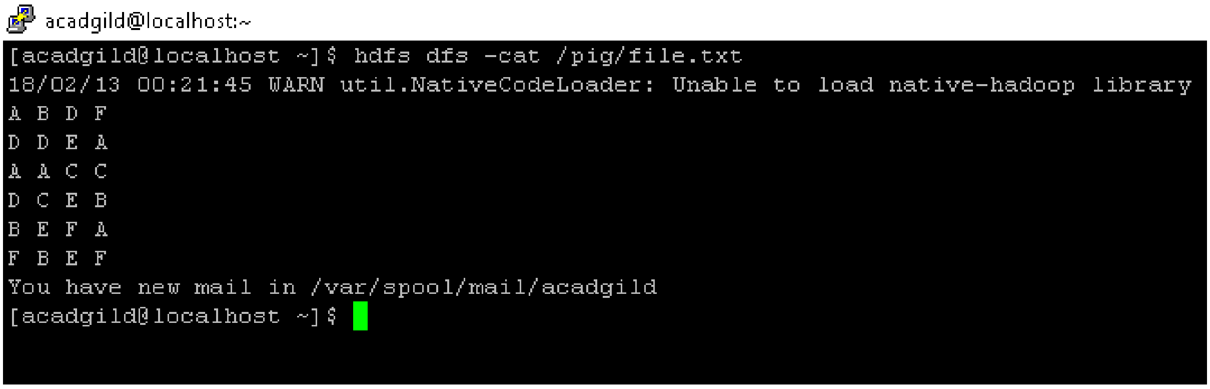
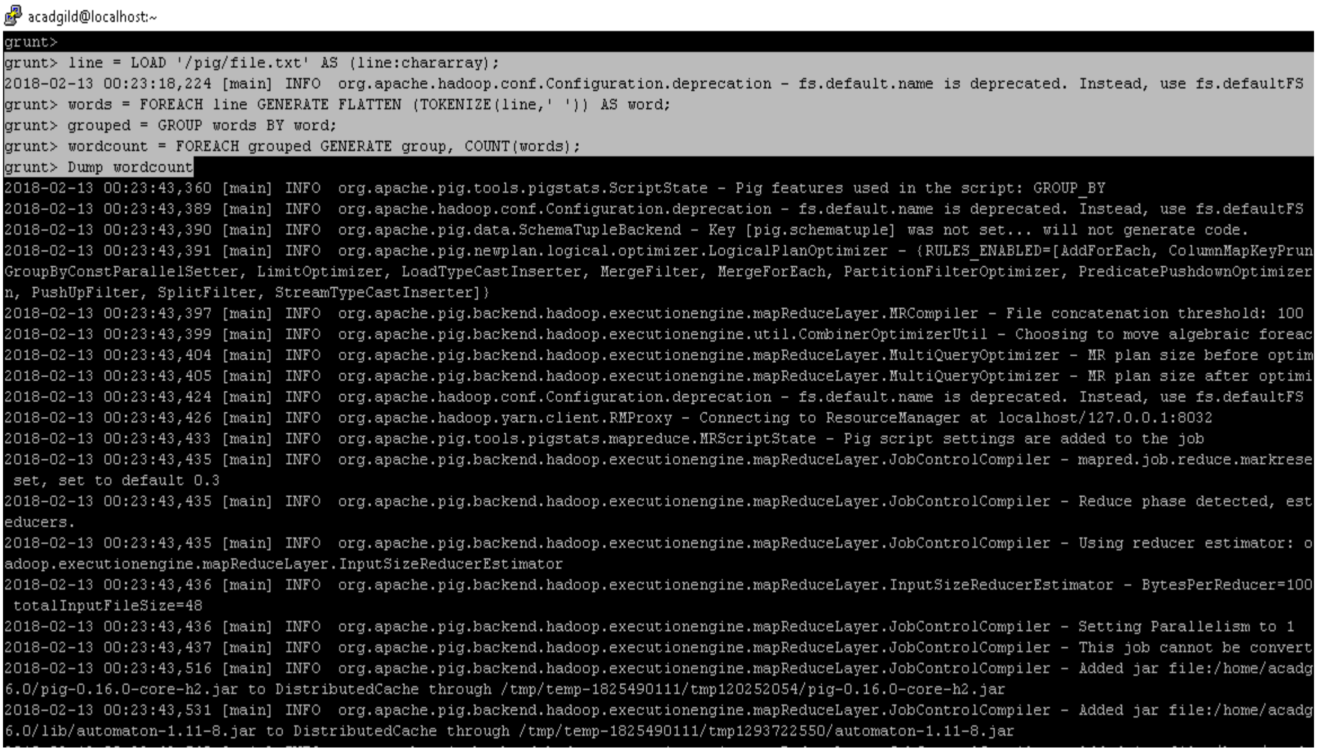
ASSIGNMENT 7.1

Task 1: Write a program to implement wordcount using Pig.

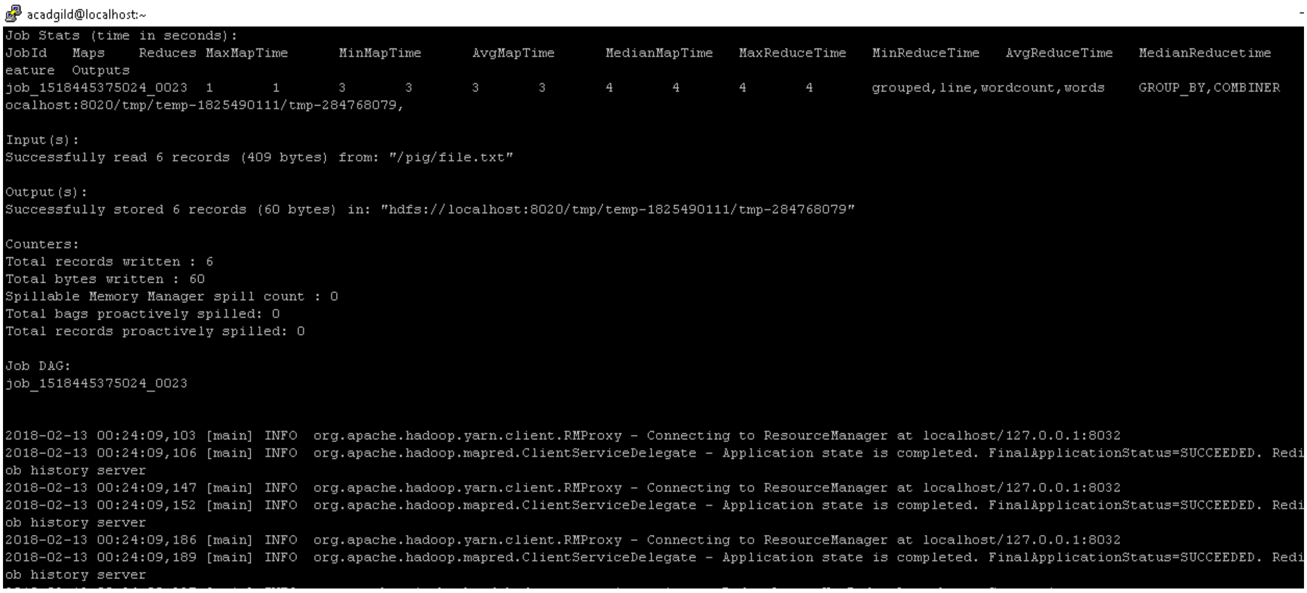
Input:-



```
[acadgild@localhost ~]$ hdfs dfs -cat /pig/file.txt
18/02/13 00:21:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library
A B D F
D D E A
A A C C
D C E B
B E F A
F B E F
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```



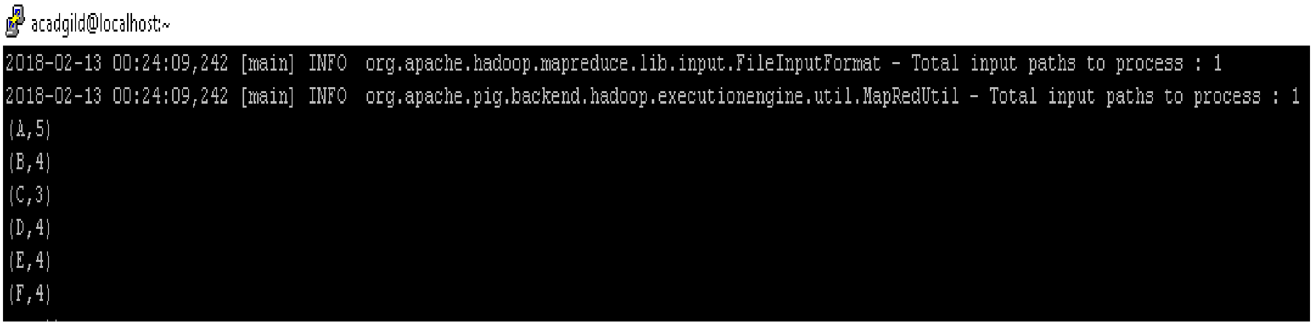
```
grunt>
grunt> line = LOAD '/pig/file.txt' AS (line:chararray);
2018-02-13 00:23:18,224 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> words = FOREACH line GENERATE FLATTEN (TOKENIZE(line,' ')) AS word;
grunt> grouped = GROUP words BY word;
grunt> wordcount = FOREACH grouped GENERATE group, COUNT(words);
grunt> Dump wordcount
2018-02-13 00:23:43,360 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2018-02-13 00:23:43,389 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-02-13 00:23:43,390 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-02-13 00:23:43,391 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrun
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-02-13 00:23:43,397 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100
2018-02-13 00:23:43,399 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreac
2018-02-13 00:23:43,404 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optim
2018-02-13 00:23:43,405 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimi
2018-02-13 00:23:43,424 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-02-13 00:23:43,426 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-02-13 00:23:43,433 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-02-13 00:23:43,435 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markrese
 set, set to default 0.3
2018-02-13 00:23:43,435 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, est
educers.
2018-02-13 00:23:43,435 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: o
adoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2018-02-13 00:23:43,436 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=100
 totalInputFileSize=48
2018-02-13 00:23:43,436 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2018-02-13 00:23:43,437 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be convert
2018-02-13 00:23:43,516 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/acadg
6.0/pig-0.16.0-core-h2.jar to DistributedCache through /tmp/temp-1825490111/tmp120252054/pig-0.16.0-core-h2.jar
2018-02-13 00:23:43,531 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/acadg
6.0/lib/automaton-1.11-8.jar to DistributedCache through /tmp/temp-1825490111/tmp1293722550/automaton-1.11-8.jar
```



acadgild@localhost:~

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime
eature  Outputs
job_1518445375024_0023  1       1       3       3       3       3       4       4       4       4       grouped,line,wordcount,words    GROUP_BY,COMBINER
ocalhost:8020/tmp/temp-1825490111/tmp-284768079,

Input(s):
Successfully read 6 records (409 bytes) from: "/pig/file.txt"

Output(s):
Successfully stored 6 records (60 bytes) in: "hdfs://localhost:8020/tmp/temp-1825490111/tmp-284768079"

Counters:
Total records written : 6
Total bytes written : 60
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1518445375024_0023


2018-02-13 00:24:09,103 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-02-13 00:24:09,106 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
ob history server
2018-02-13 00:24:09,147 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-02-13 00:24:09,152 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
ob history server
2018-02-13 00:24:09,186 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-02-13 00:24:09,189 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redi
ob history server
```

