

19162121031

SMIT R PATEL

PRACTICAL 6

BIG DATA AND ANALYTICS

CODES OR COMMANDS:-

```
$hadoop fs -mkdir smitrpatel19 ## make folder which name smitrpatel19
```

```
$hadoop fs -put SalesJan2009.csv smitrpatel19 ## move csv file into folder
```

```
$hadoop fs -ls smitrpatel19 ## for go inside to folder
```

new terminal and type --> pig

```
grunt> salesTable = LOAD 'smitrpatel19/SalesJan2009.csv' USING PigStorage(',') AS  
(Transaction_date:chararray,Product:chararray,Price:chararray,Payment_Type:chararray,Name:char  
array,City:chararray,State:chararray,Country:chararray,Account_Created:chararray,Last_Login:chara  
rray,Latitude:chararray,Longitude:chararray);
```

```
grunt> GroupbyCountry = GROUP salesTable by Country; ##load the file
```

```
grunt> CountbyCountry = FOREACH GroupbyCountry GENERATE  
CONCAT((chararray)$0,CONCAT(':', (chararray)COUNT($1)));
```

```
grunt> STORE CountbyCountry INTO 'pig_output_sales1' USING PigStorage('\t');
```

new terminal and type -->

```
$hdfs dfs -cat pig_output_sales1/part-r-00000
```

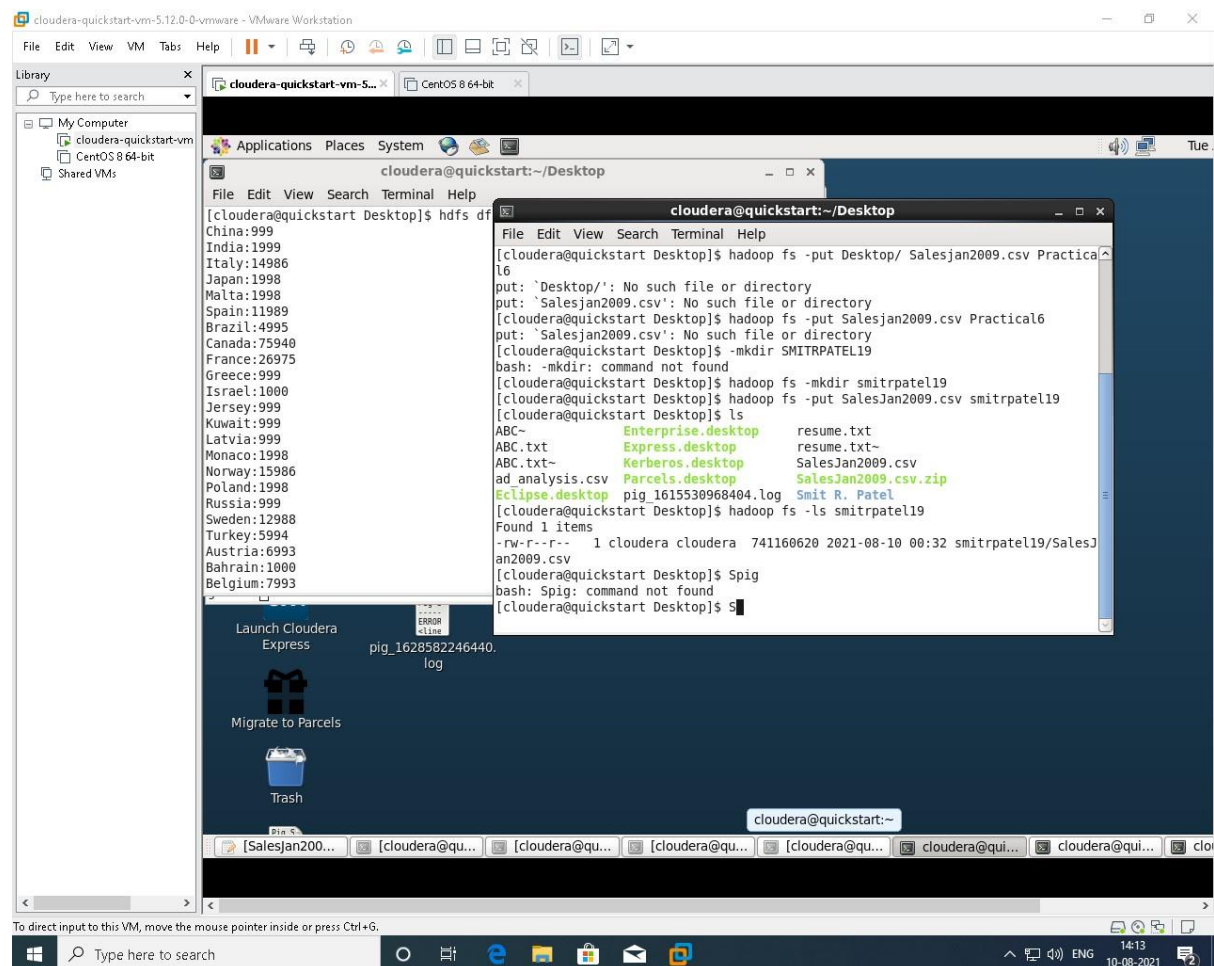
Code with output :-

Step 1 :-

\$hadoop fs -mkdir smit Patel19 ## make folder which name smit Patel19

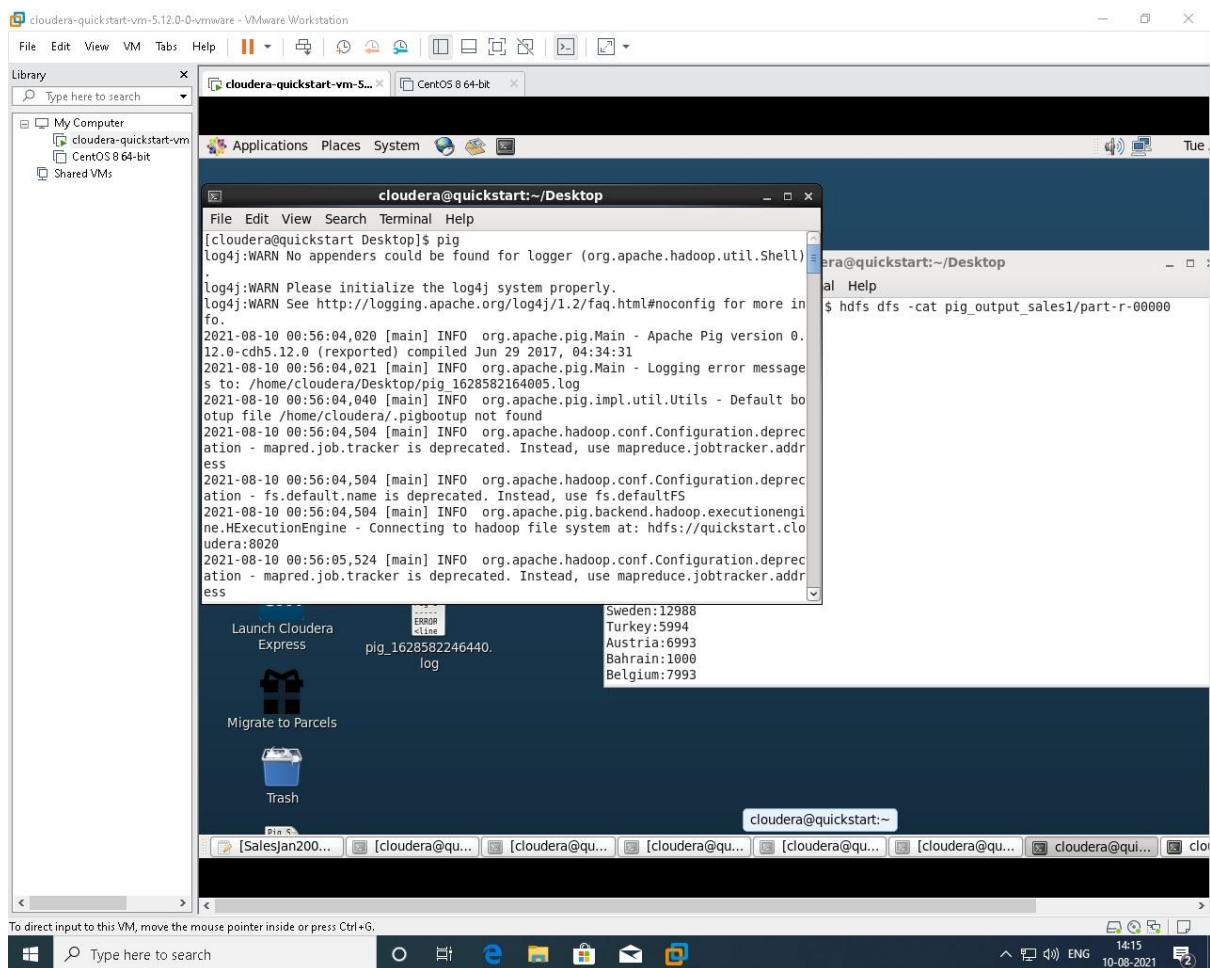
\$hadoop fs -put SalesJan2009.csv smit Patel19 ## move csv file into folder

