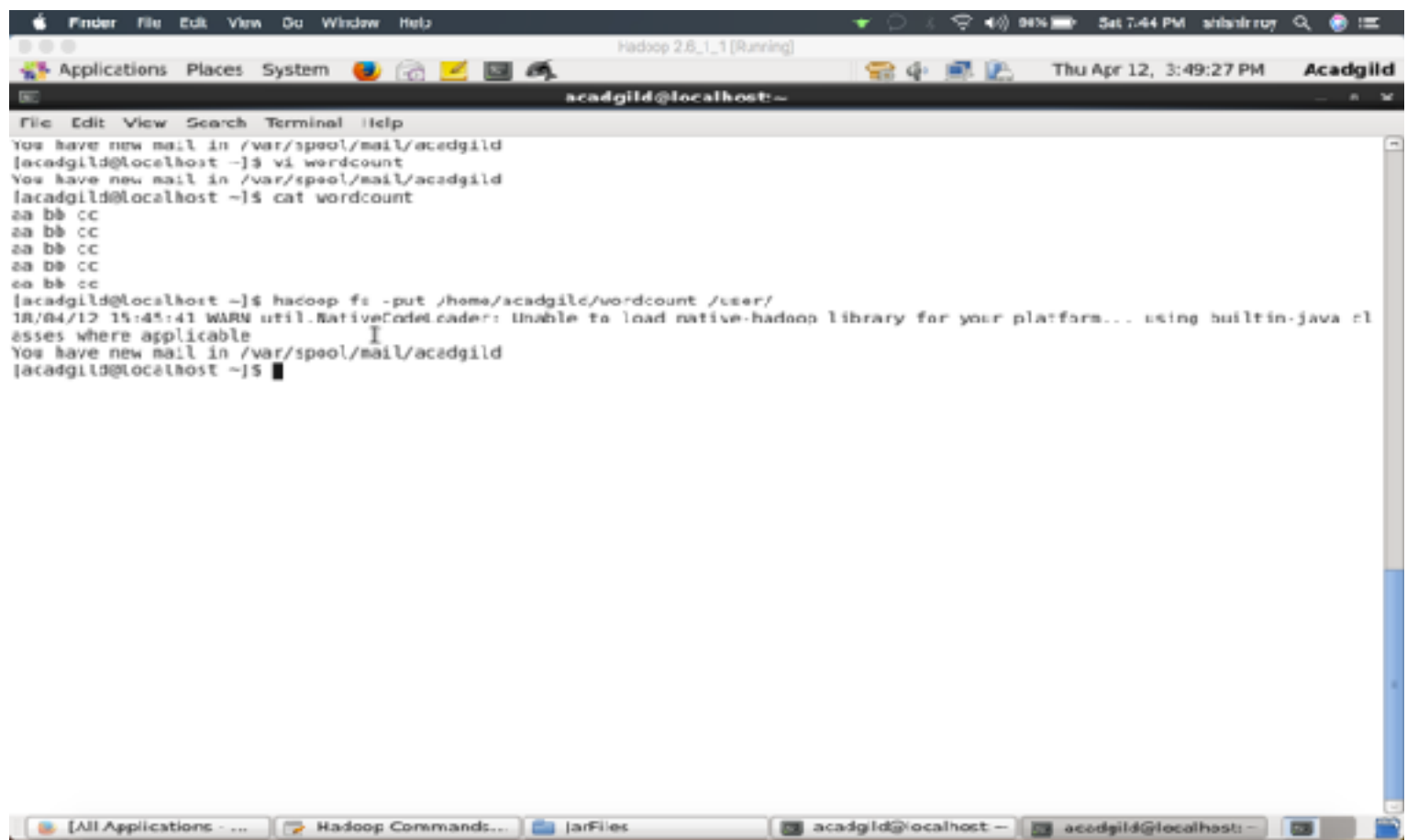