Output:

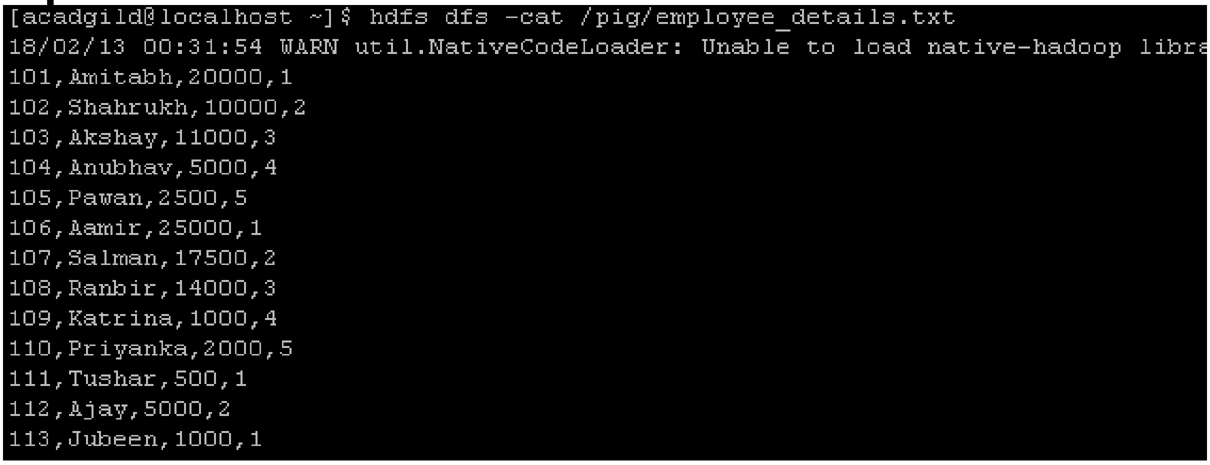


acadgild@localhost:~

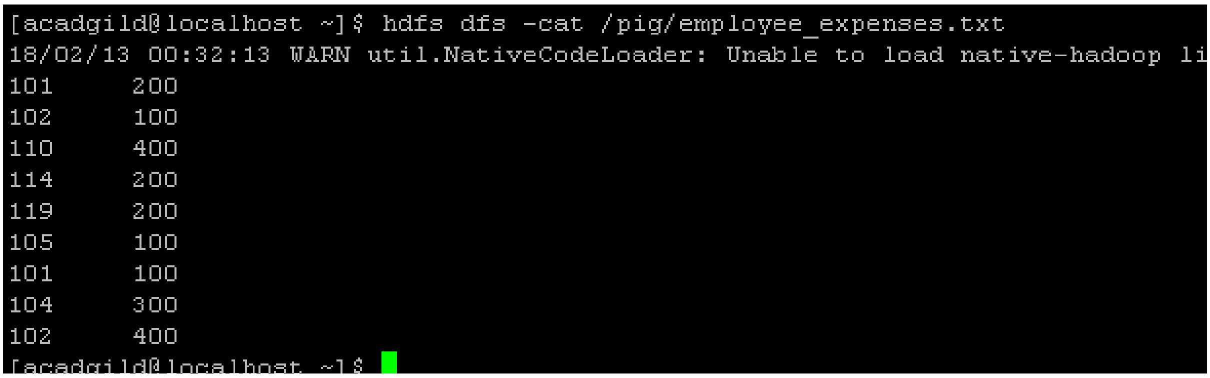
```
2018-02-13 00:24:09,242 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 00:24:09,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(A,5)
(B,4)
(C,3)
(D,4)
(E,4)
(F,4)
```

Task2

Input:

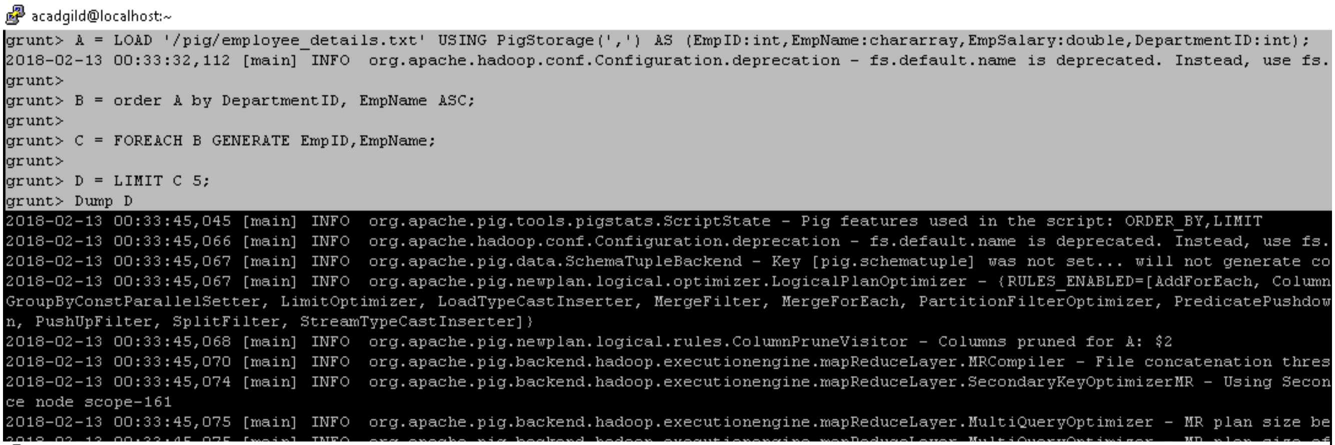


```
[acadgild@localhost ~]$ hdfs dfs -cat /pig/employee_details.txt
18/02/13 00:31:54 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
101,Amitabh,20000,1
102,Shahrukh,10000,2
103,Akshay,11000,3
104,Anubhav,5000,4
105,Pawan,2500,5
106,Aamir,25000,1
107,Salman,17500,2
108,Ranbir,14000,3
109,Katrina,1000,4
110,Priyanka,2000,5
111,Tushar,500,1
112,Ajay,5000,2
113,Jubeen,1000,1
```



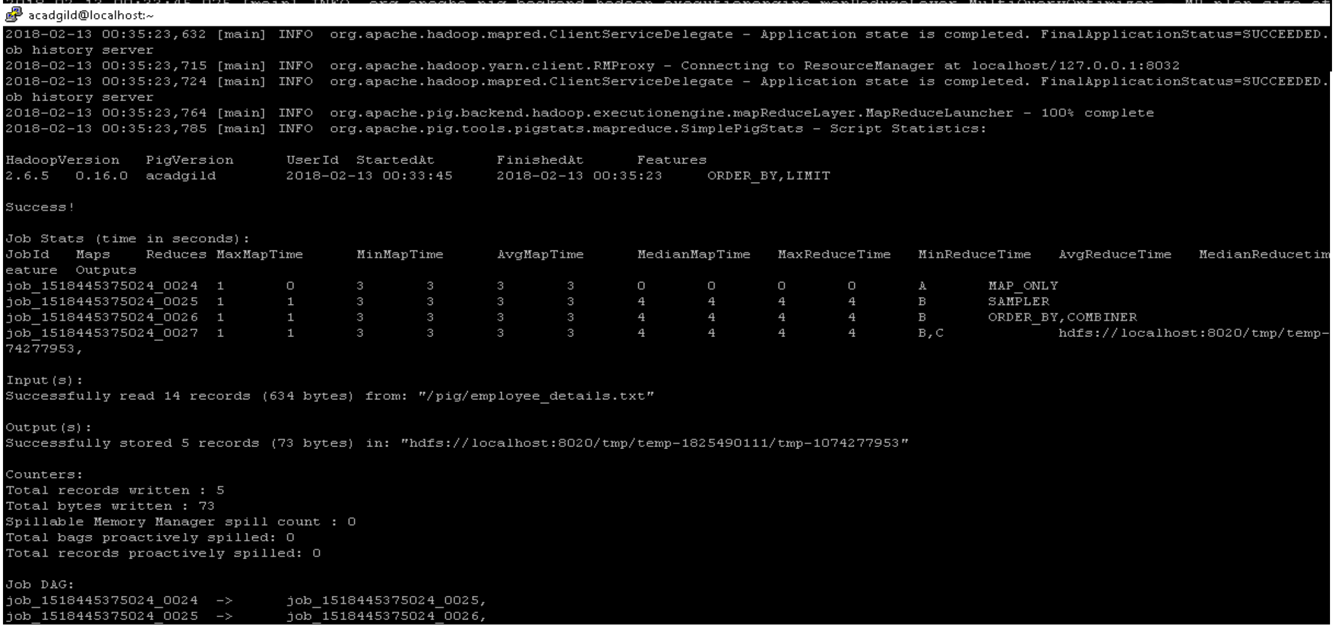
```
[acadgild@localhost ~]$ hdfs dfs -cat /pig/employee_expenses.txt
18/02/13 00:32:13 WARN util.NativeCodeLoader: Unable to load native-hadoop li
101     200
102     100
110     400
114     200
119     200
105     100
101     100
104     300
102     400
[acadgild@localhost ~]$
```

