercard,Whitney,Dumbleton,England,United Kingdom,7/31/08 13:46,2/6/09 0:04,5
2.0166667,-1.9666667),(1/25/09 9:27,Product1,1200,Visa,erin,London,England,United Kingdom,9/4/06 14:35,2/24/09 13:37,51.52721,0.14559),(1/2/09 6:07,Product1,1200,Visa,Cindy ,Kemble,England,United Kingdom
,12/30/08 10:21,1/12/09 2:24,51.6766667,-2.0180556),(1/17/09 6:39,Product1,1200,Diners,Tatjana and Martin,Birmingham,England,United Kingdom,1/12/08 7:00,2/19/09 23:30,52.4666667,-1.9166667),(1/26/09 8:58
,Product1,1200,Diners,Miss,London,England,United Kingdom,1/26/09 8:46,1/26/09 11:14,51.52721,0.14559),(1/5/09 5:33,Product1,1200,Mastercard,Kim,London,England,United Kingdom,12/8/05 13:22,1/27/09 4:49,51
.52721,0.14559),(1/14/09 1:28,Product1,1200,Mastercard,Sabine,Milton Keynes,England,United Kingdom,6/24/06 15:10,2/7/09 0:26,52.0333333,-0.7),(1/5/09 15:23,Product1,1200,Visa,maggie,Stockton,England,Unit
ed Kingdom,6/12/08 10:59,2/7/09 11:08,52.25,-2.7),(1/25/09 8:48,Product1,1200,Mastercard,Malinda,Oxford,England,United Kingdom,4/5/08 13:20,2/21/09 1:36,51.75,-1.25)})
(Dominican Republic,{(1/20/09 2:17,Product1,1200,Diners,Sarah,Santo Domingo City,Distrito Nacional,Dominican Republic,11/9/08 1:26,2/22/09 11:21,18.4666667,-69.9)})
(United Arab Emirates,{(1/25/09 2:46,Product2,3600,Visa,Family,Dubai,Dubayy,United Arab Emirates,1/8/09 1:19,2/8/09 2:06,25.2522222,55.28),(1/14/09 11:19,Product1,1200,Visa,Jennifer,Jumeira,Dubayy,United
 Arab Emirates,1/14/09 10:44,1/14/09 21:26,25.2097222,55.2477778),(1/10/09 7:55,Product1,1200,Visa,Laura,Jumeira,Dubayy,United Arab Emirates,1/7/09 9:47,2/1/09 10:18,25.2097222,55.2477778),(1/17/09 20:46
,Product2,3600,Visa,Michelle,Dubai,Dubayy,United Arab Emirates,4/13/08 2:36,2/8/09 2:12,25.2522222,55.28),(1/20/09 20:09,Product1,1200,Visa,Jeannette,Dubai,Dubayy,United Arab Emirates,1/18/09 9:58,2/23/0
9 7:01,25.2522222,55.28),(1/17/09 9:49,Product1,1200,Diners,Kathrin,Abu Dhabi,Abu Zaby,United Arab Emirates,2/27/08 20:02,1/30/09 8:45,24.4666667,54.3666667)})
grunt>
```

grunt>GroupByCountry = GROUP salesTable BY Country;



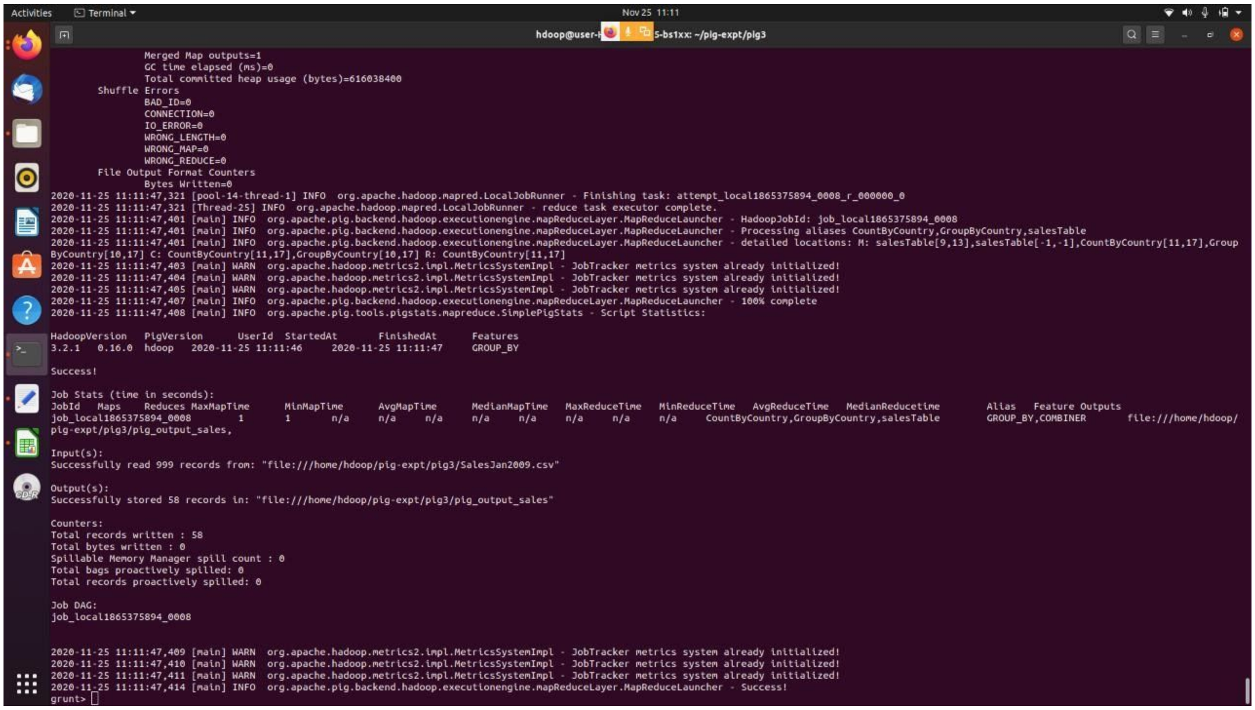
```
(1/23/09 11:23,Product1,1200,Mastercard,Lauren,Poughkeepsie                  ,NY,United States,8/26/08 9:37,2/26/09 12:17,41.70028,-73.92139)
(1/7/09 20:55,Product1,1200,Mastercard,Esther,Cincinnati                    ,OH,United States,1/7/09 20:25,2/26/09 12:24,39.16194,-84.45694)
(1/13/09 9:40,Product1,1200,Visa,Elizabeth,North Eastham                 ,MA,United States,1/26/05 19:09,2/26/09 12:45,41.865,-69.99167)
(1/25/09 15:18,Product2,3600,Visa,marie-louise,Sitka                         ,AK,United States,3/2/08 15:27,2/26/09 12:58,57.05306,-135.33)
(1/30/09 20:58,Product2,3600,Mastercard,Tanya,Carcross,Yukon Territory,Canada,1/30/09 14:05,2/26/09 13:06,60.1833333,-134.7166667)
(1/31/09 4:50,Product1,1200,Mastercard,Detlev,Taby,Stockholm,Sweden,1/12/09 1:03,2/26/09 13:12,59.5,18.05)
(1/20/09 10:05,Product2,3600,Visa,Stephen,Calgary,Alberta,Canada,1/9/09 10:36,2/26/09 13:28,51.0833333,-114.0833333)
(1/9/09 5:15,Product1,1200,Visa,Roza,Alcobendas,Madrid,Spain,9/9/08 9:08,2/26/09 14:24,40.5333333,-3.6333333)
(1/13/09 7:09,Product2,3600,Amex,John,Chicago                       ,IL,United States,1/8/09 13:49,2/26/09 15:14,41.85,-87.65)
(1/3/09 11:03,Product1,1200,Mastercard,paula,Sammamish                    ,WA,United States,1/1/09 22:43,2/26/09 15:30,47.64194,-122.07917)
(1/17/09 20:51,Product1,1200,Visa,Kathryn,Waiau Pa,Auckland,New Zealand,1/6/09 21:52,2/26/09 16:19,-37.1333333,174.75)
(1/5/09 7:55,Product1,1200,Mastercard,Isabelle,Ewa Beach                   ,HI,United States,12/31/08 7:40,2/26/09 17:47,21.31556,-158.00722)
(1/25/09 14:40,Product1,1200,Visa,Maureen,Statesville                 ,NC,United States,1/25/09 13:52,2/26/09 18:38,35.7825,-80.8875)
(1/25/09 17:55,Product1,1200,Visa,David,Kapaa                      ,HI,United States,1/24/09 12:18,2/26/09 20:21,22.07833,-159.32194)
(1/3/09 12:54,Product1,1200,Diners,caterina,Burbank                    ,CA,United States,12/30/08 8:39,2/26/09 20:22,34.18083,-118.30806)
(1/22/09 3:57,Product2,3600,Mastercard,Parth,Oegstgeest,Zuid-Holland,Netherlands,1/20/09 14:36,2/27/09 1:59,52.1833333,4.4666667)
(1/23/09 5:11,Product1,1200,Visa,Alexia,Blackrock,Dublin,Ireland,7/15/08 7:20,2/27/09 4:00,53.3030556,-6.1830556)
(1/25/09 11:52,Product2,3600,Mastercard,Gloria Gail ,WalcProduct3ee,Tyrol,Austria,1/25/09 5:03,2/27/09 5:37,47.65,12.3166667)
(1/10/09 4:12,Product1,1200,Visa,Gloria Gail ,Foxrock,Dublin,Ireland,1/7/09 7:39,2/27/09 6:33,53.2666667,-6.1741667)
