

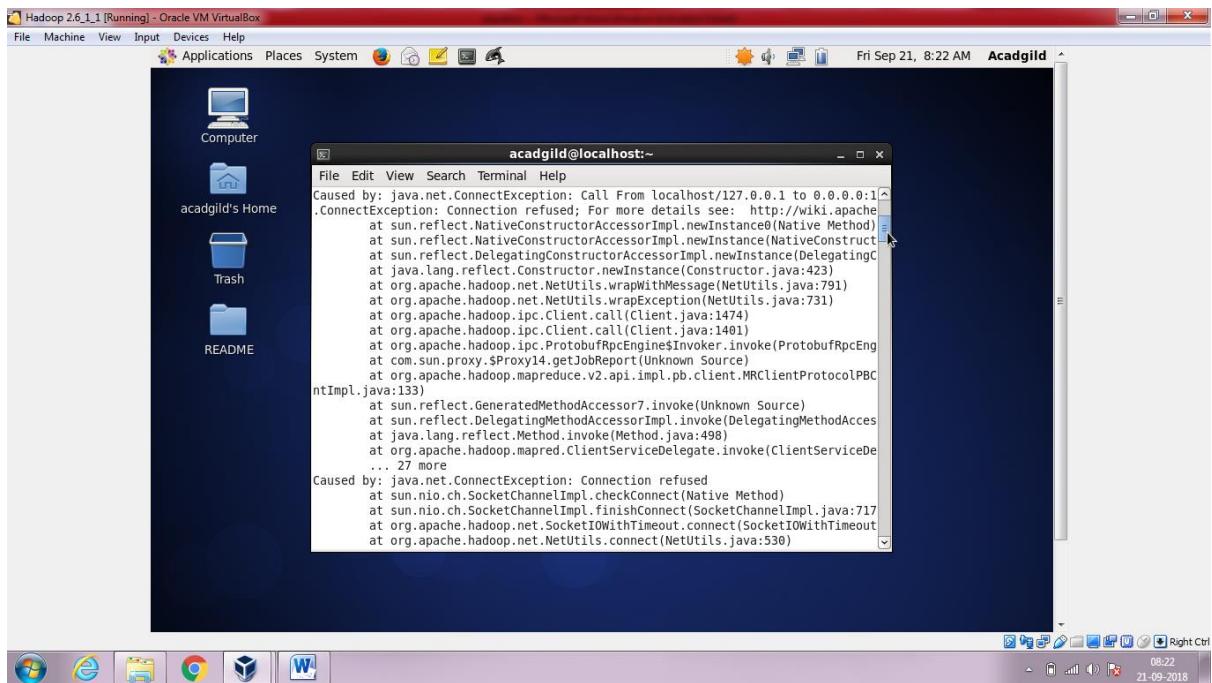
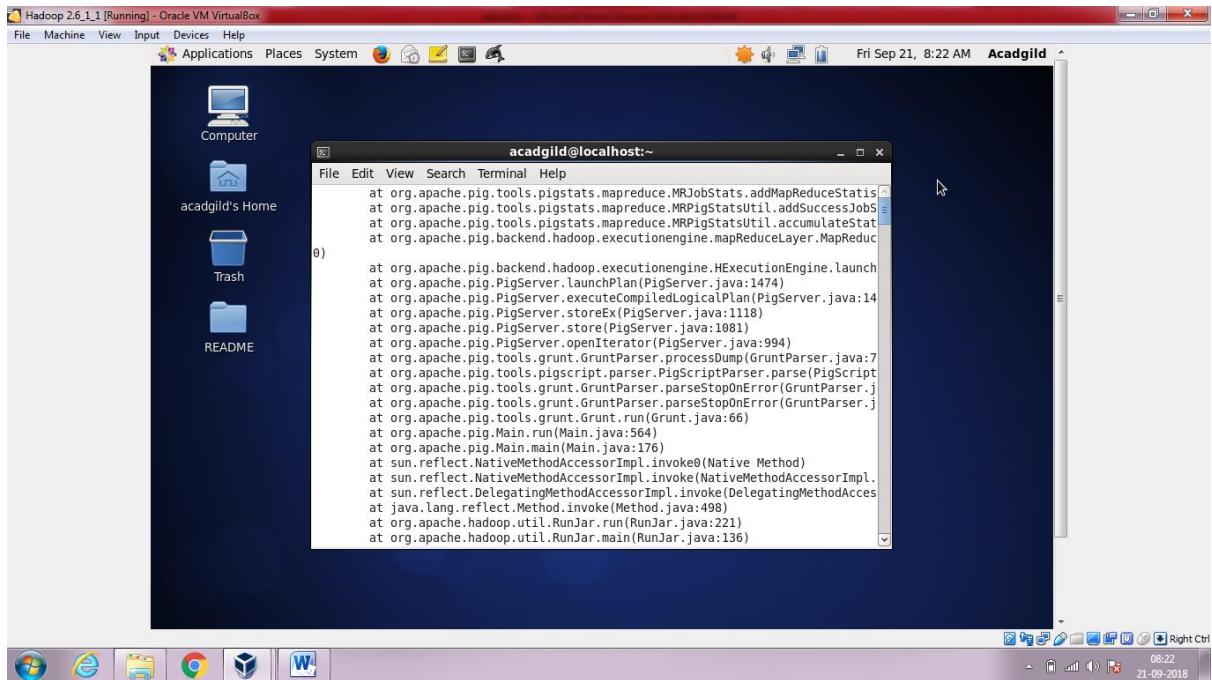
Task 1