(a) **:** Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)



acadgild@localhost:~

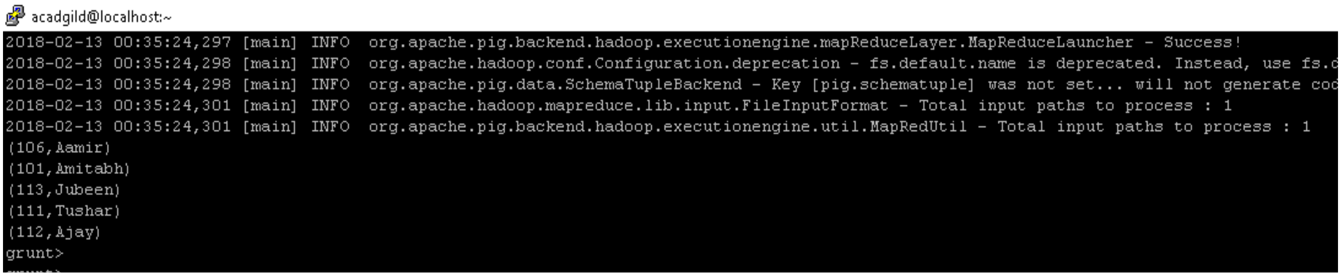
```
grunt> A = LOAD '/pig/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:double,DepartmentID:int);
2018-02-13 00:33:32,112 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
grunt>
grunt> B = order A by DepartmentID, EmpName ASC;
grunt>
grunt> C = FOREACH B GENERATE EmpID,EmpName;
grunt>
grunt> D = LIMIT C 5;
grunt> Dump D
2018-02-13 00:33:45,045 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,LIMIT
2018-02-13 00:33:45,066 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
2018-02-13 00:33:45,067 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate co
2018-02-13 00:33:45,067 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Column
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdow
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-02-13 00:33:45,068 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for A: $2
2018-02-13 00:33:45,070 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thres
2018-02-13 00:33:45,074 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secon
ce node scope-161
2018-02-13 00:33:45,075 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size be
2018-02-13 00:33:45,075 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size af
```



acadgild@localhost:~

```
2018-02-13 00:35:23,632 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED.
ob history server
2018-02-13 00:35:23,715 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-02-13 00:35:23,724 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED.
ob history server
2018-02-13 00:35:23,764 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-02-13 00:35:23,785 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2018-02-13 00:33:45     2018-02-13 00:35:23     ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetim
eature  Outputs
job_1518445375024_0024  1       0       3       3       3       3       0       0       0       0       A       MAP_ONLY
job_1518445375024_0025  1       1       3       3       3       3       4       4       4       4       B       SAMPLER
job_1518445375024_0026  1       1       3       3       3       3       4       4       4       4       B       ORDER_BY,COMBINER
job_1518445375024_0027  1       1       3       3       3       3       4       4       4       4       B,C             hdfs://localhost:8020/tmp/temp-
74277953,

Input(s):
Successfully read 14 records (634 bytes) from: "/pig/employee_details.txt"

Output(s):
Successfully stored 5 records (73 bytes) in: "hdfs://localhost:8020/tmp/temp-1825490111/tmp-1074277953"

Counters:
Total records written : 5
Total bytes written : 73
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1518445375024_0024  ->      job_1518445375024_0025,
job_1518445375024_0025  ->      job_1518445375024_0026,
```

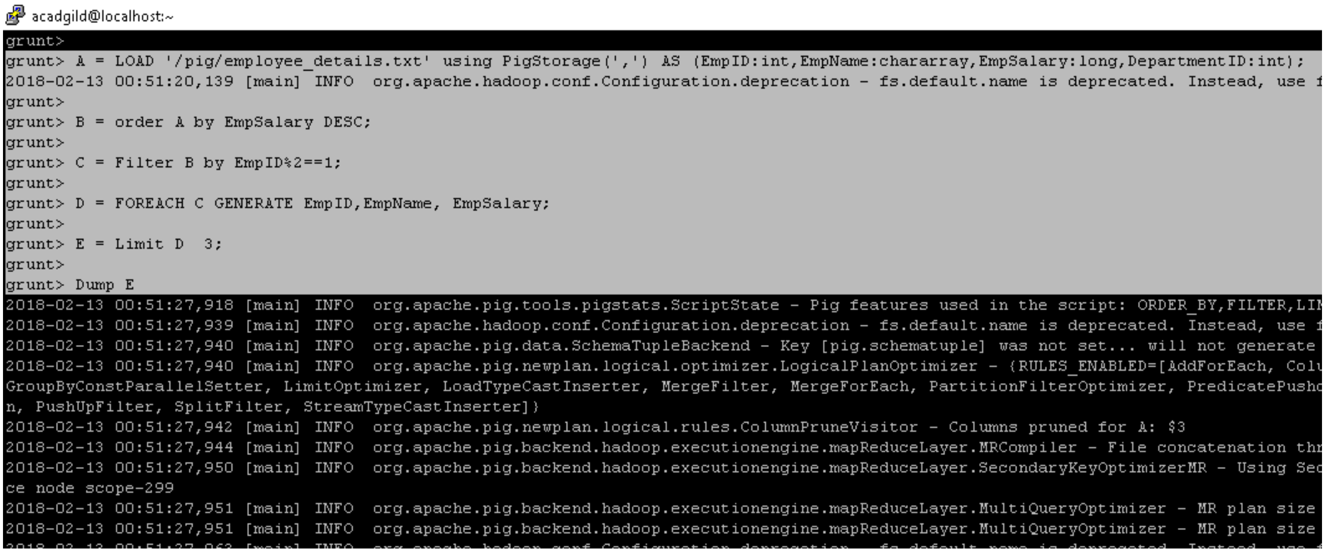
Output:-



acadgild@localhost:~

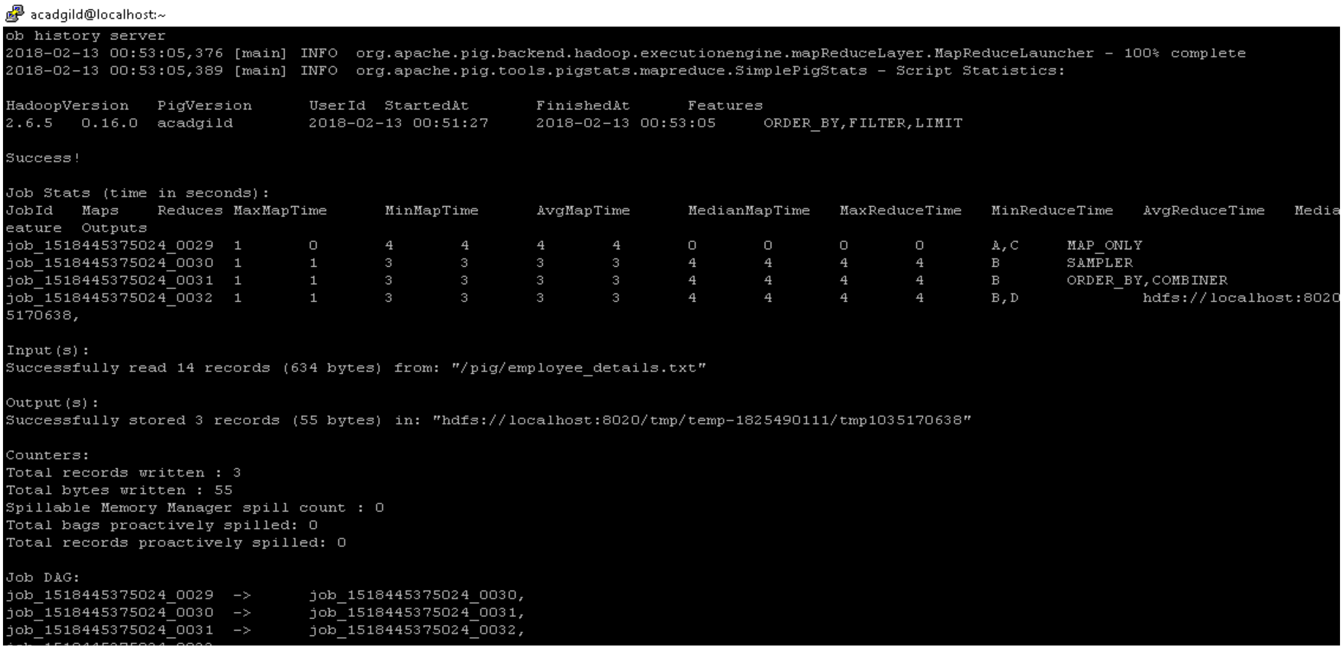
```
2018-02-13 00:35:24,297 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-02-13 00:35:24,298 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
2018-02-13 00:35:24,298 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate cod
2018-02-13 00:35:24,301 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 00:35:24,301 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(106,Aamir)
(101,Amitabh)
(113,Jubeen)
(111,Tushar)
(112,Ajay)
grunt>
```

(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference)



acadgild@localhost:~

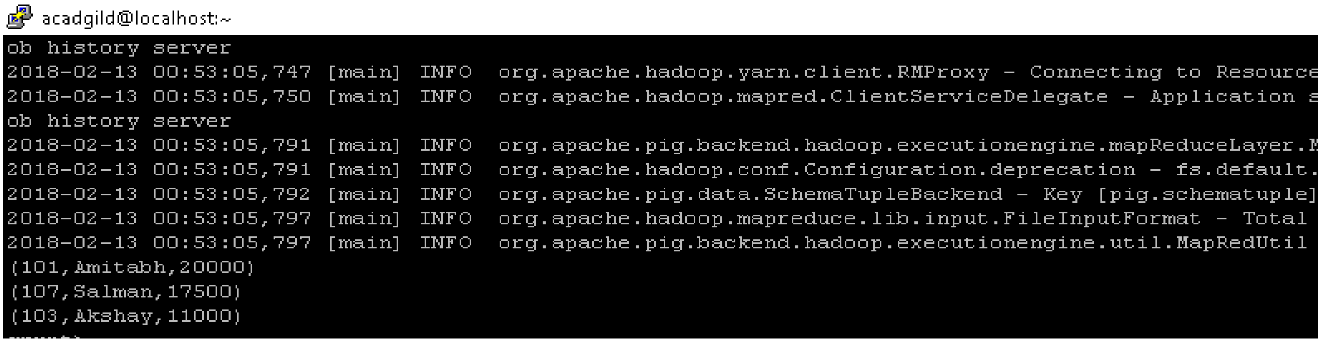
```
grunt>
grunt> A = LOAD '/pig/employee_details.txt' using PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:long,DepartmentID:int);
2018-02-13 00:51:20,139 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
grunt>
grunt> B = order A by EmpSalary DESC;
grunt>
grunt> C = Filter B by EmpID%2==1;
grunt>
grunt> D = FOREACH C GENERATE EmpID,EmpName, EmpSalary;
grunt>
grunt> E = Limit D  3;
grunt>
grunt> Dump E
2018-02-13 00:51:27,918 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIM
2018-02-13 00:51:27,939 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
2018-02-13 00:51:27,940 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
2018-02-13 00:51:27,940 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Colu
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePusho
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-02-13 00:51:27,942 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for A: $3
2018-02-13 00:51:27,944 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thr
2018-02-13 00:51:27,950 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Sec
ce node scope-299
2018-02-13 00:51:27,951 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
2018-02-13 00:51:27,951 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
2018-02-13 00:51:27,963 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
```



