

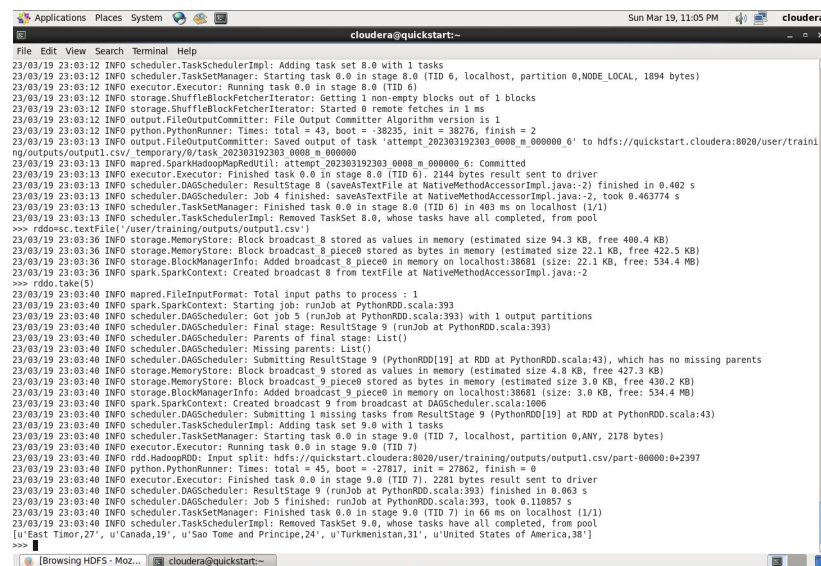
PYSPARK SALES ANALYSIS PROJECT

1] Display the number of countries present in the data

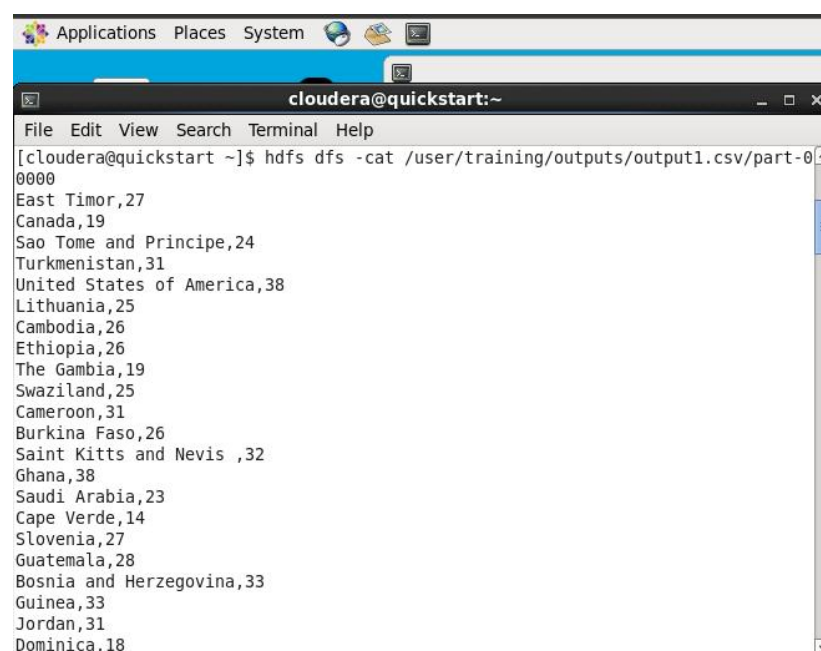
```

rdd1=sc.textFile('/user/training/sales.csv')
rdd1.take(5)
rdd2=rdd1.map(lambda x:x.split(',')).map(lambda x:(x[1],1)).reduceByKey(lambda a,b:a+b)
rdd2.getNumPartitions()
rdd2=rdd2.coalesce(1)
rdd2.getNumPartitions()
def toCSVLine(data):
    return ','.join(str(d) for d in data)

rdd3=rdd2.map(toCSVLine)
rdd3.take(5)
rdd3.saveAsTextFile('/user/training/outputs/output1.csv')
```