Write program to implement word count using pig.

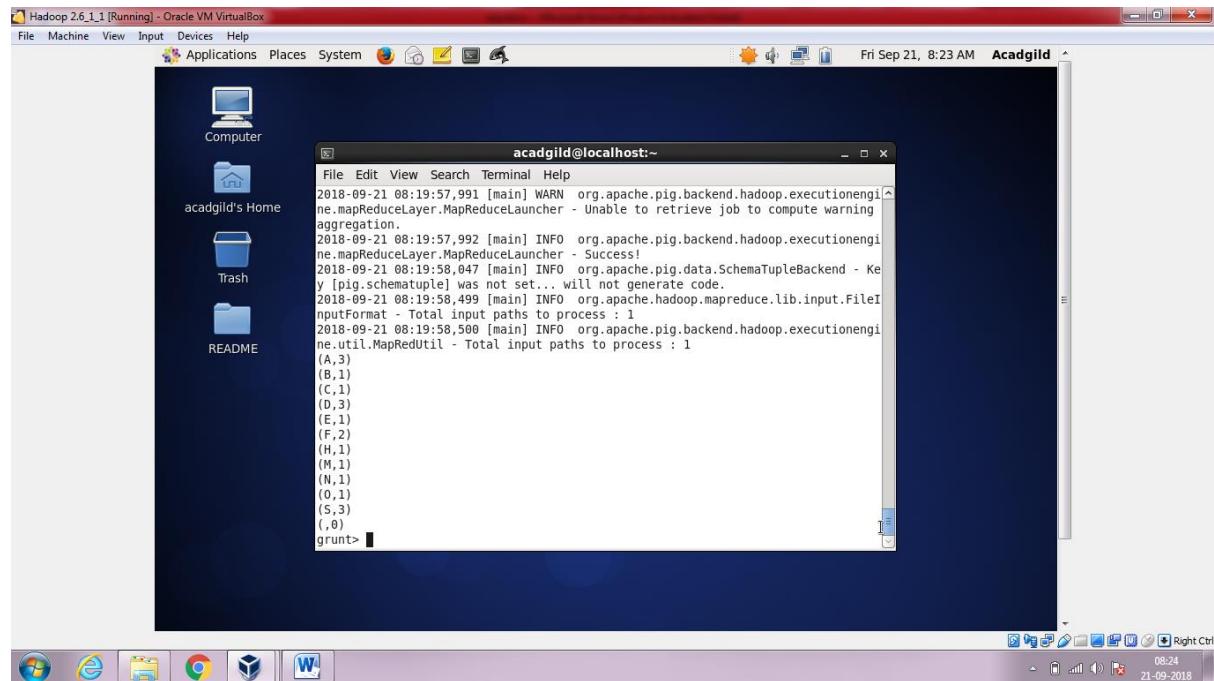
```
[acadgild@localhost ~]$ hadoop fs -cat /file.txt
18/09/20 10:46:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
A S D
S D F
A B C
A E D
S F H
M N O

[acadgild@localhost ~]$ pig
18/09/20 10:47:21 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/20 10:47:21 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/20 10:47:21 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-20 10:47:21,982 [main] INFO org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-09-20 10:47:21,982 [main] INFO org.apache.pig.Main - Logging error message
```