(1/10/09 9:32,Product2,3600,Mastercard,Cheryl,Hilton Head                 ,SC,United States,1/10/05 21:39,2/27/09 7:33,32.21611,-80.75278)
(1/22/09 12:45,Product3,7500,Visa,Frank and Christelle,Valley Center                 ,CA,United States,1/22/09 10:25,2/27/09 10:49,33.21833,-117.03333)
(1/6/09 18:11,Product2,3600,Mastercard,Triona,Swampscott                  ,MA,United States,6/9/06 8:57,2/27/09 13:00,42.47083,-70.91806)
(1/19/09 5:56,Product1,1200,Visa,Allison,Edinburgh,Scotland,United Kingdom,7/27/08 11:58,2/27/09 14:35,55.95,-3.2)
(1/25/09 10:50,Product1,1200,Diners,Phillip,San Antonio                 ,TX,United States,5/29/08 19:51,2/27/09 14:44,29.42389,-98.49333)
(1/14/09 8:38,Product1,1200,Visa,jingyan,Bristol                    ,RI,United States,3/12/07 15:14,2/27/09 17:13,41.67694,-71.26667)
(        7:24,Product2,3600,Diners,clara,Perth,Western Australia,Australia,1/1/09 21:20,2/27/09 18:43,-31.9333333,115.8333333)
(        8:15,Product1,1200,Mastercard,Selma,Greenville                  ,TX,United States,1/5/09 18:16,2/27/09 19:07,33.13833,-96.11056)
(1/3/09 21:19,Product1,1200,Visa,Doug and Tina,Pls Vrds Est                  ,CA,United States,12/24/07 22:59,2/27/09 21:40,33.80056,-118.38917)
(1/4/09 5:13,Product1,1200,Visa,Lauren,Hradec Kralove,East Bohemia,Czech Republic,3/5/06 0:51,2/27/09 23:30,50.2116667,15.8441667)
(1/4/09 9:28,Product1,1200,Visa,Jen,Maidstone,England,United Kingdom,7/18/05 14:17,2/28/09 2:28,51.2666667,0.5166667)
(1/2/09 4:34,Product1,1200,Visa,gladys,Saint Albans,England,United Kingdom,1/10/06 4:20,2/28/09 3:43,51.75,-0.3333333)
(1/22/09 11:10,Product1,1200,Mastercard,Gustavo,Voluntari,Bucuresti,Romania,7/28/08 11:12,2/28/09 7:52,44.4666667,26.1333333)
(1/7/09 12:06,Product1,1200,Visa,Erica,Nadur, ,Malta,3/30/06 2:02,2/28/09 7:59,36.0377778,14.2941667)
(1/29/09 13:25,Product1,1200,Diners,Hale,Morrison                   ,CO,United States,11/3/05 18:14,2/28/09 8:27,39.65361,-105.19056)
(1/27/09 2:57,Product1,1200,Visa,Rosemary,Cobham,England,United Kingdom,1/27/09 2:50,2/28/09 9:22,51.3833333,0.4)
(1/7/09 13:19,Product1,1200,Visa,Darian,Izmir,Izmir,Turkey,8/26/08 7:33,2/28/09 9:54,38.4072222,27.1502778)
(1/23/09 10:04,Product1,1200,Mastercard,Kevin,Hollywood                   ,CA,United States,5/30/06 20:25,2/28/09 9:59,34.09833,-118.32583)
(1/19/09 4:55,Product1,1200,Visa,Alyssa,Brighton,England,United Kingdom,2/19/08 9:27,2/28/09 10:06,50.8333333,-0.15)
(1/28/09 22:02,Product1,1200,Visa,Hale,Hawera,Taranaki,New Zealand,1/23/09 22:31,2/28/09 12:43,-39.5916667,174.2833333)
(1/4/09 18:57,Product1,1200,Mastercard,KELI,Worongary,Queensland,Australia,12/23/08 15:17,2/28/09 14:00,-28.05,153.35)
(1/12/09 20:31,Product1,1200,Visa,Glen,Atlantida,Guatemala,Guatemala,1/6/09 16:53,2/28/09 14:39,14.65,-90.4833333)
(1/24/09 12:00,Product1,1200,Visa,T,El Escorial,Madrid,Spain,12/30/08 15:19,2/28/09 15:17,40.5833333,-4.1166667)
(1/28/09 11:19,Product1,1200,Visa,christal,Morrison                   ,CO,United States,6/20/04 17:16,2/28/09 17:18,39.65361,-105.19056)
(1/7/09 17:48,Product1,1200,Mastercard,Alex,Augusta                    ,GA,United States,6/10/05 20:25,2/28/09 19:57,33.51722,-82.07583)
(1/23/09 12:42,Product2,3600,Mastercard,Anke,Avalon,New South Wales,Australia,3/3/08 17:38,2/28/09 22:26,-33.6333333,151.3333333)
(1/7/09 19:48,Product2,3600,Mastercard,TRICIA,Sydney,New South Wales,Australia,9/21/08 20:49,3/1/09 0:14,-33.8833333,151.2166667)
(1/26/09 11:19,Product1,1200,Mastercard,smith,Lahti,Etela-Suomen Laani,Finland,1/4/09 5:25,3/1/09 0:39,60.9666667,25.6666667)
(1/5/09 13:23,Product1,1200,Visa,Macy,Inner City,Vienna,Austria,1/5/09 11:28,3/1/09 2:28,48.2166667,16.3666667)
(1/26/09 13:41,Product1,1200,Mastercard,Lesleigh,Baden,Aargau,Switzerland,10/23/05 9:23,3/1/09 3:11,47.4666667,8.3)
(1/20/09 10:42,Product2,3600,Diners,esther,Huddersfield,England,United Kingdom,1/20/09 9:15,3/1/09 3:29,53.65,-1.7833333)
(1/22/09 14:25,Product1,1200,Visa,Hans-Joerg,Belfast,Northern Ireland,United Kingdom,11/10/08 12:15,3/1/09 3:37,54.5833333,-5.9333333)
(1/28/09 5:36,Product2,3600,Visa,Christiane,Black River,Black River,Mauritius,1/9/09 8:10,3/1/09 4:40,-20.3602778,57.3661111)
(1/1/09 4:24,Product3,7500,Amex,Pamela,Skaneateles                 ,NY,United States,12/28/08 17:28,3/1/09 7:21,42.94694,-76.42944)
(1/8/09 11:55,Product1,1200,Diners,julie,Haverhill,England,United Kingdom,11/29/06 13:31,3/1/09 7:28,52.0833333,0.4333333)
(1/12/09 21:30,Product1,1200,Visa,Julia ,Madison                    ,WI,United States,11/17/08 22:24,3/1/09 10:14,43.07306,-89.40111)
grunt> 
```

grunt>CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0,CONCAT(':',(chararray)COUNT($1)));