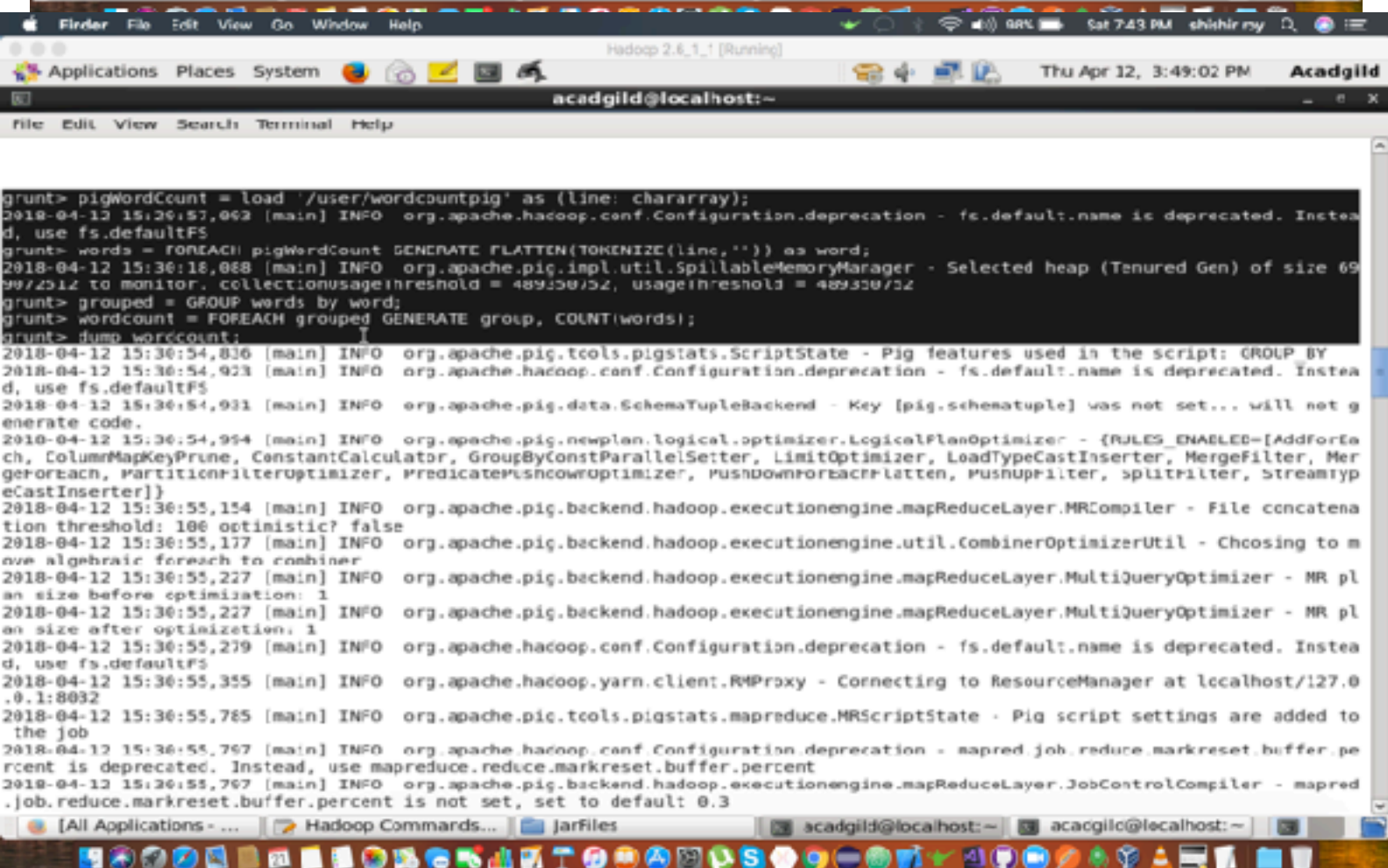