acadgild@localhost:~

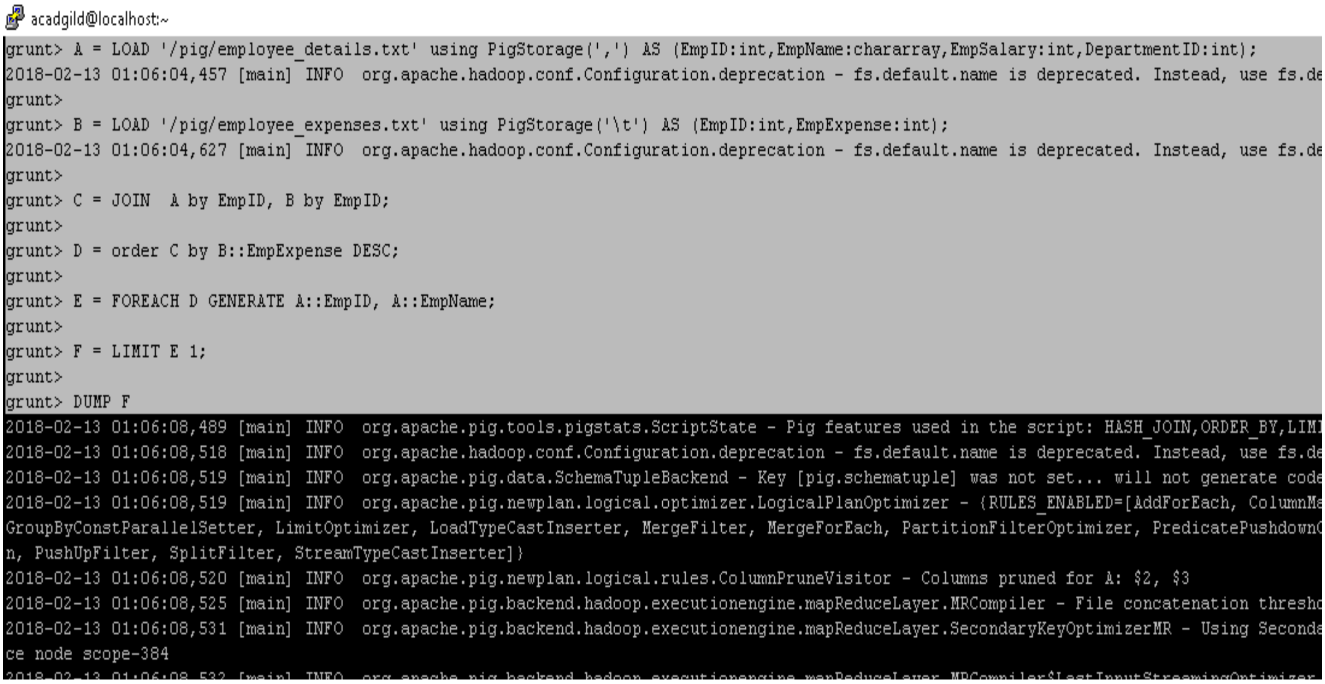
```
ob history server
2018-02-13 00:53:05,376 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-02-13 00:53:05,389 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2018-02-13 00:51:27     2018-02-13 00:53:05     ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   Media
eature  Outputs
job_1518445375024_0029  1       0       4       4       4       4       0       0       0       0       A,C     MAP_ONLY
job_1518445375024_0030  1       1       3       3       3       3       4       4       4       4       B       SAMPLER
job_1518445375024_0031  1       1       3       3       3       3       4       4       4       4       B       ORDER_BY,COMBINER
job_1518445375024_0032  1       1       3       3       3       3       4       4       4       4       B,D             hdfs://localhost:8020
5170638,

Input(s):
Successfully read 14 records (634 bytes) from: "/pig/employee_details.txt"

Output(s):
Successfully stored 3 records (55 bytes) in: "hdfs://localhost:8020/tmp/temp-1825490111/tmp1035170638"

Counters:
Total records written : 3
Total bytes written : 55
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1518445375024_0029  ->      job_1518445375024_0030,
job_1518445375024_0030  ->      job_1518445375024_0031,
job_1518445375024_0031  ->      job_1518445375024_0032,
```

Output:-

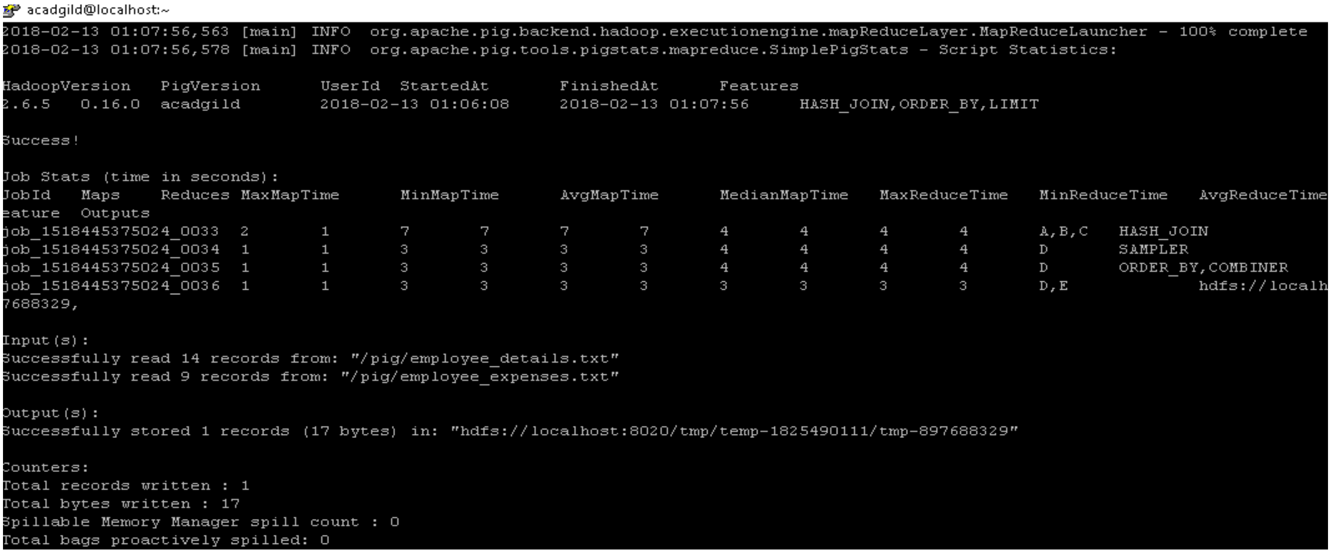


```
ob history server
2018-02-13 00:53:05,747 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to Resource
2018-02-13 00:53:05,750 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application s
ob history server
2018-02-13 00:53:05,791 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.M
2018-02-13 00:53:05,791 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.
2018-02-13 00:53:05,792 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple]
2018-02-13 00:53:05,797 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total
2018-02-13 00:53:05,797 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
(101,Amitabh,20000)
(107,Salman,17500)
(103,Akshay,11000)
```

(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)

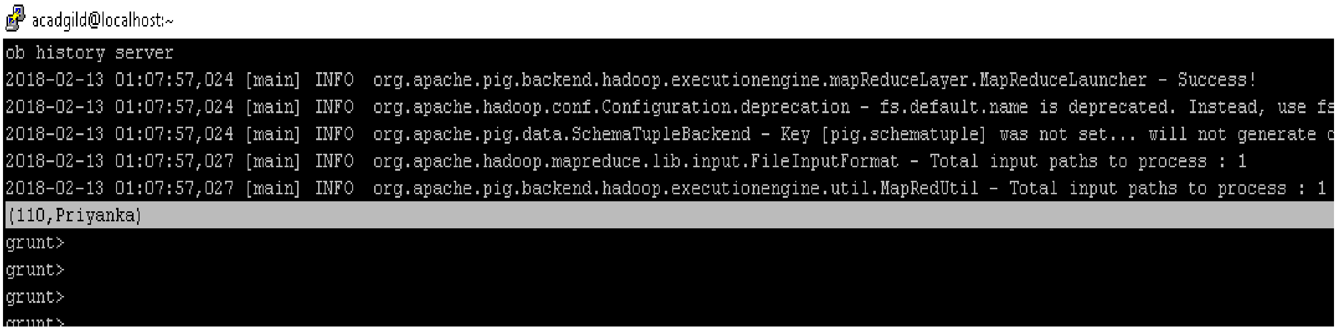


```
grunt> A = LOAD '/pig/employee_details.txt' using PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:int,DepartmentID:int);
2018-02-13 01:06:04,457 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
grunt>
grunt> B = LOAD '/pig/employee_expenses.txt' using PigStorage('\t') AS (EmpID:int,EmpExpense:int);
2018-02-13 01:06:04,627 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
grunt>
grunt> C = JOIN  A by EmpID, B by EmpID;
grunt>
grunt> D = order C by B::EmpExpense DESC;
grunt>
grunt> E = FOREACH D GENERATE A::EmpID, A::EmpName;
grunt>
grunt> F = LIMIT E 1;
grunt>
grunt> DUMP F
2018-02-13 01:06:08,489 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,ORDER_BY,LIMI
2018-02-13 01:06:08,518 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
2018-02-13 01:06:08,519 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
2018-02-13 01:06:08,519 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMa
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownO
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-02-13 01:06:08,520 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for A: $2, $3
2018-02-13 01:06:08,525 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresho
2018-02-13 01:06:08,531 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Seconda
ce node scope-384
2018-02-13 01:06:08,532 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer
```