```
(Italy:15)
(Japan:2)
(Malta:2)
(Spain:12)
(Brazil:5)
(Canada:76)
(France:27)
(Greece:1)
(Israel:1)
(Jersey:1)
(Kuwait:1)
(Latvia:1)
(Monaco:2)
(Norway:16)
(Poland:2)
(Russia:1)
(Sweden:13)
(Turkey:6)
(Austria:7)
(Bahrain:1)
(Belgium:8)
(Bermuda:1)
(Country:1)
(Denmark:15)
(Finland:2)
(Germany:25)
(Hungary:3)
(Iceland:1)
(Ireland:49)
(Moldova:1)
(Romania:1)
(Ukraine:1)
(Bulgaria:1)
(Malaysia:1)
(Thailand:2)
(Argentina:1)
(Australia:38)
(Guatemala:1)
(Hong Kong:1)
(Mauritius:1)
(Costa Rica:1)
(Luxembourg:1)
(Cayman Isls:1)
(Netherlands:22)
(New Zealand:6)
(Philippines:2)
(South Korea:1)
(Switzerland:36)
(The Bahamas:2)
(South Africa:5)
(United States:462)
(Czech Republic:3)
(United Kingdom:100)
(Dominican Republic:1)
(United Arab Emirates:6)
grunt> 
```

grunt>STORE CountByCountry INTO 'pig_output_sales' USING PigStorage('\t');