Task 1

Write a program to implement wordcount using Pig.



```
Finder File Edit View Go Window Help
hadoop 2.6.1_1 [Running]
Applications Places System
acacgild@localhost:~
File Edit View Search Terminal Help
You have new mail in /var/spool/mail/acacgild
[acacgild@localhost ~]$ vi wordcount
You have new mail in /var/spool/mail/acacgild
[acacgild@localhost ~]$ cat wordcount
aa bb cc
aa bb cc
aa bb cc
aa bb cc
aa bb cc
[acacgild@localhost ~]$ hadoop fs -put /home/acacgild/wordcount /user/
18/04/12 15:45:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
You have new mail in /var/spool/mail/acacgild
[acacgild@localhost ~]$
```



```
Finder File Edit View Go Window Help
hadoop 2.6.1_1 [Running]
Applications Places System
acacgild@localhost:~
File Edit View Search Terminal Help
grunt> pigWordCount = load '/user/wordcount/pig' as (line: chararray);
2018-04-12 15:36:57,962 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> words = FOREACH pigWordCount GENERATE FLATTEN(TOKENIZE(line, '')) as word;
2018-04-12 15:36:18,068 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 69
9972512 to monitor. collectionUsageThreshold = 48933832, usageThreshold = 48933832
grunt> grouped = GROUP words by word;
grunt> wordcount = FOREACH grouped GENERATE group, COUNT(words);
grunt> dump wordcount;
2018-04-12 15:36:54,836 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2018-04-12 15:36:54,923 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-12 15:36:54,931 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2018-04-12 15:36:54,994 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Mer
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTyp
eCastInserter]}
2018-04-12 15:36:55,154 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2018-04-12 15:36:55,177 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to m
ove algebraic foreach to combiner
2018-04-12 15:36:55,227 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2018-04-12 15:36:55,227 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2018-04-12 15:36:55,279 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-04-12 15:36:55,355 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0
.0.1:8032
2018-04-12 15:36:55,785 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to
the job
2018-04-12 15:36:55,797 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.p
ercent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2018-04-12 15:36:55,797 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default: 0.3
```

```
Input(s):
Successfully read 5 records (408 bytes) from: "/user/wordcount"

Output(s):
Successfully stored 1 records (17 bytes) in: "hdfs://localhost:8020/tmp/temp2063247015/tmp-975112111"

Counters:
Total records written : 1
Total bytes written : 17
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1521868459622_0034

2018-04-12 15:47:44,420 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-12 15:47:44,438 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-12 15:47:44,537 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-12 15:47:44,546 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-12 15:47:44,664 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-12 15:47:44,675 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-12 15:47:44,752 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2018-04-12 15:47:44,798 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-12 15:47:44,798 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-04-12 15:47:44,816 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-12 15:47:44,817 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(aa bb cc,5)
```

Task 2

We have employee_details and employee_expenses files. Use local mode while running Pig and write Pig Latin script to get below results:

employee_details (EmpID,Name,Salary,DepartmentID) https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_details.txt

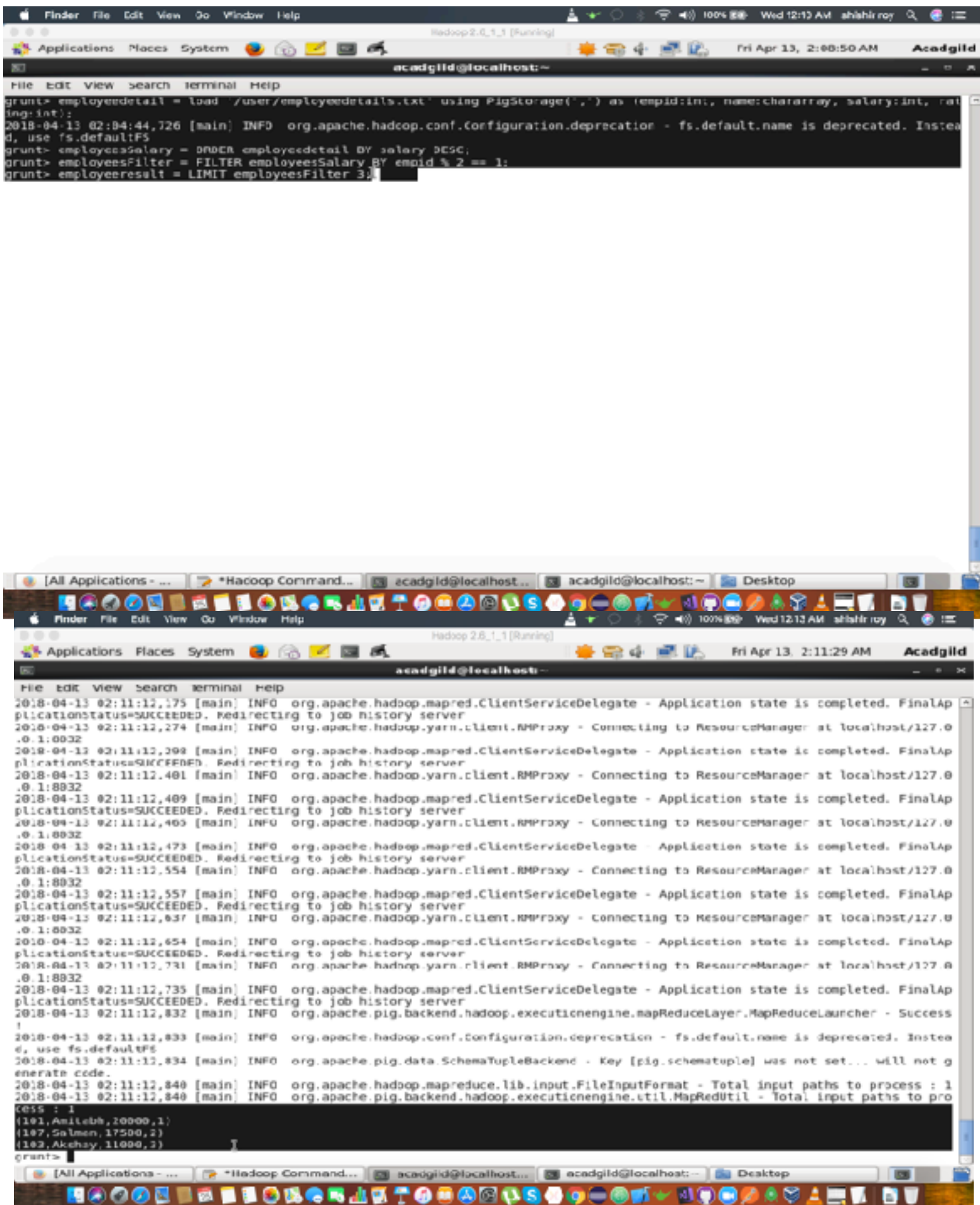
employee_expenses(EmpID,Expense) https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_expenses.txt