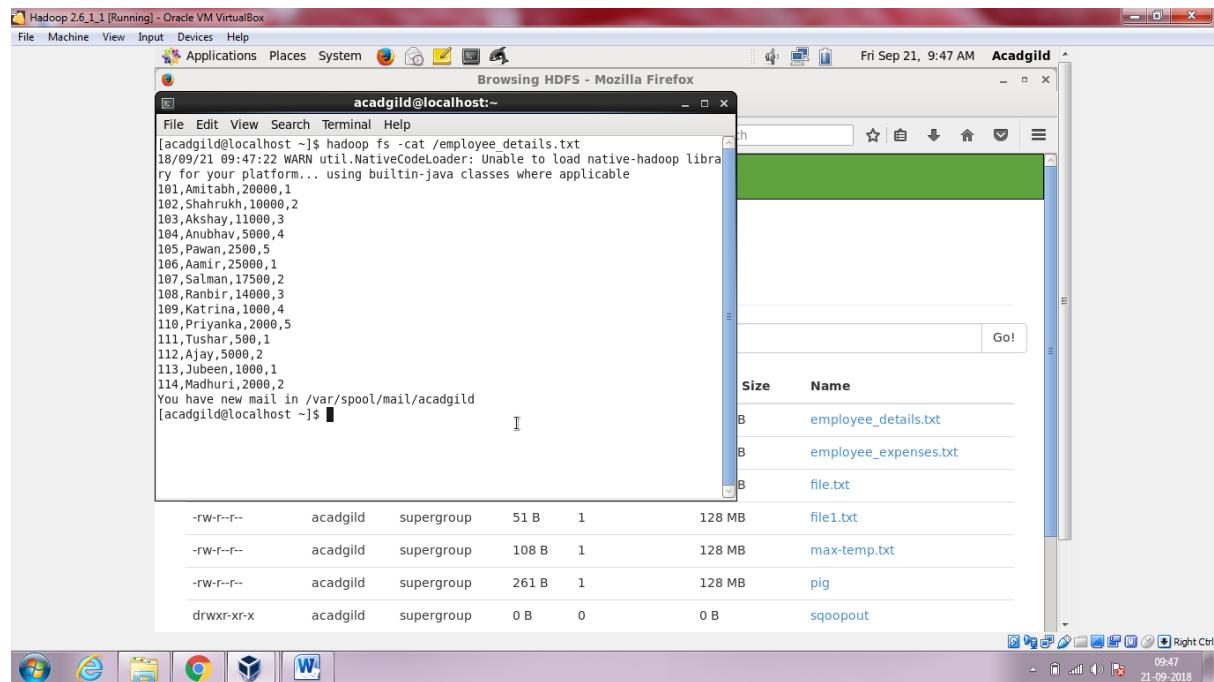
```
grunt> line = LOAD '/file.txt' AS (line:chararray);
2018-09-21 08:15:23,308 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> words = FOREACH line GENERATE FLATTEN (TOKENIZE(line, ' ')) AS word;
2018-09-21 08:15:29,716 [main] INFO org.apache.pig.impl.util.SplittableMemoryManager - Selected heap (Tenured Gen) of size 699972512 to monitor. collectionUsage Threshold = 489350752, usageThreshold = 489350752
grunt> grouped = GROUP words BY word;
grunt> wordcount = FOREACH grouped GENERATE group, COUNT(words);
grunt> Dump wordcount
2018-09-21 08:15:45,378 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2018-09-21 08:15:45,546 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-21 08:15:45,565 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-21 08:15:45,668 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-09-21 08:15:45,983 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-09-21 08:15:46,052 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2018-09-21 08:15:46,154 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2018-09-21 08:15:46,154 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2018-09-21 08:15:46,296 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-21 08:15:46,445 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-09-21 08:15:47,502 [main] INFO org.apache.pig.tools.pigcerver.Master - MPC
```