\$hadoop fs -ls smit Patel19 ## for go inside to folder



Step 2 :-

new terminal and type --> pig

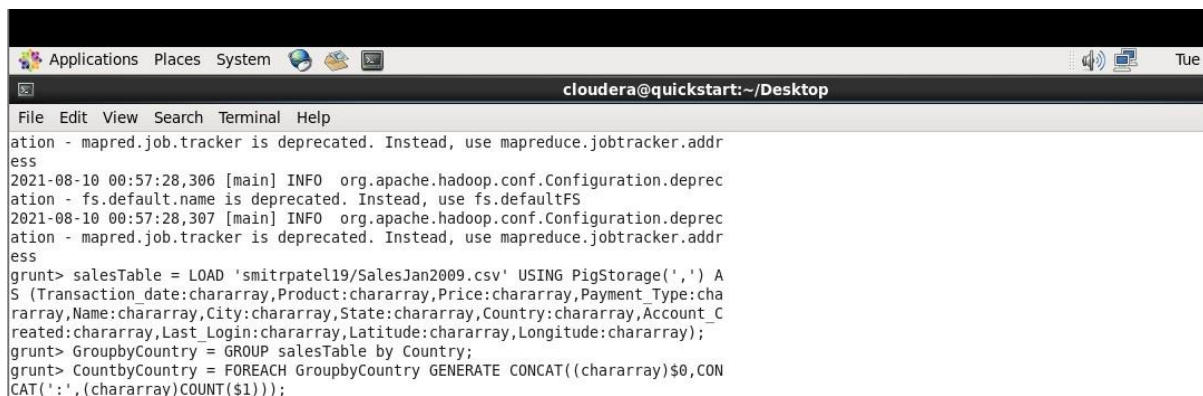


Step 3 :-

```
grunt> salesTable = LOAD 'smitrpatel19/SalesJan2009.csv' USING PigStorage(',') AS  
(Transaction_date:chararray,Product:chararray,Price:chararray,Payment_Type:chararray,Name:char  
array,City:chararray,State:chararray,Country:chararray,Account_Created:chararray,Last_Login:chara  
rray,Latitude:chararray,Longitude:chararray);
```

```
grunt> GroupbyCountry = GROUP salesTable by Country; ##load the file
```

```
grunt> CountbyCountry = FOREACH GroupbyCountry GENERATE  
CONCAT((chararray)$0,CONCAT(':', (chararray)COUNT($1)));
```