(a) Top 5 employees (employee id and employee name) with highest rating.

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.8.1_1 [Running]
Applications  Places  System
acadgild@localhost:~
File Edit View Search Terminal Help
grunt> employeeDetail = load '/user/employeeDetails.txt' using PigStorage(',') as (empid:int, name:chararray, salary:int, rating:int);
2018-04-13 01:59:41,547 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> employeeOrder = ORDER employeeDetail BY rating DESC;
grunt> employeeResult = LIMIT employeeOrder 5;
grunt> dump employeeResult;
```

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.8.1_1 [Running]
Applications  Places  System
acadgild@localhost:~
File Edit View Search Terminal Help
2018-04-13 02:02:55,214 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 02:02:55,229 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 02:02:55,337 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 02:02:55,343 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 02:02:55,457 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 02:02:55,475 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 02:02:55,619 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 02:02:55,647 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 02:02:55,797 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 02:02:55,893 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 02:02:55,993 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 02:02:55,939 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 02:02:56,058 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2018-04-13 02:02:56,062 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-13 02:02:56,064 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-04-13 02:02:56,072 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-13 02:02:56,072 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka,2000,5)
(105,Ravan,2500,5)
(100,Katrina,1000,4)
(104,Anubhav,5000,4)
(108,Rashir,14000,3)
grunt>
```


(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number.



The image shows two screenshots of a Mac terminal window. The top screenshot shows the execution of Pig commands to load employee data, sort it by salary, filter for odd employee IDs, and limit the results to 3. The bottom screenshot shows the output of the Pig job, including status messages and the final result of the query.