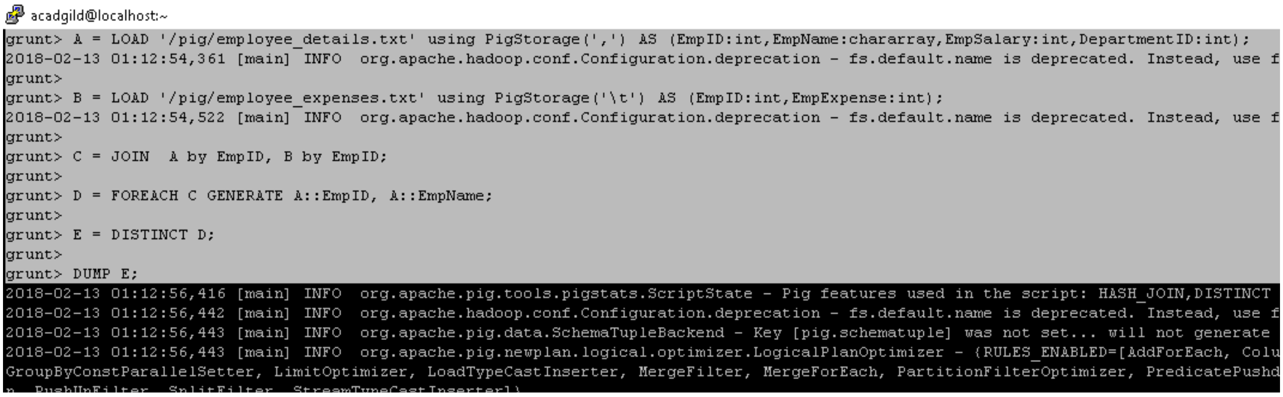
```
acadgild@localhost:~
2018-02-13 01:07:56,563 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-02-13 01:07:56,578 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2018-02-13 01:06:08     2018-02-13 01:07:56     HASH_JOIN,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime
eature  Outputs
job_1518445375024_0033  2       1       7       7       7       7       4       4       4       4       A,B,C   HASH_JOIN
job_1518445375024_0034  1       1       3       3       3       3       4       4       4       4       D       SAMPLER
job_1518445375024_0035  1       1       3       3       3       3       4       4       4       4       D       ORDER_BY,COMBINER
job_1518445375024_0036  1       1       3       3       3       3       3       3       3       3       D,E             hdfs://localh
7688329,

Input(s):
Successfully read 14 records from: "/pig/employee_details.txt"
Successfully read 9 records from: "/pig/employee_expenses.txt"

Output(s):
Successfully stored 1 records (17 bytes) in: "hdfs://localhost:8020/tmp/temp-1825490111/tmp-897688329"

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

Output:-

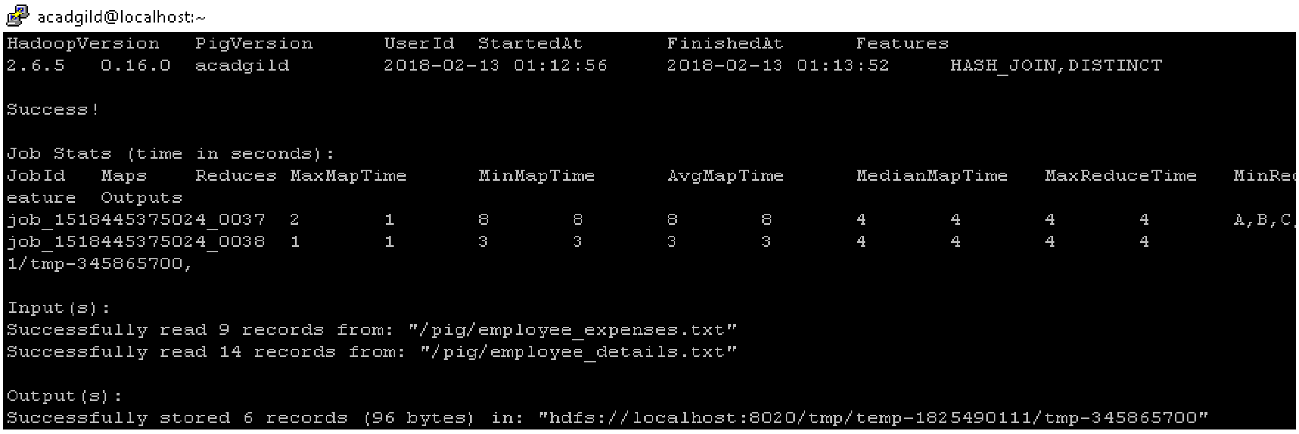


```
acadgild@localhost:~
ob history server
2018-02-13 01:07:57,024 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-02-13 01:07:57,024 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs
2018-02-13 01:07:57,024 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate c
2018-02-13 01:07:57,027 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 01:07:57,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
grunt>
grunt>
grunt>
grunt>
```

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

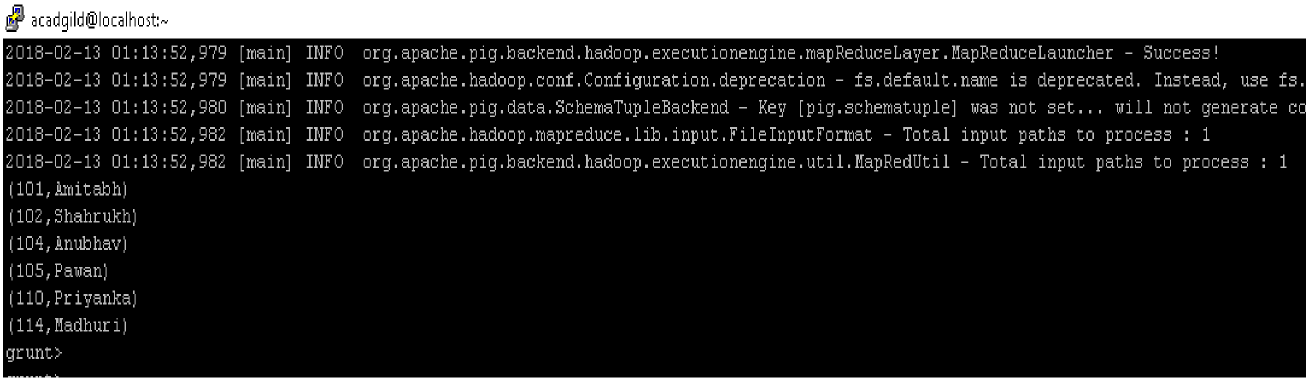


```
acadgild@localhost:~
grunt> A = LOAD '/pig/employee_details.txt' using PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:int,DepartmentID:int);
2018-02-13 01:12:54,361 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
grunt>
grunt> B = LOAD '/pig/employee_expenses.txt' using PigStorage('\t') AS (EmpID:int,EmpExpense:int);
2018-02-13 01:12:54,522 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
grunt>
grunt> C = JOIN  A by EmpID, B by EmpID;
grunt>
grunt> D = FOREACH C GENERATE A::EmpID, A::EmpName;
grunt>
grunt> E = DISTINCT D;
grunt>
grunt> DUMP E;
2018-02-13 01:12:56,416 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,DISTINCT
2018-02-13 01:12:56,442 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
2018-02-13 01:12:56,443 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
2018-02-13 01:12:56,443 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Colu
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushd
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
```



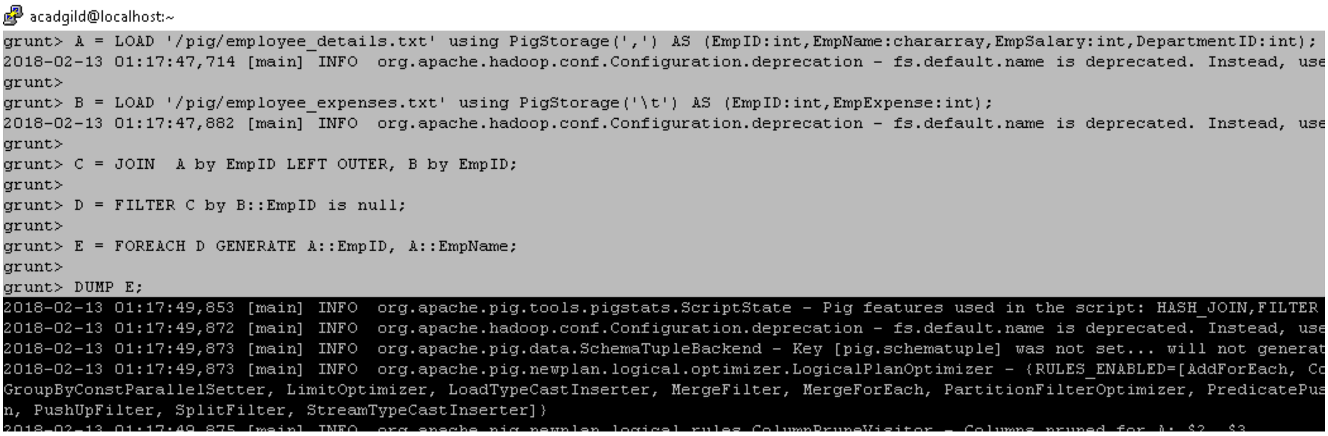
```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2018-02-13 01:12:56     2018-02-13 01:13:52     HASH_JOIN,DISTINCT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinRe
eature  Outputs
job_1518445375024_0037  2       1       8       8       8       8       4       4       4       4       A,B,C,
job_1518445375024_0038  1       1       3       3       3       3       4       4       4       4
1/tmp-345865700,

Input(s):
Successfully read 9 records from: "/pig/employee_expenses.txt"
Successfully read 14 records from: "/pig/employee_details.txt"