Output:



Task2:input file



Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Fri Sep 21, 9:48 AM Acadgild

```
acadgild@localhost:~
```

```

File Edit View Search Terminal Help
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
114,Madhuri,2000,2
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /employee_expenses.txt
18/09/21 09:48:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
101    200
102    100
110    400
114    200
119    200
105    100
101    100
104    300
102    400
[acadgild@localhost ~]$
```

09:48 21-09-2018 Right Ctrl

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

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

```
\\\" ...
"run" ...
"exec" ...
"OrDefault" ...
"DECLARE" ...
"scriptDone" ...
"%" ...
"%" ...
<EOL> ...
";" ...
```

```
Details at logfile: /home/acadgild/pig_1537503661588.log
grunt> ALOAD "/employee_details.txt" USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:double,DepartmentID:int)
words = FOREACH line GENERATE FLATTEN (TOKENIZE(line,'')) AS word;
2018-09-21 10:06:07,504 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse:
<line 1, column 16> Undefined alias: line
Details at logfile: /home/acadgild/pig_1537503661588.log
grunt>
```

```
A = LOAD "/employee_details.txt" USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:double,DepartmentID:int);
<line 1, column 9> Unexpected character '"'
2018-09-21 10:07:23,941 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 1, column 9> Unexpected character '"'
Details at logfile: /home/acadgild/pig_1537503661588.log
grunt> A = LOAD "/employee_details.txt" USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:double,DepartmentID:int);
words = FOREACH line GENERATE FLATTEN (TOKENIZE(line,'')) AS word;
2018-09-21 10:07:36,256 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse:
<line 1, column 16> Undefined alias: line
Details at logfile: /home/acadgild/pig_1537503661588.log
grunt> A = LOAD '/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:double,DepartmentID:int);
2018-09-21 10:08:38,283 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = order A by DepartmentID, EmpName ASC;
grunt> C = FOREACH B GENERATE EmpID,EmpName;
grunt> D = LIMIT C 5;
grunt> Dump D
```

[Browsing HDFS - Mozilla Firefox acadgild@localhost:~]

10:12 21-09-2018 Right Ctrl

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Fri Sep 21, 10:18 AM Acadgild

Computer

acadgild's Home

Trash

README

acadgild@localhost:~

```
Dump D
2018-09-21 10:16:20,964 [main] INFO org.apache.pig.tools.pigstats.ScriptState -
  Pig features used in the script: ORDER BY,LIMIT
2018-09-21 10:16:21,115 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
  fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-21 10:16:21,150 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-21 10:16:21,281 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-09-21 10:16:21,369 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for A: $2
2018-09-21 10:16:21,498 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2018-09-21 10:16:21,639 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-09-21 10:16:21,777 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-23
```

Output:

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Fri Sep 21, 10:31 AM Acadgild