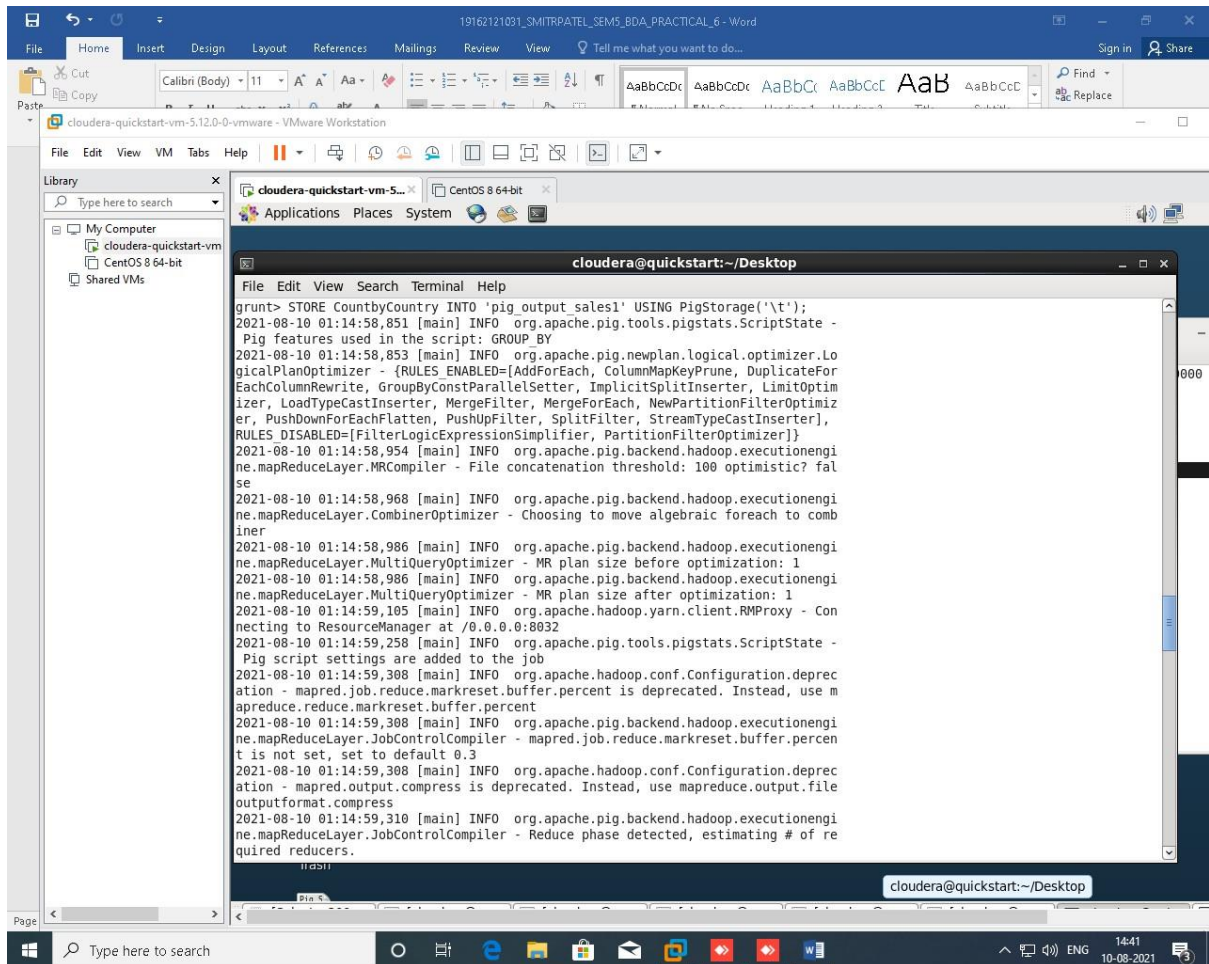


The screenshot shows a terminal window titled "cloudera@quickstart: ~/Desktop". The terminal displays the execution of Pig Latin commands. It starts with a deprecation warning about "mapred.job.tracker" and "fs.default.name". Then, the command to load the CSV file is executed, followed by the GROUP and FOREACH commands. The output shows the commands being processed successfully.

```
File Edit View Search Terminal Help  
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr  
ess  
2021-08-10 00:57:28,306 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2021-08-10 00:57:28,307 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr  
ess  
grunt> salesTable = LOAD 'smitrpatel19/SalesJan2009.csv' USING PigStorage(',') A  
S (Transaction_date:chararray,Product:chararray,Price:chararray,Payment_Type:cha  
rarray,Name:chararray,City:chararray,State:chararray,Country:chararray,Account_C  
reated:chararray,Last_Login:chararray,Latitude:chararray,Longitude:chararray);  
grunt> GroupbyCountry = GROUP salesTable by Country;  
grunt> CountbyCountry = FOREACH GroupbyCountry GENERATE CONCAT((chararray)$0,CON  
CAT(':', (chararray)COUNT($1)));
```