The screenshot shows a terminal window titled 'cloudera@quickstart:~'. It displays a series of Spark logs for a PySpark job. The logs indicate the execution of a map-reduce operation to count countries. Key log messages include: 'INFO scheduler.TaskSchedulerImpl: Adding task set 8.0 with 1 tasks', 'INFO executor.Executor: Running task 0.0 in stage 8.0 (TID 6)', 'INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms', 'INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1', 'INFO output.FileOutputCommitter: Saved output of task 'attempt_202303192303_0008_m_000000_6' to hdfs://quickstart.cloudera:8020/user/training/outputs/output1.csv/part-000000', 'INFO scheduler.DAGScheduler: ResultStage 8 (saveAsTextFile at NativeMethodAccessorImpl.java:2) finished in 0.402 s', 'INFO scheduler.DAGScheduler: Job 4 finished: saveAsTextFile at NativeMethodAccessorImpl.java:2, took 0.463774 s', 'INFO scheduler.TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool', 'INFO mapred.FileInputFormat: Total input paths to process : 1', 'INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:393', 'INFO scheduler.DAGScheduler: Got job 5 (runJob at PythonRDD.scala:393) with 1 output partitions', 'INFO scheduler.DAGScheduler: Final stage: ResultStage 9 (runJob at PythonRDD.scala:393)', 'INFO scheduler.DAGScheduler: Parents of Final stage: List()', 'INFO scheduler.DAGScheduler: Missing parents: List()', 'INFO scheduler.DAGScheduler: Submitting ResultStage 9 (PythonRDD[19] at RDD at PythonRDD.scala:43), which has no missing parents', 'INFO storage.MemoryStore: Block broadcast 9 stored as values in memory (estimated size 4.8 KB, free 427.3 KB)', 'INFO storage.BlockManagerInfo: Added broadcast 9 piece0 in memory on localhost:38681 (size: 3.0 KB, free: 534.4 MB)', 'INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1066', 'INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (PythonRDD[19] at RDD at PythonRDD.scala:43)', 'INFO scheduler.TaskSchedulerImpl: Adding task set 9.0 with 1 tasks', 'INFO scheduler.TaskSchedulerImpl: Starting task 0.0 in stage 9.0 (TID 7), localhost, partition 0,ANY, 2178 bytes)', 'INFO executor.Executor: Running task 0.0 in stage 9.0 (TID 7)', 'INFO rdd.HadoopRDD: Input split: hdfs://quickstart.cloudera:8020/user/training/outputs/output1.csv/part-000000:0+2397', 'INFO python.PythonRunner: Times: total = 45, boot = 27817, init = 27862, finish = 0', 'INFO executor.Executor: Finished task 0.0 in stage 9.0 (TID 7): 2201 bytes result sent to driver', 'INFO scheduler.DAGScheduler: ResultStage 9 (runJob at PythonRDD.scala:393) finished in 0.063 s', 'INFO scheduler.DAGScheduler: Job 5 finished: runJob at PythonRDD.scala:393, took 0.110857 s', 'INFO scheduler.TaskSchedulerImpl: Finished task 0.0 in stage 9.0 (TID 7) in 66 ms on localhost (1/1)', 'INFO scheduler.TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool', and finally 'INFO East Timor,27', 'Canada,19', 'Sao Tome and Principe,24', 'Turkmenistan,31', 'United States of America,38'.