```
		Merged Map outputs=1
		GC time elapsed (ms)=0
		Total committed heap usage (bytes)=616038400
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
		IO_ERROR=0
		WRONG_LENGTH=0
		WRONG_MAP=0
		WRONG_REDUCE=0
	File Output Format Counters
		Bytes Written=0
2020-11-25 11:11:47,321 [pool-14-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1865375894_0008_r_000000_0
2020-11-25 11:11:47,321 [Thread-25] INFO  org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
2020-11-25 11:11:47,401 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local1865375894_0008
2020-11-25 11:11:47,401 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases CountByCountry,GroupByCountry,salesTable
2020-11-25 11:11:47,401 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: salesTable[9,13],salesTable[-1,-1],CountByCountry[11,17],Group
ByCountry[10,17] C: CountByCountry[11,17],GroupByCountry[10,17] R: CountByCountry[11,17]
2020-11-25 11:11:47,403 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:11:47,404 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:11:47,405 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:11:47,407 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-11-25 11:11:47,408 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features
3.2.1	0.16.0	hdoop	2020-11-25 11:11:46	2020-11-25 11:11:47	GROUP_BY

Success!

Job Stats (time in seconds):
JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReducetime	Alias	Feature	Outputs
job_local1865375894_0008	1	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	CountByCountry,GroupByCountry,salesTable	GROUP_BY,COMBINER	file:///home/hdoop/
pig-expt/pig3/pig_output_sales,

Input(s):
Successfully read 999 records from: "file:///home/hdoop/pig-expt/pig3/SalesJan2009.csv"

Output(s):
Successfully stored 58 records in: "file:///home/hdoop/pig-expt/pig3/pig_output_sales"

Counters:
Total records written : 58
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1865375894_0008


2020-11-25 11:11:47,409 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:11:47,410 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:11:47,411 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2020-11-25 11:11:47,414 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

Problem Statement-8: Illustrate the concept of bucketing and partitioning and bucketing in hive.

Description:

Partitioning – Apache Hive organizes tables into partitions for grouping the same type of data together based on a column or partition key. Each table in the hive can have one or more partition keys to identify a particular partition. Using partition we can make it faster to do queries on slices of the data.

The Hive command for Partitioning is:

CREATE TABLE table_name (column1 data_type, column2 data_type) PARTITIONED BY (partition1 data_type, partition2 data_type,….);

Bucketing – In Hive Tables or partitions are subdivided into buckets based on the hash function of a column in the table to give extra structure to the data that may be used for more efficient queries.

The Hive command for Bucketing is:

CREATE TABLE table_name PARTITIONED BY (partition1 data_type, partition2 data_type,….) CLUSTERED BY (column_name1, column_name2, …) SORTED BY (column_name [ASC|DESC], …)] INTO num_buckets BUCKETS;

Code:

set hive.exec.dynamic.partition.mode=nonstrict;

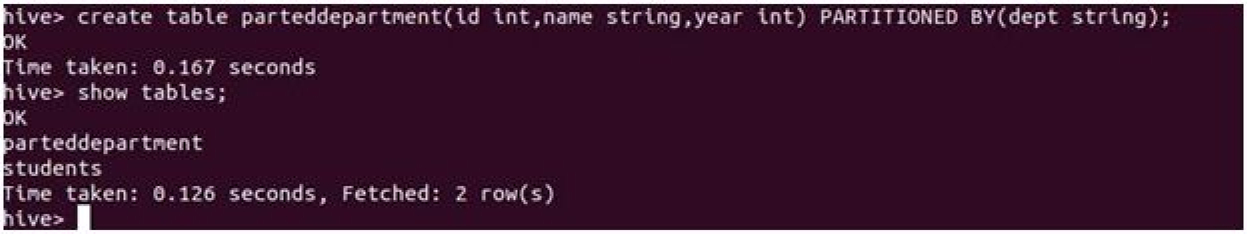
Create table students (id int, name string, year int, dept string) row format delimited fields terminated by „,‟; Load data inpath „hdfs://localhost:54310/lab10/data‟ into table students;

Create table parteddepartment (id int, name string, year int) PARTITIONED by (dept string) row format delimited fields terminated by „,‟;

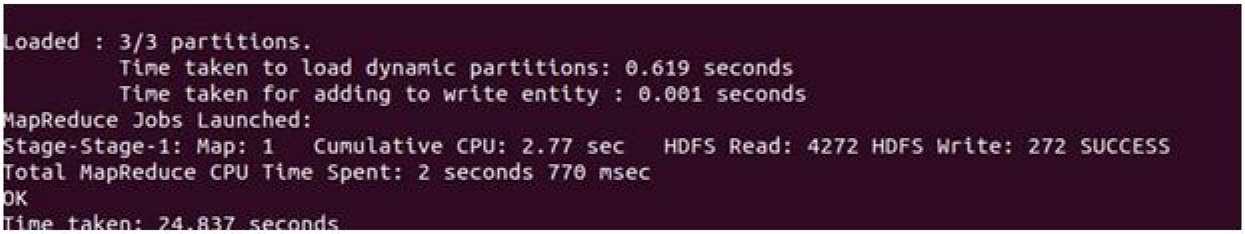
insert overwrite table parteddepartment PARTITION(dept) SELECT id,name,year,dept from students;

Create table samplebucket (id int, name string, year int) clustered by (name) into 2 buckets row format delimited fields terminated by „,‟;

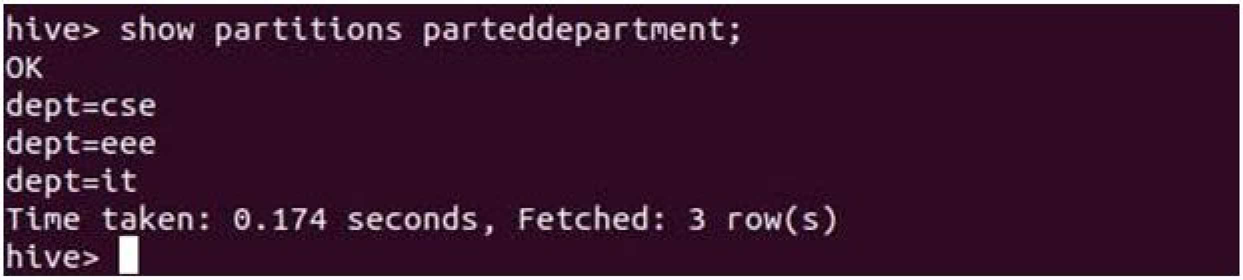
From parteddepartment insert overwrite table samplebucket select id,name,year;



```
hive> create table parteddepartment(id int,name string,year int) PARTITIONED BY(dept string);
OK
Time taken: 0.167 seconds
hive> show tables;
OK
parteddepartment
students
Time taken: 0.126 seconds, Fetched: 2 row(s)
hive> 
```



```
Loaded : 3/3 partitions.
         Time taken to load dynamic partitions: 0.619 seconds
         Time taken for adding to write entity : 0.001 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.77 sec   HDFS Read: 4272 HDFS Write: 272 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 770 msec
OK
Time taken: 24.837 seconds
```



```
hive> show partitions parteddepartment;
OK
dept=cse
dept=eee
dept=it
Time taken: 0.174 seconds, Fetched: 3 row(s)
hive> █
```