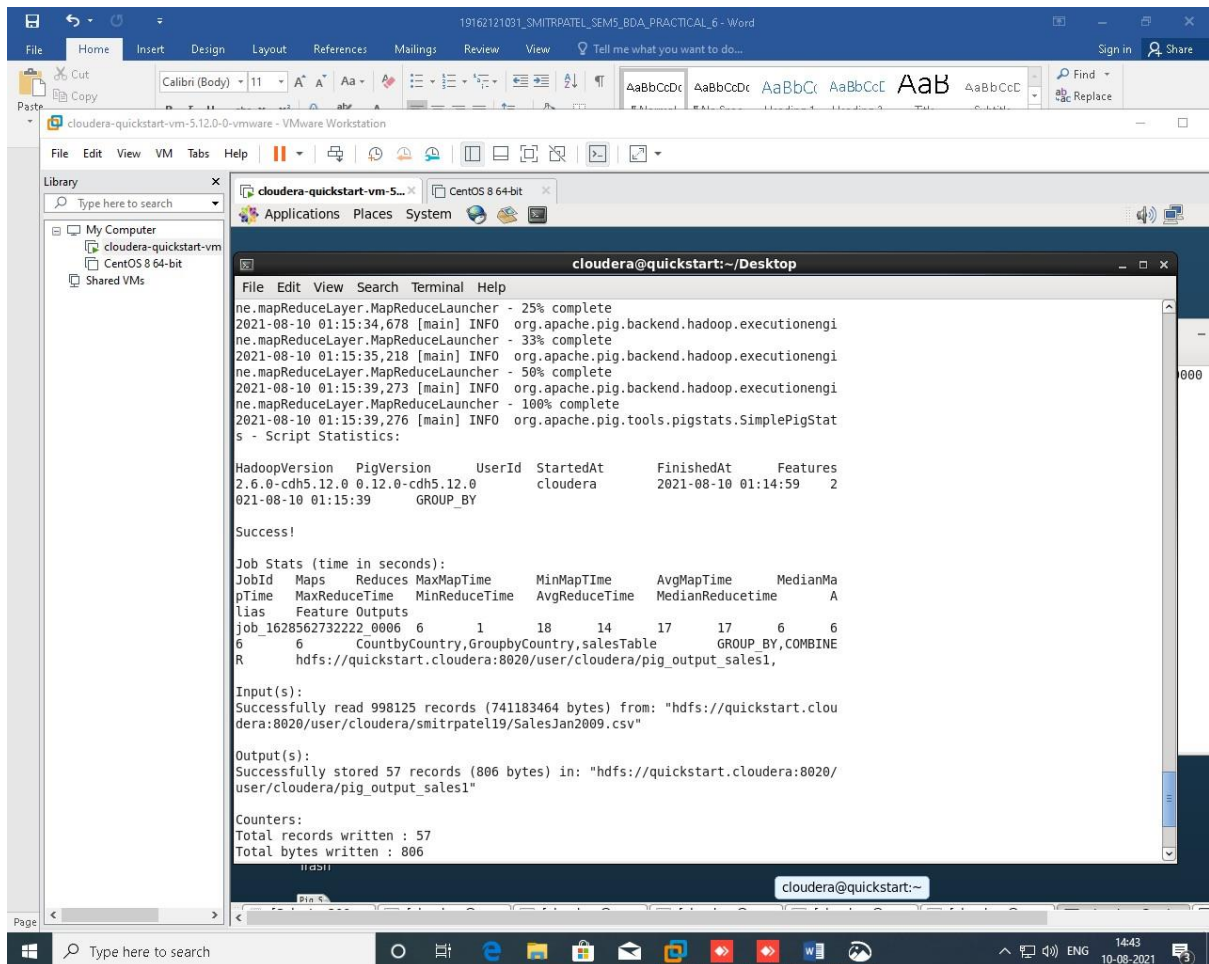
Step 4 :-

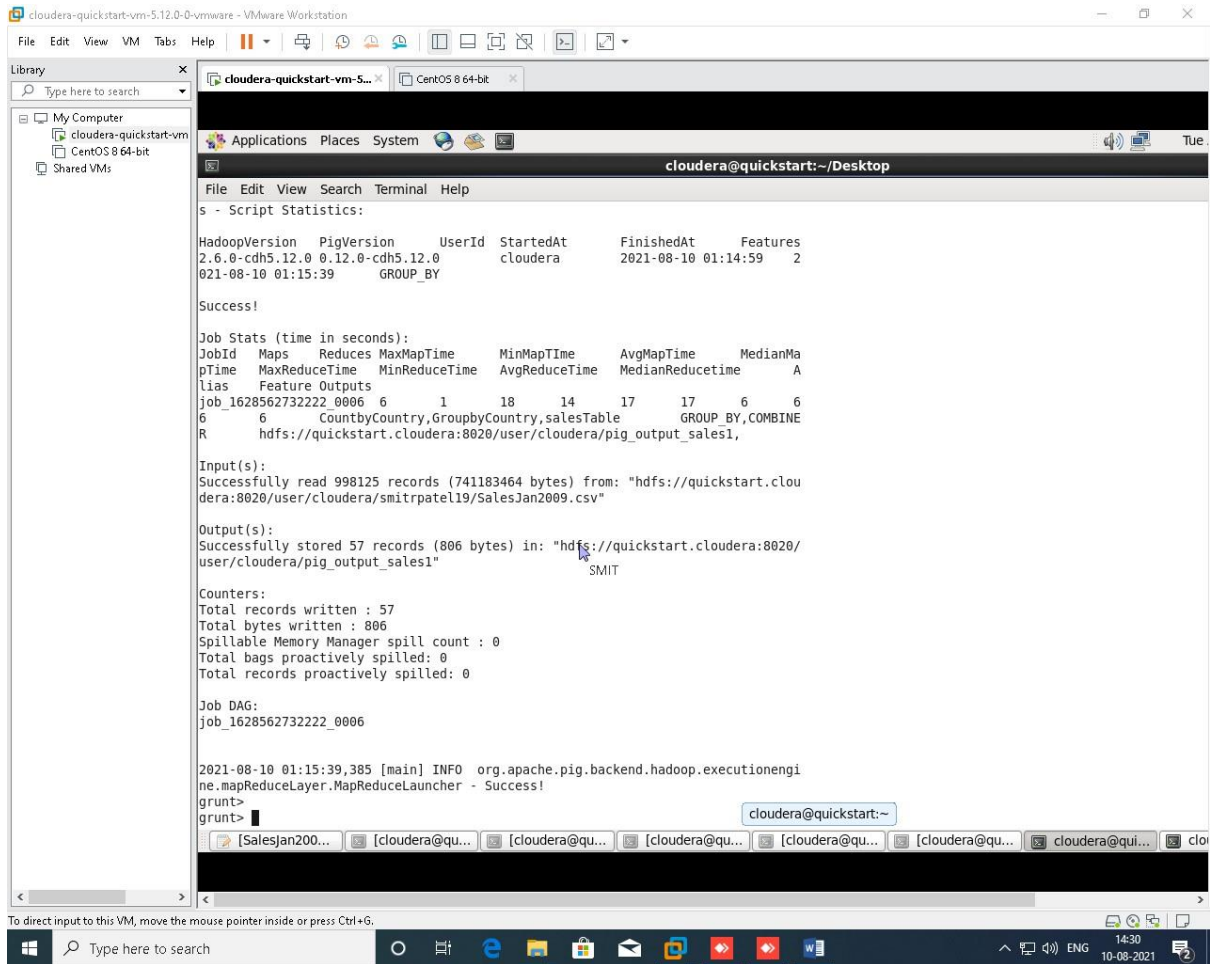
grunt> STORE CountbyCountry INTO 'pig_output_sales1' USING PigStorage('\t');



The screenshot shows a Windows desktop environment with a VMware Workstation window titled 'cloudera-quickstart-vm-5.12.0-0-vmware'. Inside the VM, a terminal window titled 'cloudera@quickstart:~/Desktop' is open, displaying the output of the command 'grunt> STORE CountbyCountry INTO 'pig_output_sales1' USING PigStorage('\t');'. The output shows various log messages from the Pig system, including information about the script state, optimizer rules, and the execution of the STORE command. The terminal window is overlaid on a file explorer showing the 'cloudera-quickstart-vm-5...' directory. The Windows taskbar at the bottom shows the time as 14:41 on 10-08-2021.