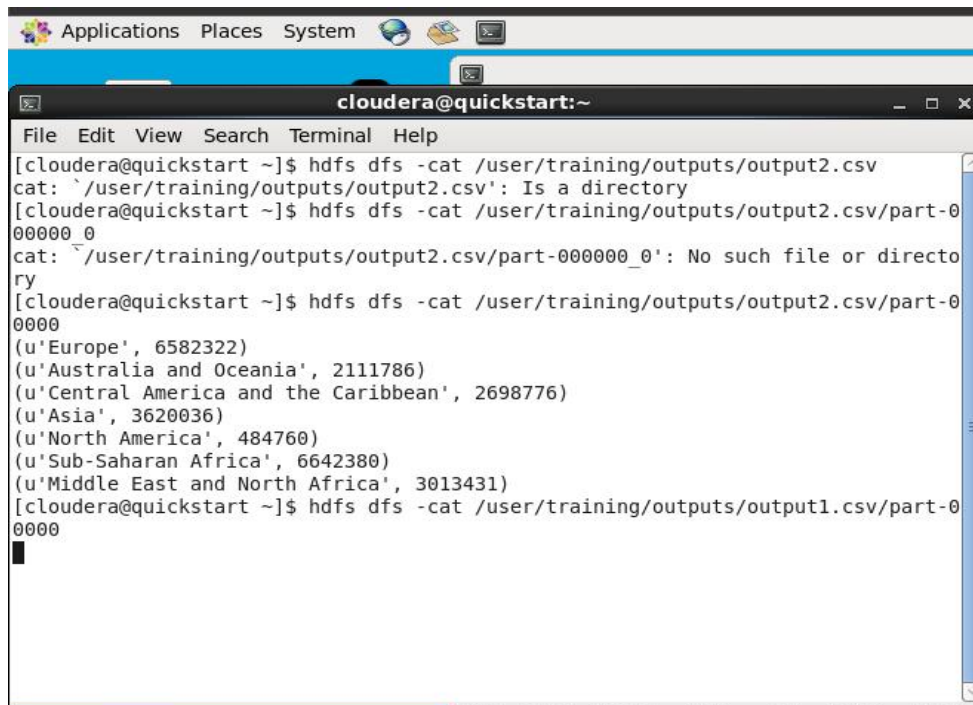
The screenshot shows a terminal window titled 'cloudera@quickstart:~'. It displays the output of the command 'hdfs dfs -cat /user/training/outputs/output1.csv/part-000000'. The output lists countries and their counts, separated by commas: 'East Timor,27', 'Canada,19', 'Sao Tome and Principe,24', 'Turkmenistan,31', 'United States of America,38', 'Lithuania,25', 'Cambodia,26', 'Ethiopia,26', 'The Gambia,19', 'Swaziland,25', 'Cameroon,31', 'Burkina Faso,26', 'Saint Kitts and Nevis ,32', 'Ghana,38', 'Saudi Arabia,23', 'Cape Verde,14', 'Slovenia,27', 'Guatemala,28', 'Bosnia and Herzegovina,33', 'Guinea,33', 'Jordan,31', and 'Dominica,18'.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
' at line 1  
mysql> select * from sp_solution1;  
+-----+-----+  
| country | count |  
+-----+-----+  
| Spain   | 26    |  
| Burundi | 26    |  
| Taiwan  | 31    |  
| Fiji    | 28    |  
| Barbados | 26    |  
| Madagascar | 24    |  
| Palau   | 25    |  
| Bhutan  | 30    |  
| Sudan   | 34    |  
| Nepal   | 28    |  
| Malta   | 32    |  
| Democratic Republic of the Congo | 23    |  
| Maldives | 20    |  
| United Kingdom | 23    |  
| Israel  | 23    |  
| Tunisia | 31    |  
| Iceland | 23    |  
| Zambia  | 31    |  
| Senegal | 30    |  
+-----+-----+
```

2] Display the number of units sold in each region.

```
rdd1=sc.textFile("/user/training/sales.csv")  
header=rdd1.first()  
rdd2=rdd1.filter(lambda row:row != header)  
rdd3=rdd2.map(lambda x:x.split(','))  
rdd4=rdd3.map(lambda x:(x[0],x[8])).reduceByKey(lambda x,y:x+y)  
  