Output(s):
Successfully stored 6 records (96 bytes) in: "hdfs://localhost:8020/tmp/temp-1825490111/tmp-345865700"
```

Output:-



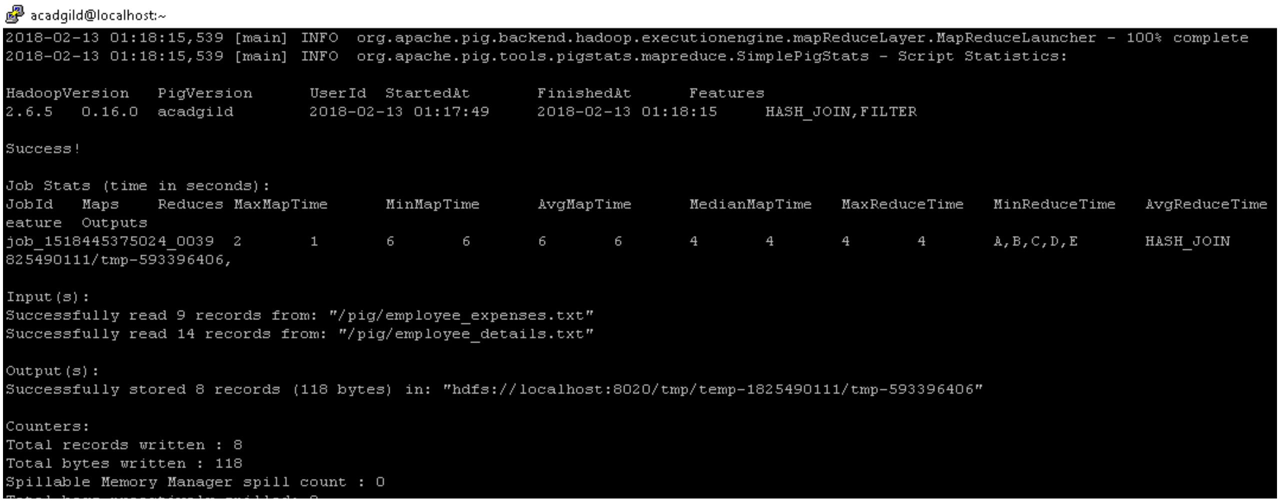
```
2018-02-13 01:13:52,979 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-02-13 01:13:52,979 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
2018-02-13 01:13:52,980 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate co
2018-02-13 01:13:52,982 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 01:13:52,982 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt>
```

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



acadgild@localhost:~

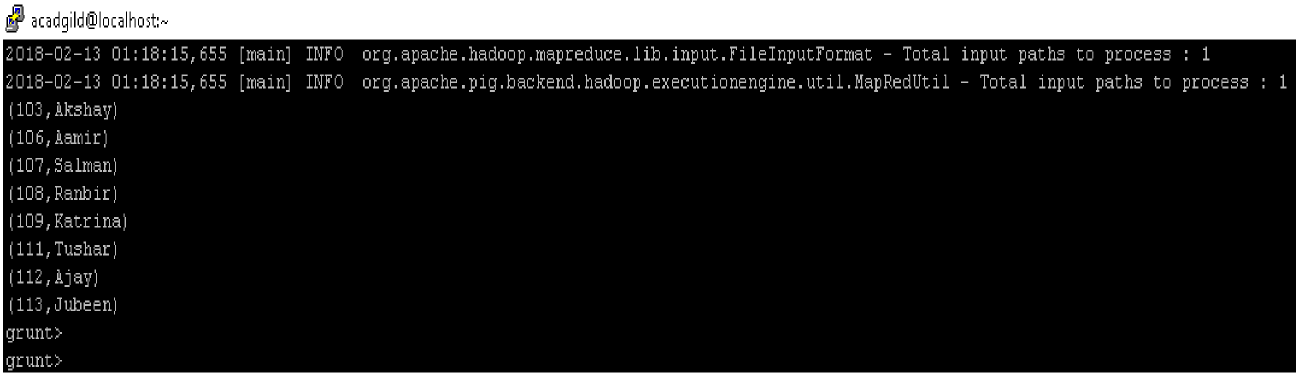
```
grunt> A = LOAD '/pig/employee_details.txt' using PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:int,DepartmentID:int);
2018-02-13 01:17:47,714 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
grunt>
grunt> B = LOAD '/pig/employee_expenses.txt' using PigStorage('\t') AS (EmpID:int,EmpExpense:int);
2018-02-13 01:17:47,882 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
grunt>
grunt> C = JOIN  A by EmpID LEFT OUTER, B by EmpID;
grunt>
grunt> D = FILTER C by B::EmpID is null;
grunt>
grunt> E = FOREACH D GENERATE A::EmpID, A::EmpName;
grunt>
grunt> DUMP E;
2018-02-13 01:17:49,853 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,FILTER
2018-02-13 01:17:49,872 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
2018-02-13 01:17:49,873 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generat
2018-02-13 01:17:49,873 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Co
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePus
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-02-13 01:17:49,875 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for A: $2, $3
```



acadgild@localhost:~

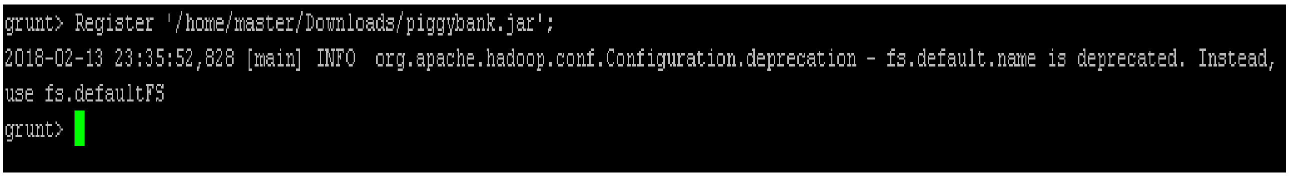
```
2018-02-13 01:18:15,539 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-02-13 01:18:15,539 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2018-02-13 01:17:49     2018-02-13 01:18:15     HASH_JOIN,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime
eature  Outputs
job_1518445375024_0039  2       1       6       6       6       6       4       4       4       4       A,B,C,D,E       HASH_JOIN
825490111/tmp-593396406,

Input(s):
Successfully read 9 records from: "/pig/employee_expenses.txt"
Successfully read 14 records from: "/pig/employee_details.txt"

Output(s):
Successfully stored 8 records (118 bytes) in: "hdfs://localhost:8020/tmp/temp-1825490111/tmp-593396406"

Counters:
Total records written : 8
Total bytes written : 118
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
```

Output:-



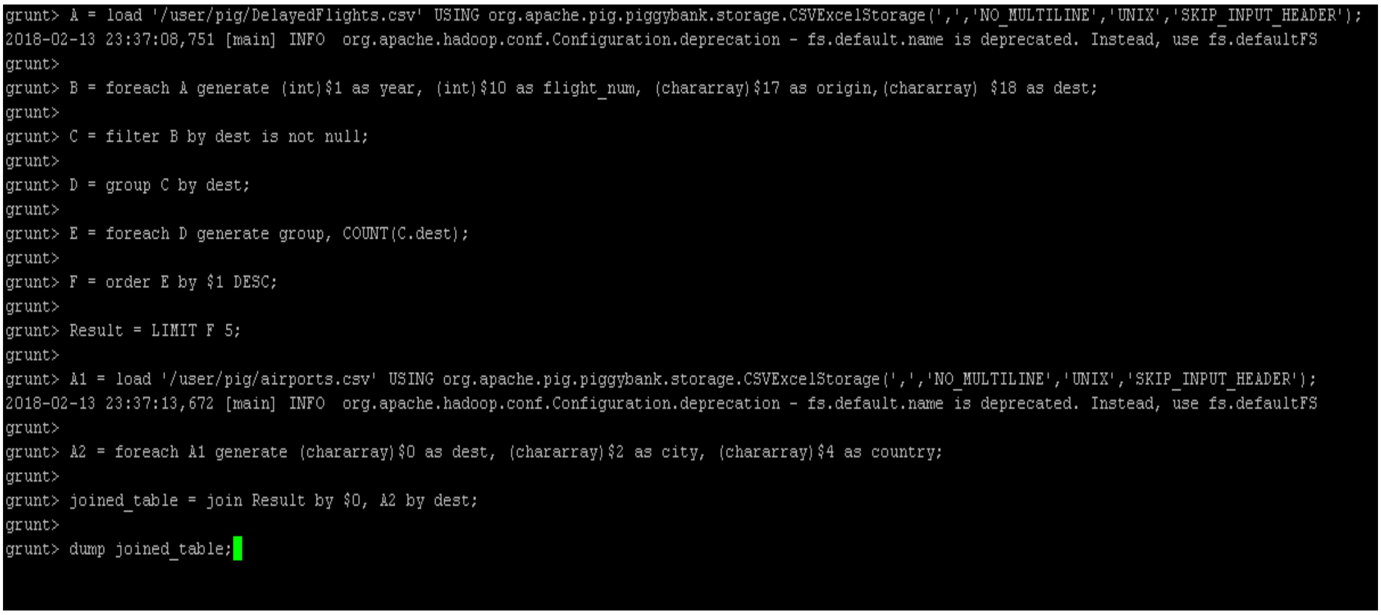
```
2018-02-13 01:18:15,655 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 01:18:15,655 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
grunt>
grunt>
```

Task 3: Implementing the use case: aviation-data-analysis-using-apache-pig.



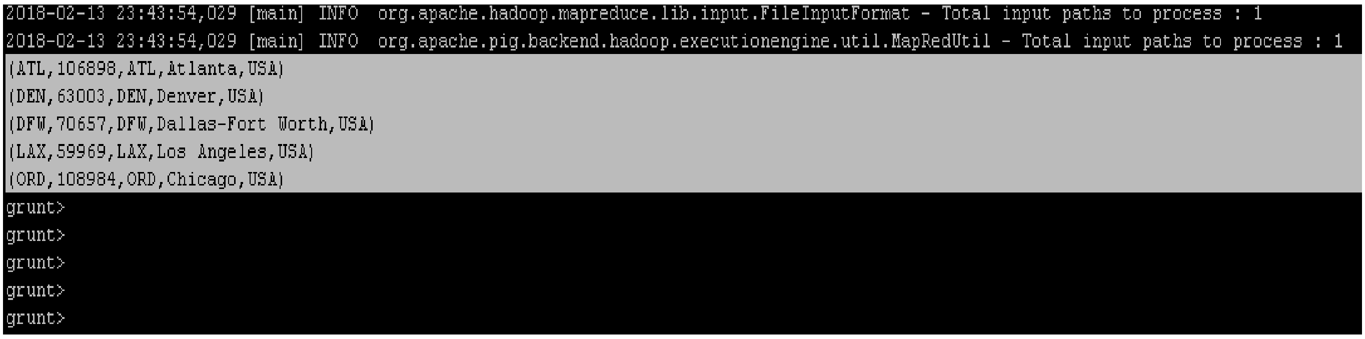
```
grunt> Register '/home/master/Downloads/piggybank.jar';
2018-02-13 23:35:52,828 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
grunt>
```

Problem1: Top 5 most visited destinations.



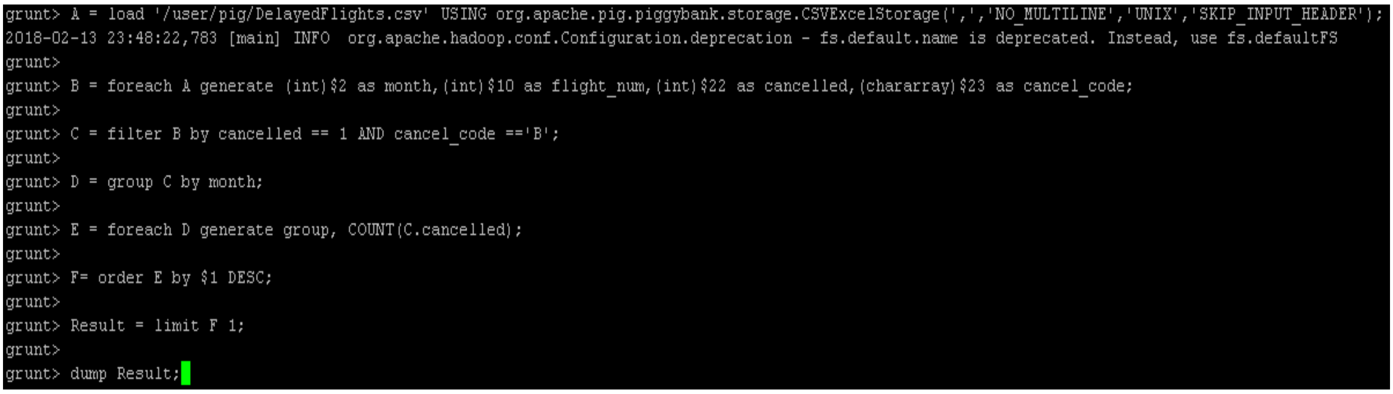
```
grunt> A = load '/user/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-02-13 23:37:08,751 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
grunt>
grunt> C = filter B by dest is not null;
grunt>
grunt> D = group C by dest;
grunt>
grunt> E = foreach D generate group, COUNT(C.dest);
grunt>
grunt> F = order E by $1 DESC;
grunt>
grunt> Result = LIMIT F 5;
grunt>
grunt> A1 = load '/user/pig/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-02-13 23:37:13,672 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
grunt>
grunt> joined_table = join Result by $0, A2 by dest;
grunt>
grunt> dump joined_table;
```

Output:-

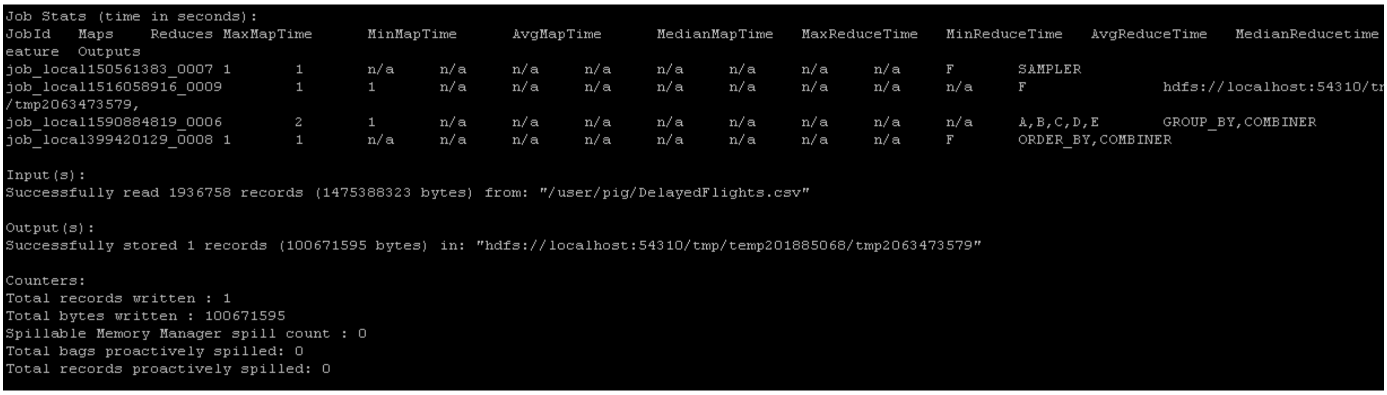