Problem Statement-9: From the given Aadhaar dataset , write the hive queries for the following.

a. Count the number of Identities generated in each state
b. Count the number of Identities generated by each Enrolment Agency
c. For how many states the Aadhaar data exists?
d. What is the % of Aadhaar being approved per state?

Description:
Hive provides SQL type querying language for the ETL purpose on top of Hadoop file system.

Hive Query language (HiveQL) provides SQL type environment in Hive to work with tables, databases, queries.

We can have a different type of Clauses associated with Hive to perform different type data manipulations and querying. For better connectivity with different nodes outside the environment. HIVE provides JDBC connectivity as well.

Hive queries provides the following features:

- Data modeling such as Creation of databases, tables, etc.
- ETL functionalities such as Extraction, Transformation, and Loading data into tables
- Joins to merge different data tables
- User specific custom scripts for ease of code
- Faster querying tool on top of Hadoop

Code:

create table aad(register string,enrolment_agency string,state string,district string,sub_district string,pincode int,gender string,age int,aad_generated int,enroll_rejected int,res_providing_email int,res_providing_number int)row format delimited fields terminated by ',' stored as textfile;

160117737006
N.Durga Sai Lakshmi

load data local inpath '/home/ak/Desktop/adata.txt' overwrite into table aad;

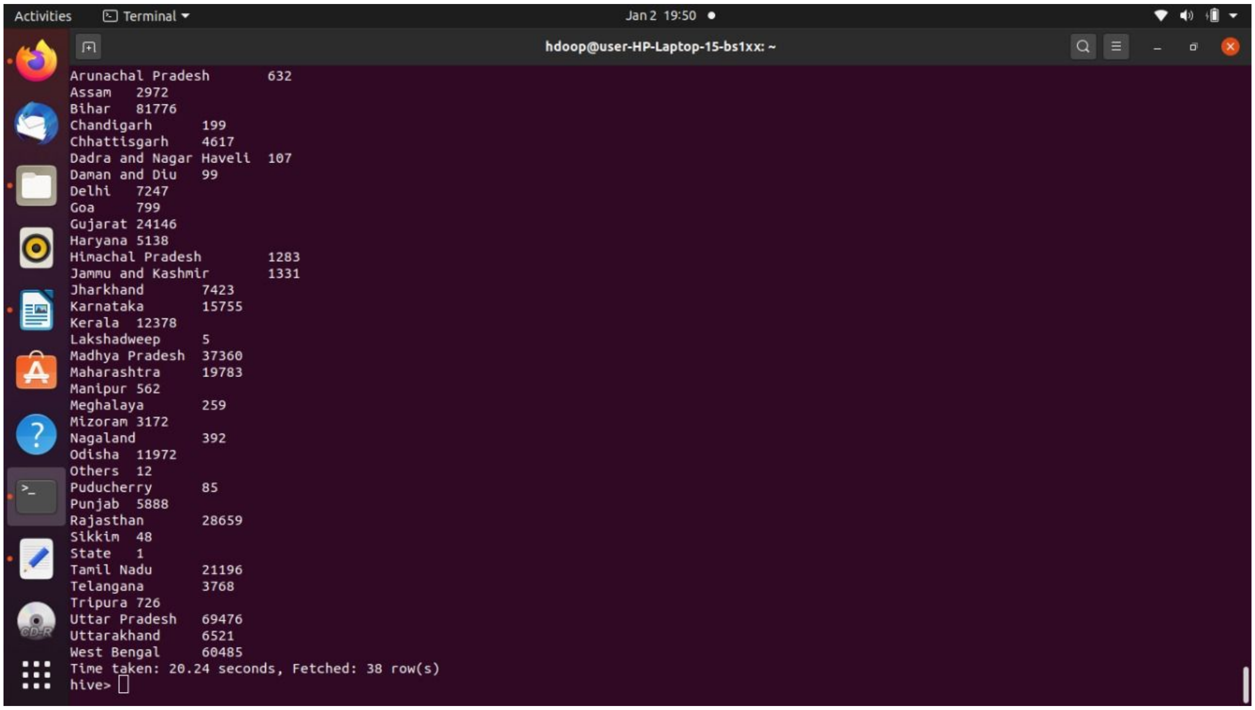
a. select state,count(*) from aad group by state;

b. select enrolment_agency,count(*) from aad group by enrolment_agency;

c. select count(distinct state) from aad;