rdd4.saveAsTextFile('/user/training/outputs/output2.csv')
```

```
File Edit View Search Terminal Help  
23/03/20 05:37:17 INFO python.PythonRunner: Times: total = 19, boot  
= -152, init = 168, finish = 3  
23/03/20 05:37:17 INFO executor.Executor: Finished task 0.0 in stag  
e 8.0 (TID 8). 1496 bytes result sent to driver  
23/03/20 05:37:17 INFO scheduler.DAGScheduler: ResultStage 8 (runJo  
b at PythonRDD.scala:393) finished in 0.085 s  
23/03/20 05:37:17 INFO scheduler.DAGScheduler: Job 5 finished: runJ  
ob at PythonRDD.scala:393, took 1.759492 s  
23/03/20 05:37:17 INFO scheduler.TaskSetManager: Finished task 0.0  
in stage 8.0 (TID 8) in 97 ms on localhost (1/1)  
23/03/20 05:37:17 INFO scheduler.TaskSchedulerImpl: Removed TaskSet  
8.0, whose tasks have all completed, from pool  
[(u'Europe', 6582322), (u'Australia and Oceania', 2111786), (u'Cent  
ral America and the Caribbean', 2698776), (u'Asia', 3620036), (u'No  
rth America', 484760), (u'Sub-Saharan Africa', 6642380), (u'Middle  
East and North Africa', 3013431)]  
>>> █
```



A terminal window titled 'cloudera@quickstart:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

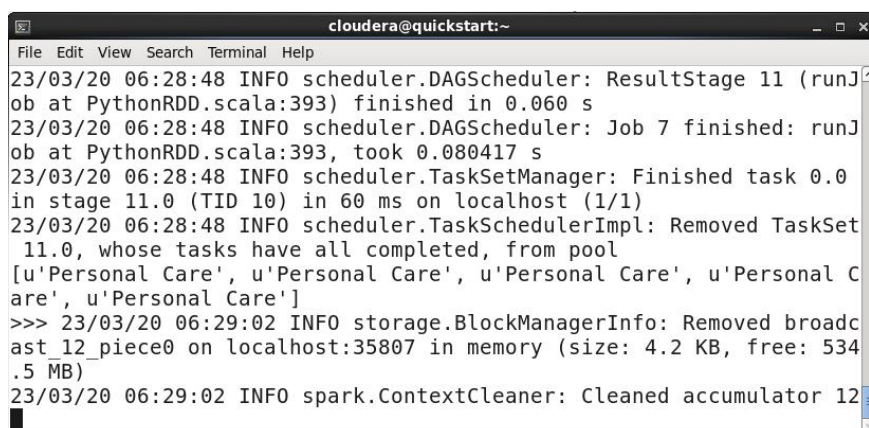
```
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output2.csv
cat: `/user/training/outputs/output2.csv': Is a directory
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output2.csv/part-000000
cat: `/user/training/outputs/output2.csv/part-000000_0': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output2.csv/part-000000
(u'Europe', 6582322)
(u'Australia and Oceania', 2111786)
(u'Central America and the Caribbean', 2698776)
(u'Asia', 3620036)
(u'North America', 484760)
(u'Sub-Saharan Africa', 6642380)
(u'Middle East and North Africa', 3013431)
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output1.csv/part-000000
```

3] Display the 10 most recent sales.

4] Display the products with atleast two occurrences of 'a'

```
rdd1=sc.textFile("/user/training/sales.csv")
header =rdd1.first()
rdd2 = rdd1.filter(lambda row:row != header)
rdd3=rdd2.map(lambda x:x.split(','))
rdd4=rdd3.map(lambda x:x[2]).filter(lambda x:x.count('a')>=2)

rdd4.coalesce(1).saveAsTextFile('/user/training/outputs/output4.csv')
```



A terminal window titled 'cloudera@quickstart:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows Spark logs:

```
23/03/20 06:28:48 INFO scheduler.DAGScheduler: ResultStage 11 (runJob at PythonRDD.scala:393) finished in 0.060 s
23/03/20 06:28:48 INFO scheduler.DAGScheduler: Job 7 finished: runJob at PythonRDD.scala:393, took 0.080417 s
23/03/20 06:28:48 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 11.0 (TID 10) in 60 ms on localhost (1/1)
23/03/20 06:28:48 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 11.0, whose tasks have all completed, from pool
[u'Personal Care', u'Personal Care', u'Personal Care', u'Personal Care', u'Personal Care']
>>> 23/03/20 06:29:02 INFO storage.BlockManagerInfo: Removed broadcast 12_piece0 on localhost:35807 in memory (size: 4.2 KB, free: 534.5 MB)
23/03/20 06:29:02 INFO spark.ContextCleaner: Cleaned accumulator 12
```

```

cat: `/user/training/outputs/output4.csv/part-000000': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output3.csv/part-000000
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care
Personal Care

```

5] Display country in each region with highest units sold

```

rdd1=sc.textFile('/FileStore/tables/5000_Sales_Records.csv')
rdd2=rdd1.map(lambda x: x.split(',')).map(lambda x:(x[0],[x[1],x[8]])).reduceByKey(lambda v1,v2: (v1
if v1>=v2 else v2))
def toCSVLine(data):
    return ','.join(str(d) for d in data)
rdd3=rdd2.map(toCSVLine) rdd3.take(5)

rdd3.coalesce(1).saveAsTextFile('/user/training/outputs/output5.csv')

```

```

[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output5.csv/part-000000
Europe,[u'Vatican City', u'9381']
Australia and Oceania,[u'Vanuatu', u'959']
Region,[u'Country', u'Units Sold']
Central America and the Caribbean,[u'Trinidad and Tobago', u'958']
Asia,[u'Vietnam', u'928']
North America,[u'United States of America', u'876']
Sub-Saharan Africa,[u'Zimbabwe', u'969']
Middle East and North Africa,[u'Yemen', u'9864']
[cloudera@quickstart ~]$

```

6] Display the unit price and unit cost of each item in ascending order