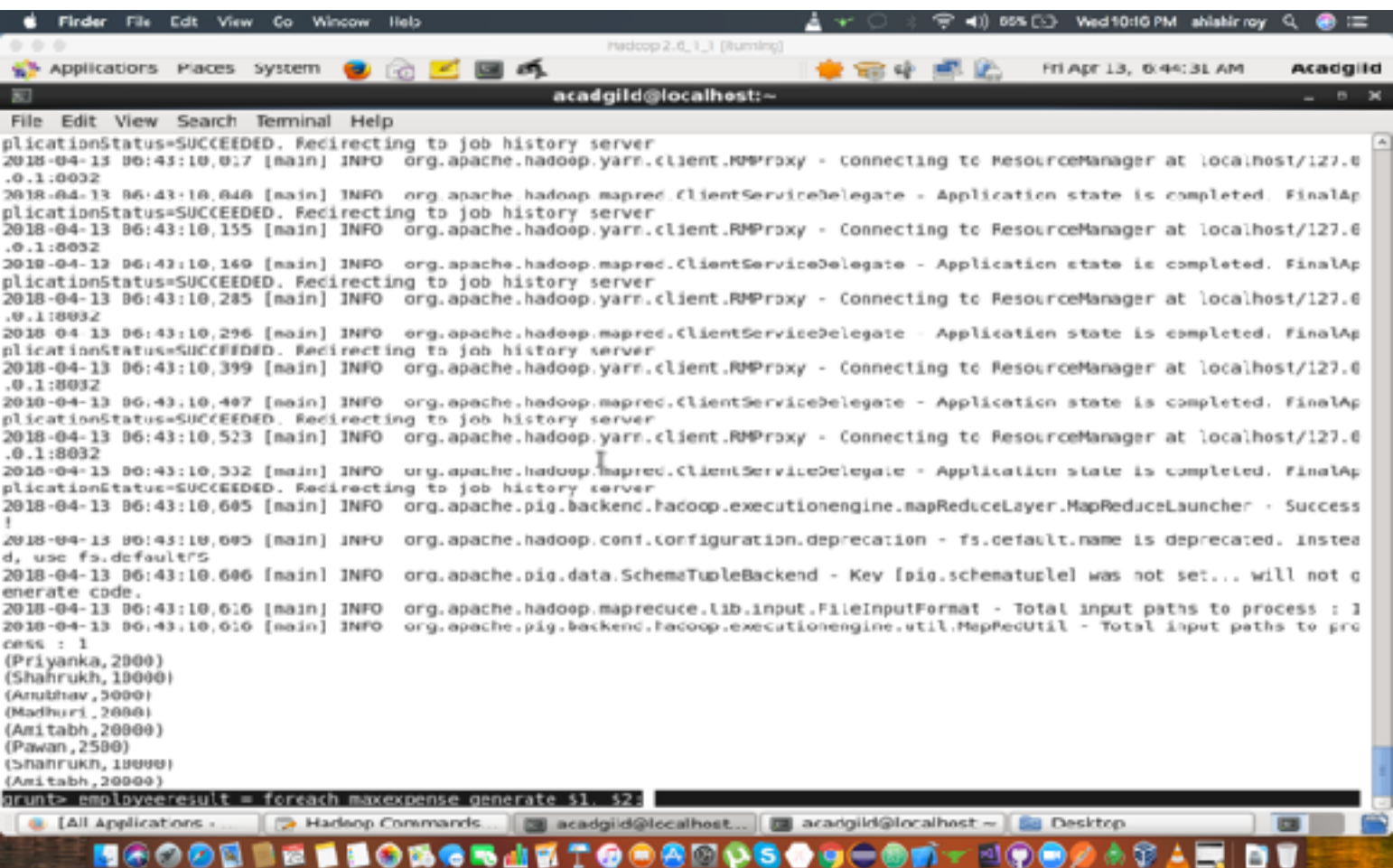
```
acadmild@localhost:~  
grunt> employeeDetail = load '/user/employeeDetails.txt' using PigStorage(',') as (empid:int, name:chararray, salary:int, rating:int);  
2018-04-13 02:04:44,726 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
grunt> employeesSalary = ORDER employeeDetail BY salary DESC;  
grunt> employeesFilter = FILTER employeesSalary BY empid % 2 == 1;  
grunt> employeesResult = LIMIT employeesFilter 3;
```

```
2018-04-13 02:11:12,175 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,274 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032  
2018-04-13 02:11:12,398 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,401 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032  
2018-04-13 02:11:12,409 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,409 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032  
2018-04-13 02:11:12,473 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,554 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032  
2018-04-13 02:11:12,557 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,637 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032  
2018-04-13 02:11:12,654 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,731 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032  
2018-04-13 02:11:12,735 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server  
2018-04-13 02:11:12,832 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success  
1  
2018-04-13 02:11:12,833 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2018-04-13 02:11:12,834 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.  
2018-04-13 02:11:12,840 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1  
2018-04-13 02:11:12,840 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1  
(101,Amitabh,20000,1)  
(107,Salman,17500,2)  
(102,Akshay,11000,2)  
grunt>
```