```
2018-02-13 23:43:54,029 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 23:43:54,029 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(ATL,106898,ATL,Atlanta,USA)
(DEN,63003,DEN,Denver,USA)
(DFW,70657,DFW,Dallas-Fort Worth,USA)
(LAX,59969,LAX,Los Angeles,USA)
(ORD,108984,ORD,Chicago,USA)
grunt>
grunt>
grunt>
grunt>
grunt>
```

Problem2: Which month has seen the most number of cancellations due to bad weather.

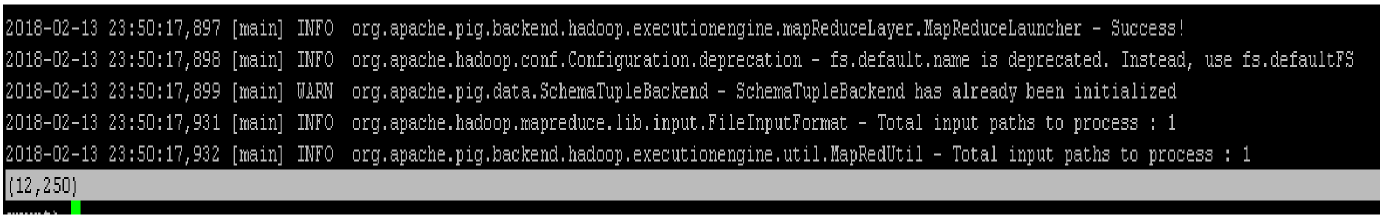


```
grunt> A = load '/user/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-02-13 23:48:22,783 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt>
grunt> C = filter B by cancelled == 1 AND cancel_code =='B';
grunt>
grunt> D = group C by month;
grunt>
grunt> E = foreach D generate group, COUNT(C.cancelled);
grunt>
grunt> F= order E by $1 DESC;
grunt>
grunt> Result = limit F 1;
grunt>
grunt> dump Result;
```



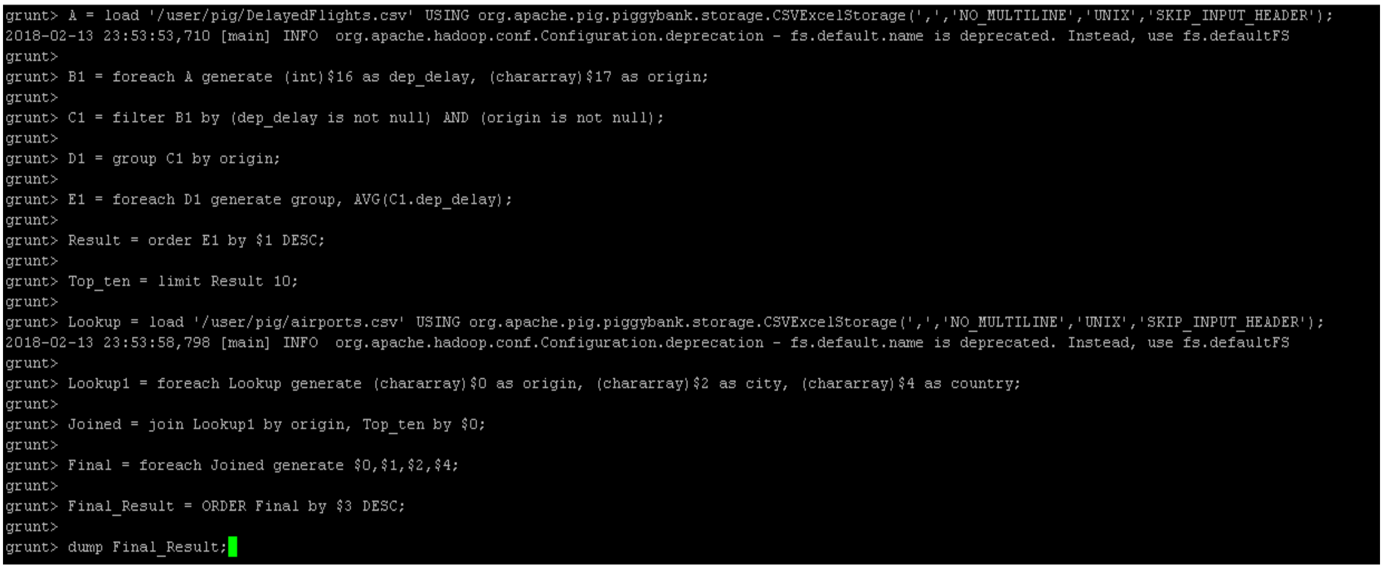
```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime
eature  Outputs
job_local150561383_0007 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F       SAMPLER
job_local1516058916_0009        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F               hdfs://localhost:54310/tr
/tmp2063473579,
job_local1590884819_0006        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     A,B,C,D,E       GROUP_BY,COMBINER
job_local399420129_0008 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F       ORDER_BY,COMBINER

Input(s):
Successfully read 1936758 records (1475388323 bytes) from: "/user/pig/DelayedFlights.csv"

Output(s):
Successfully stored 1 records (100671595 bytes) in: "hdfs://localhost:54310/tmp/temp201885068/tmp2063473579"

Counters:
Total records written : 1
Total bytes written : 100671595
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

Output:-

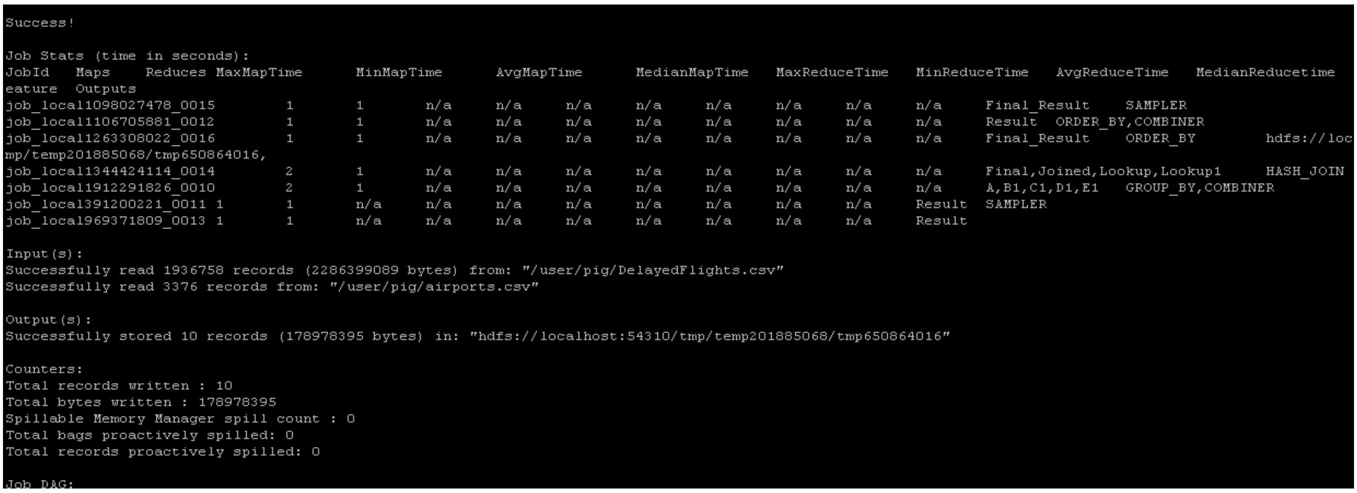


```
2018-02-13 23:50:17,897 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-02-13 23:50:17,898 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-02-13 23:50:17,899 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-02-13 23:50:17,931 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 23:50:17,932 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt>
```

Problem3: Top ten origins with the highest AVG departure delay.