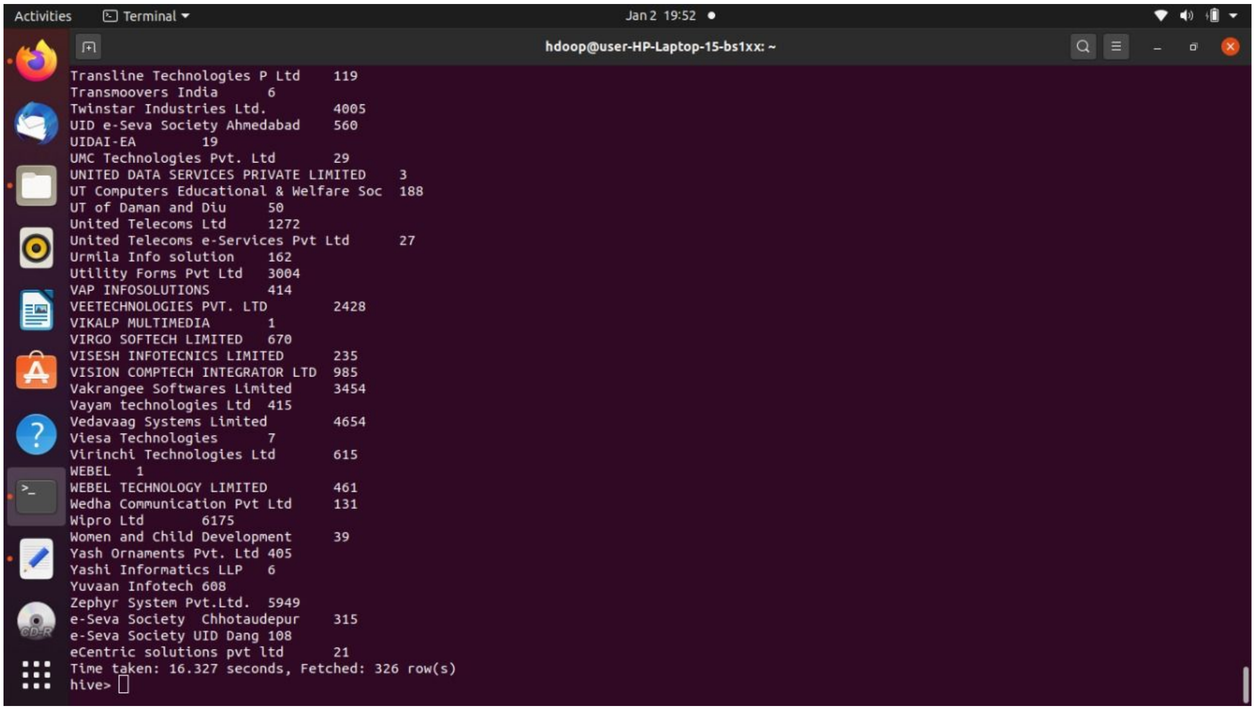
d. select state,((sum(aad_generated)-sum(enroll_rejected))/(sum(aad_generated))*100) from aad group by state;

OUTPUTS:
a)



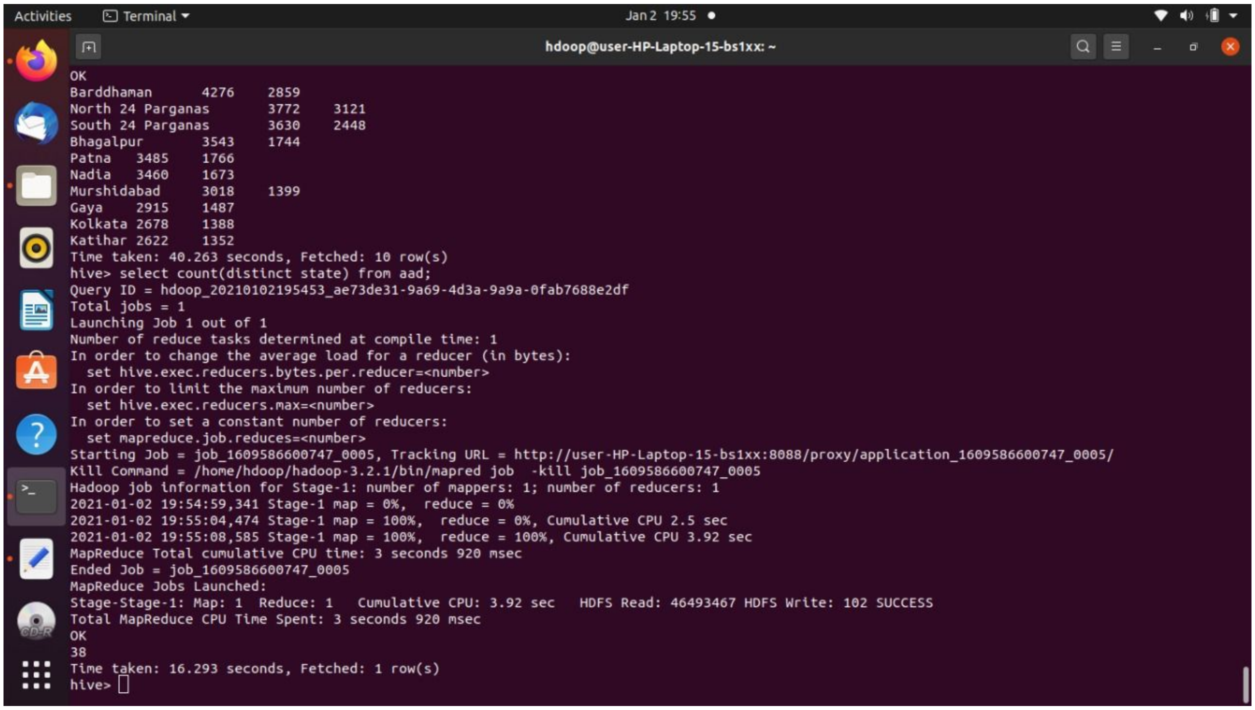
```
Arunachal Pradesh       632
Assam   2972
Bihar   81776
Chandigarh      199
Chhattisgarh    4617
Dadra and Nagar Haveli  107
Daman and Diu   99
Delhi   7247
Goa     799
Gujarat 24146
Haryana 5138
Himachal Pradesh        1283
Jammu and Kashmir       1331
Jharkhand       7423
Karnataka       15755
Kerala  12378
Lakshadweep     5
Madhya Pradesh  37360
Maharashtra     19783
Manipur 562
Meghalaya       259
Mizoram 3172
Nagaland        392
Odisha  11972
Others  12
Puducherry      85
Punjab  5888
Rajasthan       28659
Sikkim  48
State   1
Tamil Nadu      21196
Telangana       3768
Tripura 726
Uttar Pradesh   69476
Uttarakhand     6521
West Bengal     60485
Time taken: 20.24 seconds, Fetched: 38 row(s)
hive> 
```

b)