(c) Employee (employee id and employee name) with maximum expense



```
Finder File Edit View Go Window Help
Hadoop 2.6.1_1 [Running]
Applications Places System
acadmild@localhost:~
File Edit View Search Terminal Help
grunt> employeeDetail = load '/user/employeeDetails.txt' using PigStorage(',') as (empid:int, name:chararray, salary:int, rating:int);
2018-04-13 06:39:14,761 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> employeeExpense = load '/user/employeeExpenses.txt' using PigStorage(',') as (empid:int, expense:int);
2018-04-13 06:39:20,291 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> employeeData = JOIN employeeDetail BY empid, employeeExpense BY empid;
grunt> maxExpense = GROUP employeeData BY expense DESC;
```



```
Finder File Edit View Go Window Help
Hadoop 2.6.1_1 [Running]
Applications Places System
acadmild@localhost:~
File Edit View Search Terminal Help
applicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 06:43:10,017 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 06:43:10,048 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 06:43:10,155 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 06:43:10,169 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 06:43:10,285 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 06:43:10,296 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 06:43:10,399 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 06:43:10,407 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 06:43:10,523 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-13 06:43:10,532 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-13 06:43:10,605 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-13 06:43:10,605 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-13 06:43:10,606 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-04-13 06:43:10,616 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-13 06:43:10,616 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
(Priyanka,2000)
(Shahrukh,10000)
(Aruthav,5000)
(Madhuri,2000)
(Amitabh,20000)
(Pawan,2500)
(Shahrukh,10000)
(Amitabh,20000)
grunt> employeeResult = foreach maxExpense generate $1, $2;
```

(d) List of employees (employee id and employee name) having entries in employee_expenses file.

```

2018-04-14 03:32:53,696 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-14 03:32:53,718 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost:127.0
.0.1:8032
2018-04-14 03:32:54,722 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-14 03:32:54,848 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost:127.0
.0.1:8032
2018-04-14 03:32:54,844 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-14 03:32:54,943 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost:127.0
.0.1:8032
2018-04-14 03:32:53,940 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-14 03:32:54,012 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost:127.0
.0.1:8032
2018-04-14 03:32:54,020 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-14 03:32:54,110 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost:127.0
.0.1:8032
2018-04-14 03:32:54,120 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-04-14 03:32:54,101 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success
ful
2018-04-14 03:32:54,193 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-04-14 03:32:54,192 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2018-04-14 03:32:54,202 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-14 03:32:54,202 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to pro
cess : 1
[101, AnilKumar]
[102, Shanmugam]
[104, Anubhav]
[105, Pawan]
[110, Priyanka]
[114, Madhura]
grep result = DISTINCT detail:You have new mail in /var/spool/mail/acadgild
acadgild@localhost ~$

```

(e) List of employees (employee id and employee name) having no entry in employee_expenses file.

```

grep employeeDetail = load '/user/employeeDetails.txt' using PigStorage(',') as (empid:int, name:chararray, salary:int, rat
ing:int);
2018-04-14 03:37:34,096 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grep employeeExpense = load '/user/employeeexpenses.txt' using PigStorage(',') as (empid:int, expense:int);
2018-04-14 03:37:34,096 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grep employeeNoExpense = join employeeDetail by empid with employeeExpense by empid;
2018-04-14 03:37:34,096 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (reserved 4096) of size 40
96 is 1/2 of monitor. collectionUsageThreshold = 0.5/0.5, usageThreshold = 0.5/0.5
grep empfilter = filter employeeNoExpense by employeeExpense::empid is null;
grep empdata = FOREACH empfilter GENERATE employeeDetail::empid, employeeDetail::name;
DUMP;

```



```
acacgild@localhost:~$
Total bytes written : 118
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job ID:
Job ID: hdfs://localhost:9000/job_1572884545677_000000

2018-04-18 03:45:45,888 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-18 03:45:45,906 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED, Redirecting to job history server
2018-04-18 03:45:46,125 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-18 03:45:46,147 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED, Redirecting to job history server
2018-04-18 03:45:46,300 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-04-18 03:45:46,320 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED, Redirecting to job history server
2018-04-18 03:45:46,533 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success
!
2018-04-18 03:45:46,548 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-18 03:45:46,552 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-04-18 03:45:46,674 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-18 03:45:46,676 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1

101.Akshay
106.Ashley
107.Salman
108.Ramair
109.Kulrind
111.Tushar
112.Ajay
113.Siddhant
ls -l
ls -l
```

Task 3

Implement the use case present in below blog link and share the complete steps along with screenshot(s) from your end.