```

rdd1=sc.textFile("/user/training/sales.csv")
header =rdd1.first()
rdd2 = rdd1.filter(lambda row:row != header)
rdd3=rdd2.map(lambda x:x.split(','))
rdd4=rdd3.map(lambda x:(x[2],x[9],x[10])).distinct().sortBy(lambda x:x[2])

rdd4.saveAsTextFile('/home/training/outputs/output6.csv')

```

```

File Edit View Search Terminal Help
23/03/20 06:31:34 INFO executor.Executor: Finished task 0.0 in stage 13.0 (TID 12). 1715 bytes result sent to driver
23/03/20 06:31:34 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 13.0 (TID 12) in 67 ms on localhost (1/1)
23/03/20 06:31:34 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 13.0, whose tasks have all completed, from pool
23/03/20 06:31:34 INFO scheduler.DAGScheduler: ResultStage 13 (runJob at PythonRDD.scala:393) finished in 0.056 s
23/03/20 06:31:34 INFO scheduler.DAGScheduler: Job 8 finished: runJob at PythonRDD.scala:393, took 0.310541 s
[(u'Fruits', 9.330000000000001, 6.919999999999999), (u'Beverages', 47.450000000000003, 31.789999999999999), (u'Clothes', 109.28, 35.840000000000003), (u'Personal Care', 81.730000000000004, 56.670000000000002), (u'Vegetables', 154.06, 90.930000000000007), (u'Snacks', 152.58000000000001, 97.439999999999998), (u'Cereal', 205.69999999999999, 117.11), (u'Baby Food', 255.28, 159.41999999999999), (u'Cosmetics', 437.19999999999999, 263.32999999999998), (u'Meat', 421.88999999999999, 364.69), (u'Household', 668.26999999999998, 502.54000000000002), (u'Office Supplies', 651.21000000000004, 524.96000000000004)]
>>>

```



```

File Edit View Search Terminal Help
0000
Europe,[u'Vatican City', u'9381']
Australia and Oceania,[u'Vanuatu', u'959']
Region,[u'Country', u'Units Sold']
Central America and the Caribbean,[u'Trinidad and Tobago', u'958']
Asia,[u'Vietnam', u'928']
North America,[u'United States of America', u'876']
Sub-Saharan Africa,[u'Zimbabwe', u'969']
Middle East and North Africa,[u'Yemen', u'9864']
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output6.csv/part-0
0000
(u'Fruits', 9.330000000000001, 6.919999999999999)
(u'Beverages', 47.450000000000003, 31.789999999999999)
(u'Clothes', 109.28, 35.840000000000003)
(u'Personal Care', 81.730000000000004, 56.670000000000002)
(u'Vegetables', 154.06, 90.930000000000007)
(u'Snacks', 152.58000000000001, 97.439999999999998)
(u'Cereal', 205.69999999999999, 117.11)
(u'Baby Food', 255.28, 159.41999999999999)
(u'Cosmetics', 437.19999999999999, 263.32999999999998)
(u'Meat', 421.88999999999999, 364.69)
(u'Household', 668.26999999999998, 502.540000000000002)
(u'Office Supplies', 651.21000000000004, 524.96000000000004)
[cloudera@quickstart ~]$

```

7] Display the number of sales yearwise

```

def frmt_dt(dt):
    updt=dt.split('/')[2]+dt.split('/')[0]+dt.split('/')[1]
    return int(updt)
rdd1=sc.textFile("/user/training/sales.csv")
header =rdd1.first()
rdd2 = rdd1.filter(lambda row:row != header)
rdd3=rdd2.map(lambda
x:(x.split(',')[0],x.split(',')[1],x.split(',')[2],x.split(',')[3],x.split(',')[4],frmt_dt(x.split(',')[5]),x.split(',')[6],fr
mt_dt(x.split(',')[7]),int(x.split(',')[8]),float(x.split(',')[9]),float(x.split(',')[10]),float(x.split(',')[11]),float(x
.split(',')[12]),float(x.split(',')[13])))