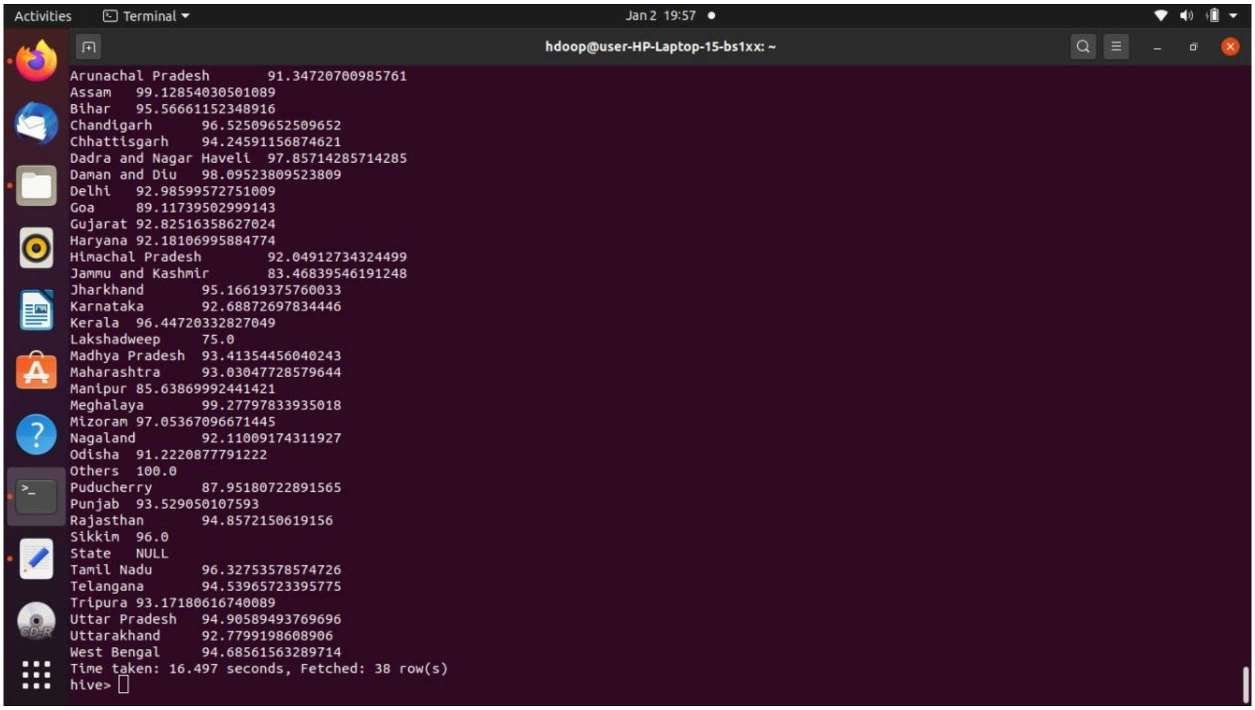
```
Transline Technologies P Ltd    119
Transmoovers India      6
Twinstar Industries Ltd.        4005
UID e-Seva Society Ahmedabad    560
UIDAI-EA        19
UMC Technologies Pvt. Ltd       29
UNITED DATA SERVICES PRIVATE LIMITED    3
UT Computers Educational & Welfare Soc  188
UT of Daman and Diu     50
United Telecoms Ltd     1272
United Telecoms e-Services Pvt Ltd      27
Urmila Info solution    162
Utility Forms Pvt Ltd   3004
VAP INFOSOLUTIONS       414
VEETECHNOLOGIES PVT. LTD        2428
VIKALP MULTIMEDIA       1
VIRGO SOFTECH LIMITED   670
VISESH INFOTECNICS LIMITED      235
VISION COMPTECH INTEGRATOR LTD  985
Vakrangee Softwares Limited     3454
Vayam technologies Ltd  415
Vedavaag Systems Limited        4654
Viesa Technologies      7
Virinchi Technologies Ltd       615
WEBEL   1
WEBEL TECHNOLOGY LIMITED        461
Wedha Communication Pvt Ltd     131
Wipro Ltd       6175
Women and Child Development     39
Yash Ornaments Pvt. Ltd 405
Yashi Informatics LLP   6
Yuvaan Infotech 608
Zephyr System Pvt.Ltd.  5949
e-Seva Society  Chhotaudepur    315
e-Seva Society UID Dang 108
eCentric solutions pvt ltd      21
Time taken: 16.327 seconds, Fetched: 326 row(s)
hive> 
```

c)



```
OK
Barddhaman      4276    2859
North 24 Parganas       3772    3121
South 24 Parganas       3630    2448
Bhagalpur       3543    1744
Patna   3485    1766
Nadia   3460    1673
Murshidabad     3018    1399
Gaya    2915    1487
Kolkata 2678    1388
Katihar 2622    1352
Time taken: 40.263 seconds, Fetched: 10 row(s)
hive> select count(distinct state) from aad;
Query ID = hdoop_20210102195453_ae73de31-9a69-4d3a-9a9a-0fab7688e2df
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_1609586600747_0005, Tracking URL = http://user-HP-Laptop-15-bs1xx:8088/proxy/application_1609586600747_0005/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1609586600747_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-01-02 19:54:59,341 Stage-1 map = 0%,  reduce = 0%
2021-01-02 19:55:04,474 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2021-01-02 19:55:08,585 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.92 sec
MapReduce Total cumulative CPU time: 3 seconds 920 msec
Ended Job = job_1609586600747_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.92 sec   HDFS Read: 46493467 HDFS Write: 102 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 920 msec
OK
38
Time taken: 16.293 seconds, Fetched: 1 row(s)
hive> 
```

d)



```
Arunachal Pradesh       91.34720700985761
Assam   99.12854030501089
Bihar   95.56661152348916
Chandigarh      96.52509652509652
Chhattisgarh    94.24591156874621
Dadra and Nagar Haveli  97.85714285714285
Daman and Diu   98.09523809523809
Delhi   92.98599572751009
Goa     89.11739502999143
Gujarat 92.82516358627024
Haryana 92.18106995884774
Himachal Pradesh        92.04912734324499
Jammu and Kashmir       83.46839546191248
Jharkhand       95.16619375760033
Karnataka       92.68872697834446
Kerala  96.44720332827049
Lakshadweep     75.0
Madhya Pradesh  93.41354456040243
Maharashtra     93.03047728579644
Manipur 85.63869992441421
Meghalaya       99.27797833935018
Mizoram 97.05367096671445
Nagaland        92.11009174311927
Odisha  91.2220877791222
Others  100.0
Puducherry      87.95180722891565
Punjab  93.529050107593
Rajasthan       94.8572150619156
Sikkim  96.0
State   NULL
Tamil Nadu      96.32753578574726
Telangana       94.53965723395775
Tripura 93.17180616740089
Uttar Pradesh   94.90589493769696
Uttarakhand     92.7799198608906
West Bengal     94.68561563289714
Time taken: 16.497 seconds, Fetched: 38 row(s)
hive> 
```