<https://acadgild.com/blog/aviation-data-analysis-using-apache-pig/>

Problem 1

```
acacgild@localhost:~$
create REGISTER /home/acacgild/install/pig/pig-0.16.0/lib/piggybank.jar;
create A = load '/home/acacgild/Desktop/Hadoop Data/airline usecase pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage[, ',', 'NO MULTILINE', 'UNIX', 'SKIP INPUT HEADER'];
2018-05-20 11:39:55,771 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-05-20 11:39:55,771 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
create B = foreach A generate (int) $1 as year, (int) $2 as flight_num, (chararray) $3 as origin, (chararray) $4 as dest;
create C = filter B by dest is not null;
create D = group C by dest;
create E = foreach D generate group, count(C.dest);
create F = order E by $1 DESC;
create Result = LIMIT F 5;
create A1 = load '/home/acacgild/Desktop/Hadoop Data/airline usecase pig/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage[, ',', 'NO MULTILINE', 'UNIX', 'SKIP INPUT HEADER'];
2018-05-20 11:39:55,972 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-05-20 11:39:55,972 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
create A2 = foreach A1 generate (chararray) $0 as dest, (chararray) $2 as city, (chararray) $4 as country;
create joined table = join Result by $0, A2 by dest;
create
```

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.6.1 [Running]
Applications  Places  System
acacgild@localhost:~$
File Edit View Search Terminal Help
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,291 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,292 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,311 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,312 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,312 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,331 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,332 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,345 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,350 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,364 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,369 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=JobTracker, sessionId= - already initialized
2018-05-20 11:41:15,407 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - Success
!
2018-05-20 11:41:15,447 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-20 11:41:15,454 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
2018-05-20 11:41:15,454 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-20 11:41:15,500 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-20 11:41:15,500 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(ATL,106000,ATL,Atlanta,USA)
(DEN,02000,DEN,Denver,USA)
(DFW,70000,DFW,Dallas-Fort Worth,USA)
(LAX,25000,LAX,Los Angeles,USA)
(ORD,100000,ORD,Chicago,USA)
gruncle
```

Problem 2

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.6.1 [Running]
Applications  Places  System
acacgild@localhost:~$
File Edit View Search Terminal Help
gruncle> REGISTER /home/acacgild/Desktop/pig/pig-0.16.0/lib/piggybank.jar
gruncle> A = load /home/acacgild/Desktop/Hadoop Data/airline_usecase.pig/DelayedFlights.csv USING org.apache.pig.piggybank.storage.CSVExcelStorage[, ',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER'];
2018-05-20 11:54:22,175 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-20 11:54:22,175 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
gruncle> A = foreach A generate (int) $ as month, (int) $ as flight_num, (int) $ as cancelled, (chararray) $ as cancel_code;
gruncle> C = filter A by cancelled == 1 and cancel_code == 'C';
gruncle> B = group C by month;
gruncle> A = foreach B generate group, count(C.cancelled);
gruncle> B = order A by $1 desc;
gruncle> Result = limit B 1;
gruncle>
```



```
2018-05-20 11:56:34,574 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,574 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,575 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,581 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,582 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,592 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,593 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,599 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,600 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,602 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,602 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,602 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e=jobTracker, sessionId= - already initialized
2018-05-20 11:56:34,603 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success
!
2018-05-20 11:56:34,604 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-20 11:56:34,604 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead
use fs.defaultFS
2018-05-20 11:56:34,604 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-20 11:56:34,640 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-20 11:56:34,640 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Total input paths to pro
cess : 1
11:56:36
grunt>
```