rdd4=rdd3.map(lambda x:(str(x[5]):4,x[6])).groupBy(lambda x:x[0]).map(lambda x:(x[0],len(x[1])))

rdd4.saveAsTextFile('/user/training/outputs/output7.csv')

```

```

File Edit View Search Terminal Help
23/03/20 06:35:25 INFO executor.Executor: Running task 0.0 in stage 21
.0 (TID 18)
23/03/20 06:35:25 INFO storage.ShuffleBlockFetcherIterator: Getting 1
non-empty blocks out of 1 blocks
23/03/20 06:35:25 INFO storage.ShuffleBlockFetcherIterator: Started 0
remote fetches in 0 ms
23/03/20 06:35:25 INFO python.PythonRunner: Times: total = 15, boot =
-27432, init = 27437, finish = 10
23/03/20 06:35:25 INFO executor.Executor: Finished task 0.0 in stage 2
1.0 (TID 18). 1406 bytes result sent to driver
23/03/20 06:35:25 INFO scheduler.TaskSetManager: Finished task 0.0 in
stage 21.0 (TID 18) in 26 ms on localhost (1/1)
23/03/20 06:35:25 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 21
.0, whose tasks have all completed, from pool
23/03/20 06:35:25 INFO scheduler.DAGScheduler: ResultStage 21 (runJob
at PythonRDD.scala:393) finished in 0.024 s
23/03/20 06:35:25 INFO scheduler.DAGScheduler: Job 13 finished: runJob
at PythonRDD.scala:393, took 0.189738 s
[('2015', 679), ('2014', 660), ('2017', 363), ('2016', 670), ('2011',
658), ('2010', 632), ('2013', 660), ('2012', 678)]
>>>

```

```
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output7.csv/part-000000
('2015', 679)
('2014', 660)
('2017', 363)
('2016', 670)
('2011', 658)
('2010', 632)
('2013', 660)
('2012', 678)
[cloudera@quickstart ~]$
```

8] Display the number of orders for each item

```
rdd1=sc.textFile("/user/training/sales.csv")
header=rdd1.first()
rdd2=rdd1.filter(lambda row:row != header)
rdd3=rdd2.map(lambda x:x.split(','))
rdd4=rdd3.map(lambda x:(x[2],x[6])).groupByKey().map(lambda x:(x[0],len(x[1])))
Rdd4.saveAsTextFile('/user/training/outputs/output8.csv')
```

```
File Edit View Search Terminal Help
23/03/20 06:37:58 INFO storage.ShuffleBlockFetcherIterator: Getting 1
non-empty blocks out of 1 blocks
23/03/20 06:37:58 INFO storage.ShuffleBlockFetcherIterator: Started 0
remote fetches in 0 ms
23/03/20 06:37:58 INFO python.PythonRunner: Times: total = 8, boot = -
53, init = 58, finish = 3
23/03/20 06:37:58 INFO executor.Executor: Finished task 0.0 in stage 2
5.0 (TID 21). 1549 bytes result sent to driver
23/03/20 06:37:58 INFO scheduler.TaskSetManager: Finished task 0.0 in
stage 25.0 (TID 21) in 19 ms on localhost (1/1)
23/03/20 06:37:58 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 25
.0, whose tasks have all completed, from pool
23/03/20 06:37:58 INFO scheduler.DAGScheduler: ResultStage 25 (runJob
at PythonRDD.scala:393) finished in 0.021 s
23/03/20 06:37:58 INFO scheduler.DAGScheduler: Job 15 finished: runJob
at PythonRDD.scala:393, took 0.244855 s
[(u'Personal Care', 415), (u'Snacks', 398), (u'Baby Food', 445), (u'Ve
getables', 410), (u'Beverages', 447), (u'Cosmetics', 424), (u'Cereal',
385), (u'Fruits', 447), (u'Clothes', 386), (u'Household', 424), (u'Of
fice Supplies', 420), (u'Meat', 399)]
>>>
```

```
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -cat /user/training/outputs/output8.csv/part-000000
(u'Personal Care', 415)
(u'Snacks', 398)
(u'Baby Food', 445)
(u'Vegetables', 410)
(u'Beverages', 447)
(u'Cosmetics', 424)
(u'Cereal', 385)
(u'Fruits', 447)
(u'Clothes', 386)
(u'Household', 424)
(u'Office Supplies', 420)
(u'Meat', 399)
[cloudera@quickstart ~]$
```

9] Display the country with highest sale