Problem Statement-10: Implement the word count job in Scala.

Description:
Scala is a general-purpose programming language. It supports object oriented, functional and imperative programming approaches. It is a strong static type language. In scala, everything is an object whether it is a function or a number. It does not have a concept of primitive data.

Procedure:

```
hdfs dfs -mkdir /spark
hdfs dfs -put /home/kgr/sparkdata.txt /spark
```

Code:

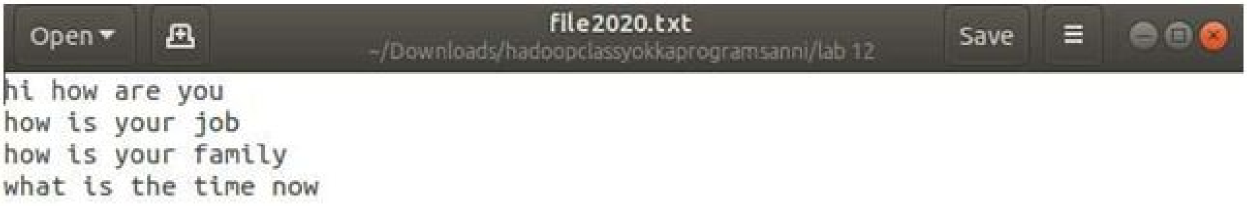
```
scala> val data=sc.textFile("sparkdata.txt");
val splitdata = data.flatMap(line => line.split(""));
val mapdata = splitdata.map(word => (word,1));
val reducedata = mapdata.reduceByKey(_+_);
reducedata.collect;
```

160117737006
N.Durga Sai Lakshmi

For dataset:

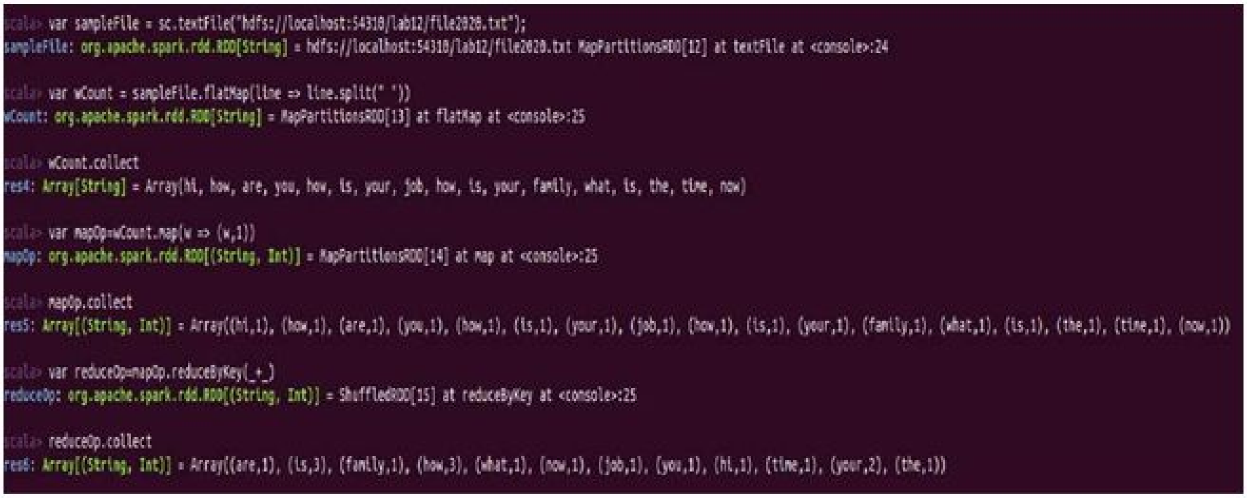
val text = sc.textFile("mytextfile.txt")
val counts = text.flatMap(line => line.split(" ")).map(word => (word,1)).reduceByKey(_+_)
counts.collect

Input file:



```
hi how are you
how is your job
how is your family
what is the time now
```

Outputs:



```
scala> var sampleFile = sc.textFile("hdfs://localhost:54310/lab12/file2020.txt");
sampleFile: org.apache.spark.rdd.RDD[String] = hdfs://localhost:54310/lab12/file2020.txt MapPartitionsRDD[12] at textFile at <console>:24

scala> var wCount = sampleFile.flatMap(line => line.split(" "))
wCount: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[13] at flatMap at <console>:25

scala> wCount.collect
res4: Array[String] = Array(hi, how, are, you, how, is, your, job, how, is, your, family, what, is, the, time, now)

scala> var mapOp=wCount.map(w => (w,1))
mapOp: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[14] at map at <console>:25

scala> mapOp.collect
res5: Array[(String, Int)] = Array((hi,1), (how,1), (are,1), (you,1), (how,1), (is,1), (your,1), (job,1), (how,1), (is,1), (your,1), (family,1), (what,1), (is,1), (the,1), (time,1), (now,1))

scala> var reduceOp=mapOp.reduceByKey(_+_)
reduceOp: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[15] at reduceByKey at <console>:25

scala> reduceOp.collect
res6: Array[(String, Int)] = Array((are,1), (is,3), (family,1), (how,3), (what,1), (now,1), (job,1), (you,1), (hi,1), (time,1), (your,2), (the,1))
```