problem 3

```
grunt> REGISTER /home/ecadgild/install/pig/pig-0.16.0/lib/piggybank.jar
grunt> A = load '/home/ecadgild/Desktop/Hadoop Data/airline_usecase pig/DelayedFlights.csv' USING org.apache.pig.piggybank.st
orage.CSVExcelStorage[, 'NO MULTILINE', 'UNIX', 'SKIP INPUT HEADER'];
2018-05-20 12:02:44,374 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-20 12:02:44,374 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> B1 = foreach A generate (int)$1 as dep_delay, (chararray)$2 as origin;
grunt> C1 = filter B1 by (dep_delay is not null) and (origin is not null);
grunt> B1 = group C1 by origin;
grunt> B1 = foreach B1 generate group, min(C1.dep_delay);
grunt> TopTen = limit B1 Result 10;
grunt> Lookup = load '/home/ecadgild/Desktop/Hadoop Data/airline_usecase pig/airports.csv' USING org.apache.pig.piggybank.st
orage.CSVExcelStorage[, 'NO MULTILINE', 'UNIX', 'SKIP INPUT HEADER'];
2018-05-20 12:03:51,777 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-05-20 12:03:51,777 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
grunt> Joined = join Lookup1 by origin, TopTen by $0;
grunt> Final = foreach Joined generate $0,$1,$2,$4;
grunt> Final Result = DROPPED Final by $3 DESC;
```

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.6.1 [Shuang]
Applications  Places  System
acadgild@localhost:~$
File  Edit  View  Search  Terminal  Media
2018-05-20 12:01:07,822 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,842 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,844 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,845 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,846 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,847 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,848 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId - already initialized
2018-05-20 12:01:07,891 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-05-20 12:01:07,922 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use fs.bytes.per.checksum
2018-05-20 12:01:07,924 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
2018-05-20 12:01:07,926 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-20 12:01:08,022 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-20 12:01:08,022 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
[INFO,Hamock,USA,136,147050235264]
[INFO,Bellston,USA,03,78190476190476]
[INFO,Springfield,USA,88,84873949579811]
[INFO,Watertown,USA,82,2258064515129]
[INFO,MA,USA,70,55605474610547]
[INFO,Atlantic City,USA,79,310344227362]
[INFO,Miami,USA,78,06103413333633]
[INFO,MA,USA,70,33002404480074]
[INFO,Bogalusa,USA,74,12891000062710]
[INFO,Birmingham,USA,73,1333350032323]
[INFO]
[acadgild@localhost:~$
```

Problem 4

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.6.1 [Shuang]
Applications  Places  System
acadgild@localhost:~$
File  Edit  View  Search  Terminal  Media
grmr> REGISTER /home/acadgild/Desktop/Hadoop Data/airline usecase pig/DelayedFlights.csv USING org.apache.pig.piggybank.storage.CSVExcelStorage;',', 'NO MULTILINE', 'UNIX', 'SKIP INPUT HEADER';
2018-05-20 12:05:54,544 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use fs.bytes.per.checksum
2018-05-20 12:05:54,544 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
grmr> M = FOREACH A GENERATE (chararray)src as origin, (chararray)dest, (int)diversion;
grmr> C = FILTER M BY (origin is not null) AND (dest is not null) AND (diversion != 1);
grmr> D = ORDER C BY (origin,dest);
grmr> E = FOREACH D GENERATE GROUP, COUNT(C.diversion);
grmr> F = ORDER E BY $1 DESC;
grmr> RESULT = LIMIT F 100;
grmr>
```

```
Finder  File  Edit  View  Go  Window  Help
Hadoop 2.6.1_1 [Sharing]
Applications  Places  System
Sun May 20, 12:07 PM  AcadGillid
acagdillid@localhost:~$
File  Edit  View  Search  Terminal  Help
2018-05-20 12:07:21,499 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,502 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,504 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,505 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,505 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,505 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,505 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,505 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,508 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,509 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,509 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,509 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName
e-JobTracker, sessionId= - already initialized
2018-05-20 12:07:21,517 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success
1
2018-05-20 12:07:21,518 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes.per.checksum
2018-05-20 12:07:21,518 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
2018-05-20 12:07:21,518 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-20 12:07:21,527 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-20 12:07:21,527 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
1 (CMD, L6A), 39)
1 (DAL, H9U), 35)
1 (DFW, L6A), 33)
1 (ATL, L6A), 22)
1 (ORD, S8A), 31)
1 (SJC, S8H), 31)
1 (MIA, L6A), 31)
1 (BNA, LFK), 29)
1 (PHL, H9U), 28)
1 (BOS, CFN), 23)
graceL#
```