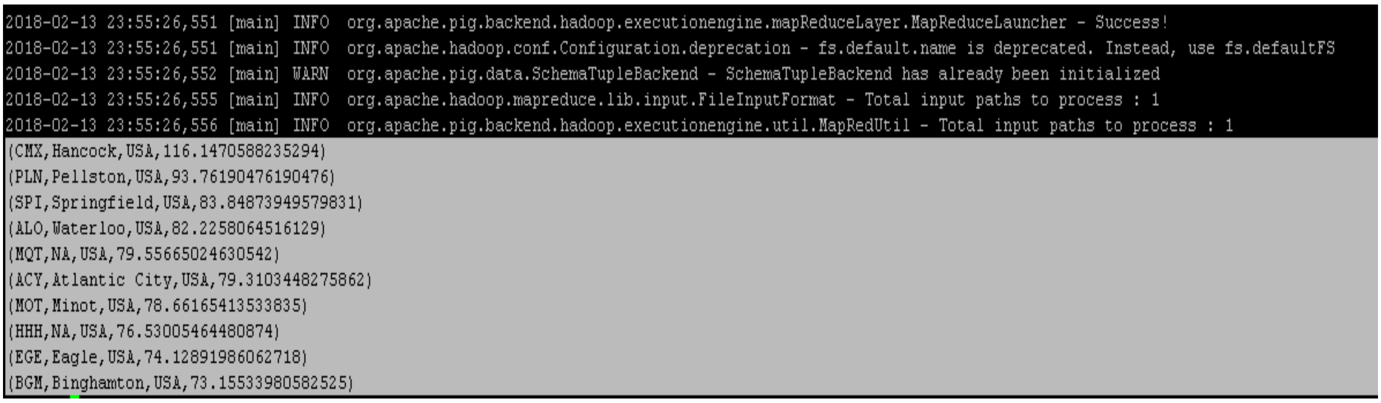


```
grunt> A = load '/user/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-02-13 23:53:53,710 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
grunt>
grunt> C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
grunt>
grunt> D1 = group C1 by origin;
grunt>
grunt> E1 = foreach D1 generate group, AVG(C1.dep_delay);
grunt>
grunt> Result = order E1 by $1 DESC;
grunt>
grunt> Top_ten = limit Result 10;
grunt>
grunt> Lookup = load '/user/pig/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-02-13 23:53:58,798 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
grunt>
grunt> Joined = join Lookup1 by origin, Top_ten by $0;
grunt>
grunt> Final = foreach Joined generate $0,$1,$2,$4;
grunt>
grunt> Final_Result = ORDER Final by $3 DESC;
grunt>
grunt> dump Final_Result;
```



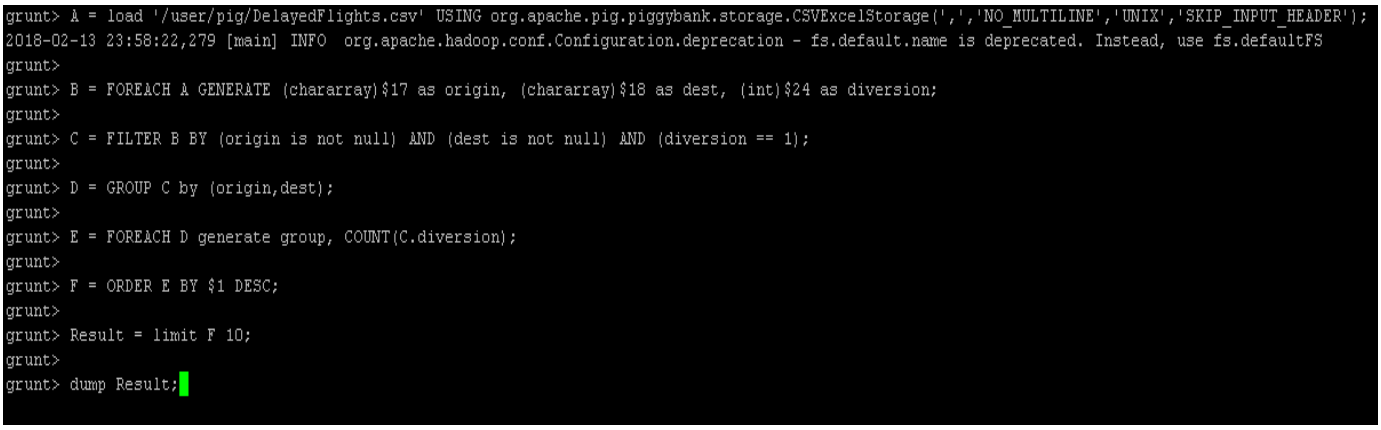
```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime
eature  Outputs
job_local1098027478_0015        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Final_Result    SAMPLER
job_local1106705881_0012        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Result  ORDER_BY,COMBINER
job_local1263308022_0016        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Final_Result    ORDER_BY        hdfs://loc
mp/temp201885068/tmp650864016,
job_local1344424114_0014        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Final,Joined,Lookup,Lookup1     HASH_JOIN
job_local1912291826_0010        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     A,B1,C1,D1,E1   GROUP_BY,COMBINER
job_local391200221_0011 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Result  SAMPLER
job_local969371809_0013 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Result

Input(s):
Successfully read 1936758 records (2286399089 bytes) from: "/user/pig/DelayedFlights.csv"
Successfully read 3376 records from: "/user/pig/airports.csv"

Output(s):
Successfully stored 10 records (178978395 bytes) in: "hdfs://localhost:54310/tmp/temp201885068/tmp650864016"

Counters:
Total records written : 10
Total bytes written : 178978395
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
```

Output:-

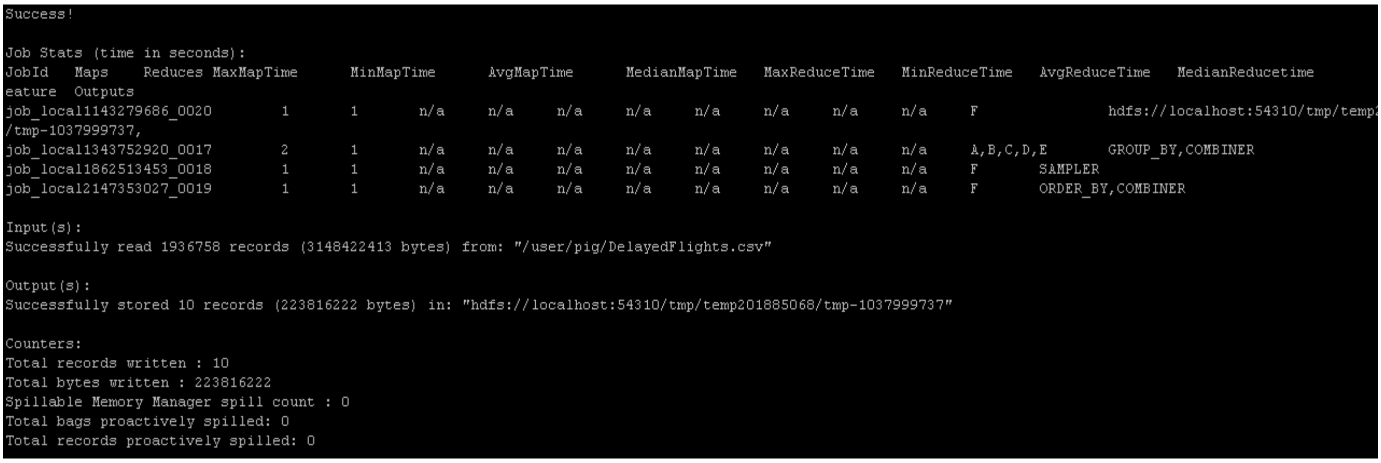


```
2018-02-13 23:55:26,551 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-02-13 23:55:26,551 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-02-13 23:55:26,552 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-02-13 23:55:26,555 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-13 23:55:26,556 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831)
(ALO,Waterloo,USA,82.2258064516129)
(MQT,NA,USA,79.55665024630542)
(ACY,Atlantic City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.53005464480874)
(EGE,Eagle,USA,74.12891986062718)
(BGM,Binghamton,USA,73.15533980582525)
```

Problem4: Which route (origin & destination) has seen the maximum diversion.

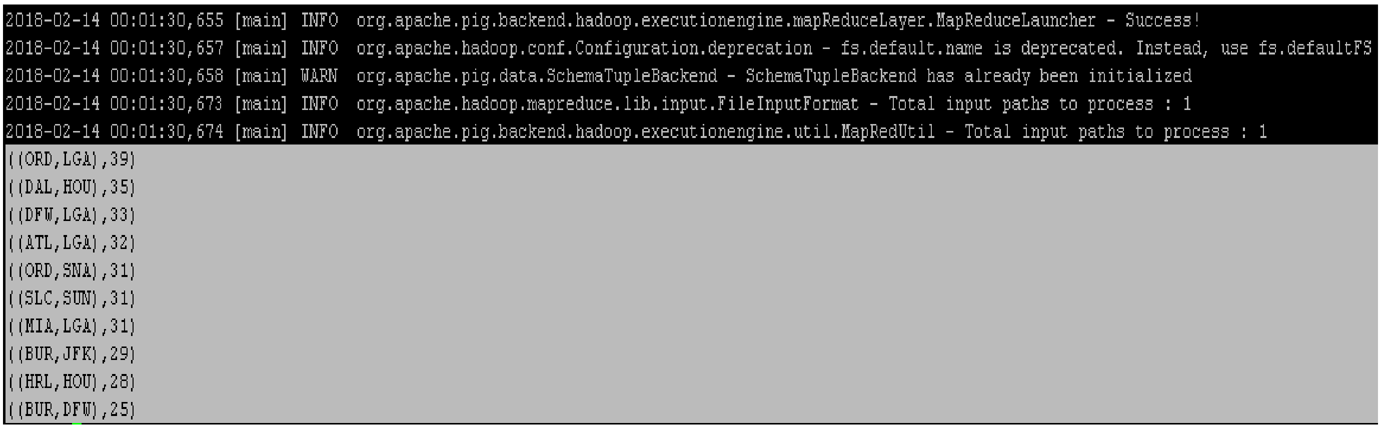


```
grunt> A = load '/user/pig/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-02-13 23:58:22,279 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
grunt>
grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
grunt>
grunt> D = GROUP C by (origin,dest);
grunt>
grunt> E = FOREACH D generate group, COUNT(C.diversion);
grunt>
grunt> F = ORDER E BY $1 DESC;
grunt>
grunt> Result = limit F 10;
grunt>
grunt> dump Result;
```



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime
eature  Outputs
job_local1143279686_0020        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F               hdfs://localhost:54310/tmp/temp2
/tmp-1037999737,
job_local1343752920_0017        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     A,B,C,D,E       GROUP_BY,COMBINER
job_local1862513453_0018        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F       SAMPLER
job_local2147353027_0019        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F       ORDER_BY,COMBINER

Input(s):
Successfully read 1936758 records (3148422413 bytes) from: "/user/pig/DelayedFlights.csv"

Output(s):
Successfully stored 10 records (223816222 bytes) in: "hdfs://localhost:54310/tmp/temp201885068/tmp-1037999737"

Counters:
Total records written : 10
Total bytes written : 223816222
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

Output:-



```
2018-02-14 00:01:30,655 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-02-14 00:01:30,657 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-02-14 00:01:30,658 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-02-14 00:01:30,673 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-02-14 00:01:30,674 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((ORD,LGA),39)
((DAL,HOU),35)
((DFW,LGA),33)
((ATL,LGA),32)
((ORD,SNA),31)
((SLC,SUN),31)
((MIA,LGA),31)
((BUR,JFK),29)
((HRL,HOU),28)
((BUR,DFW),25)
```