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
grunt> STORE CountbyCountry INTO 'pig_output_sales1' USING PigStorage('\t');
2021-08-10 01:14:58,851 [main] INFO org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: GROUP_BY
2021-08-10 01:14:58,853 [main] INFO org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-10 01:14:58,954 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-10 01:14:58,968 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to comb
iner
2021-08-10 01:14:58,986 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-10 01:14:58,986 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-10 01:14:59,105 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2021-08-10 01:14:59,258 [main] INFO org.apache.pig.tools.pigstats.ScriptState -
Pig script settings are added to the job
2021-08-10 01:14:59,308 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use m
apreduce.reduce.markreset.buffer.percent
2021-08-10 01:14:59,308 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
t is not set, set to default 0.3
2021-08-10 01:14:59,308 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.output.compress is deprecated. Instead, use mapreduce.output.file
outputformat.compress
2021-08-10 01:14:59,310 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of re
quired reducers.
11d511
cloudera@quickstart:~/Desktop
```

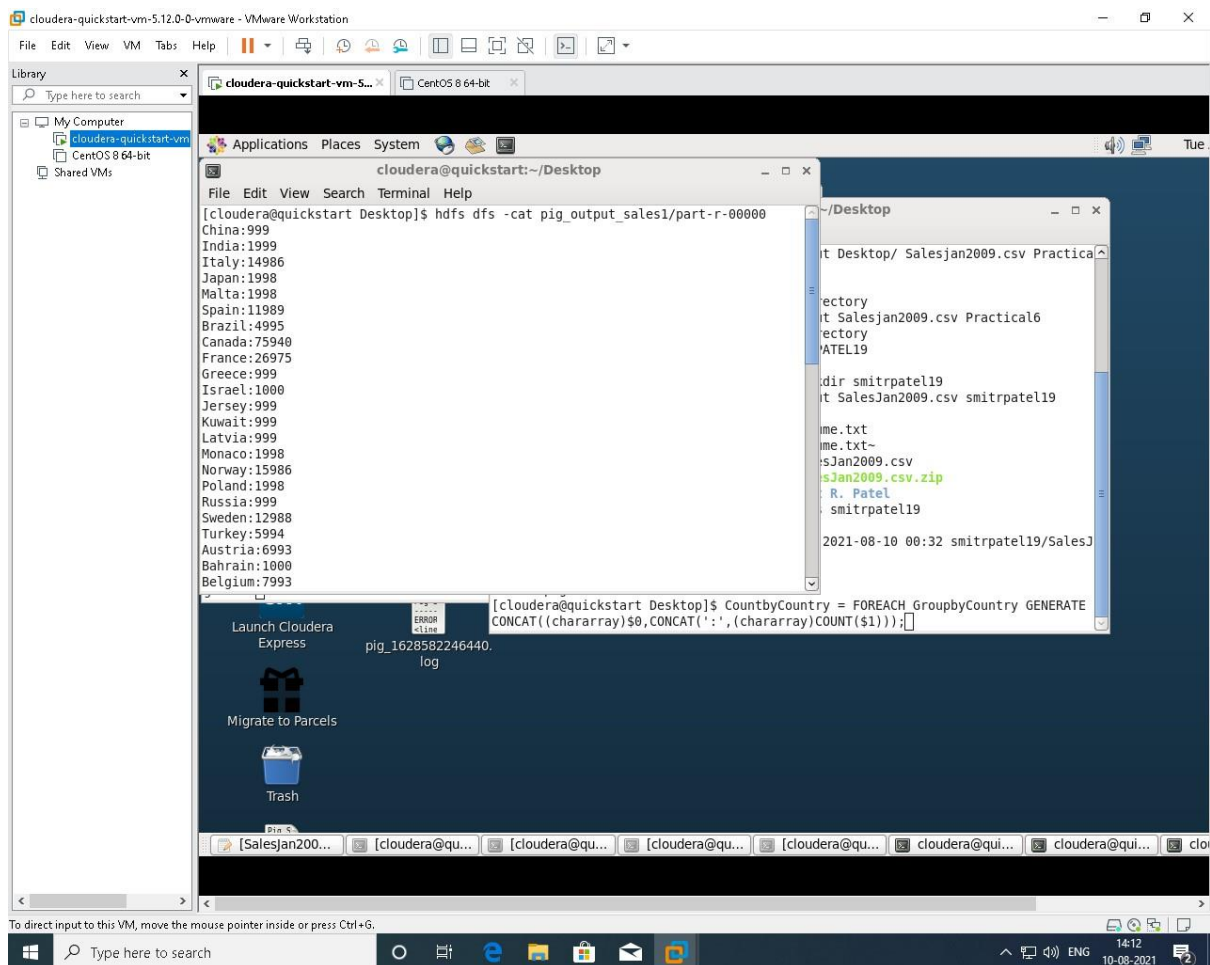




Final Step :-

new terminal and type -->

```
$hdfs dfs -cat pig_output_sales1/part-r-00000
```



Conclusion :-

Here, In this practical we learn load the file with all content – data and sort group data by field country after we generate results we show data flow in the directory 'pig_output_sales' and on hdfs, we get success all map reduce work flow and get final output.