Computer

acadgild's Home

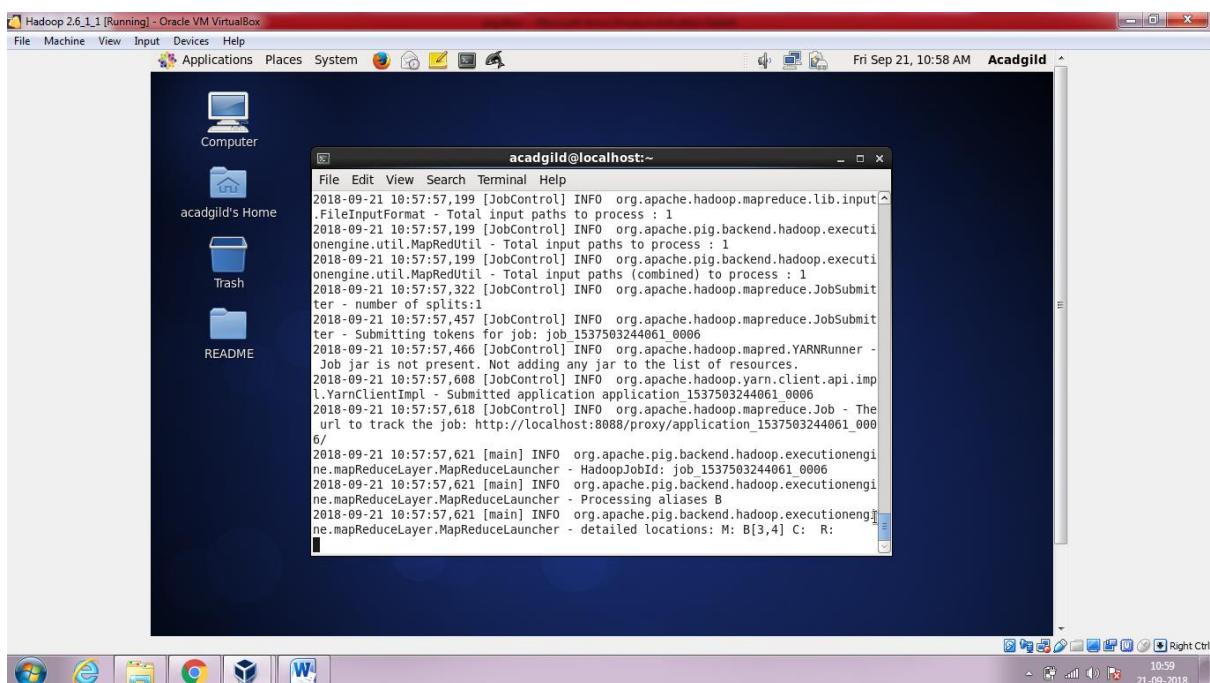
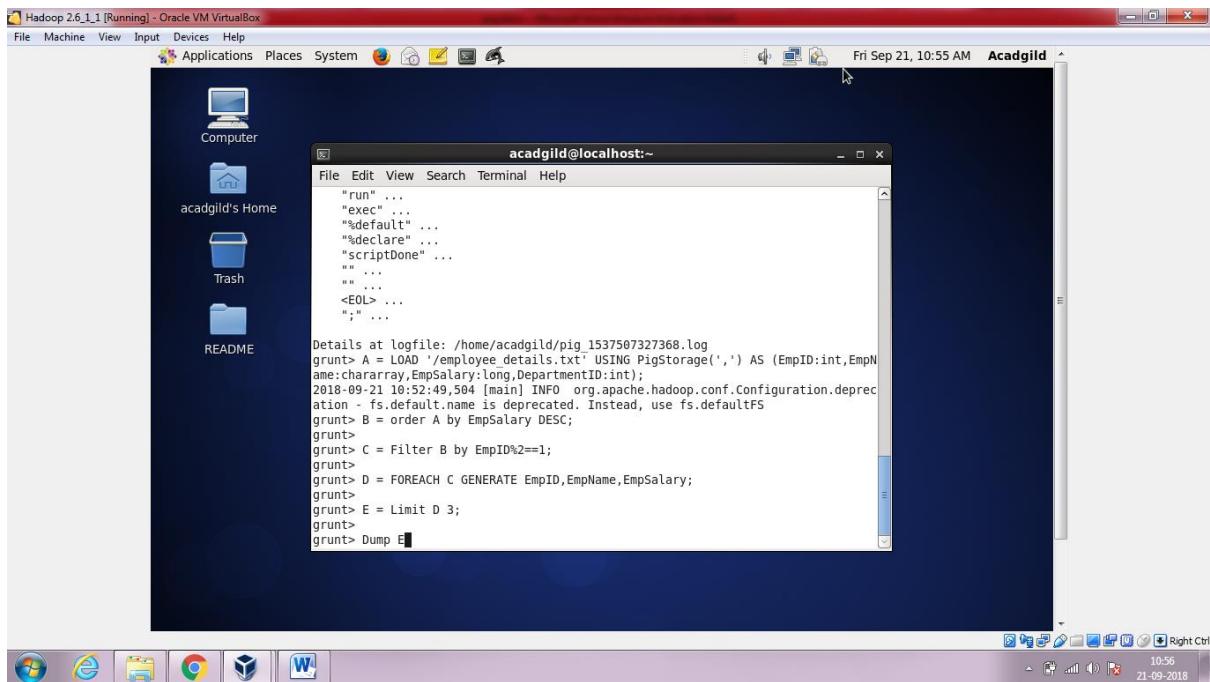
Trash

README

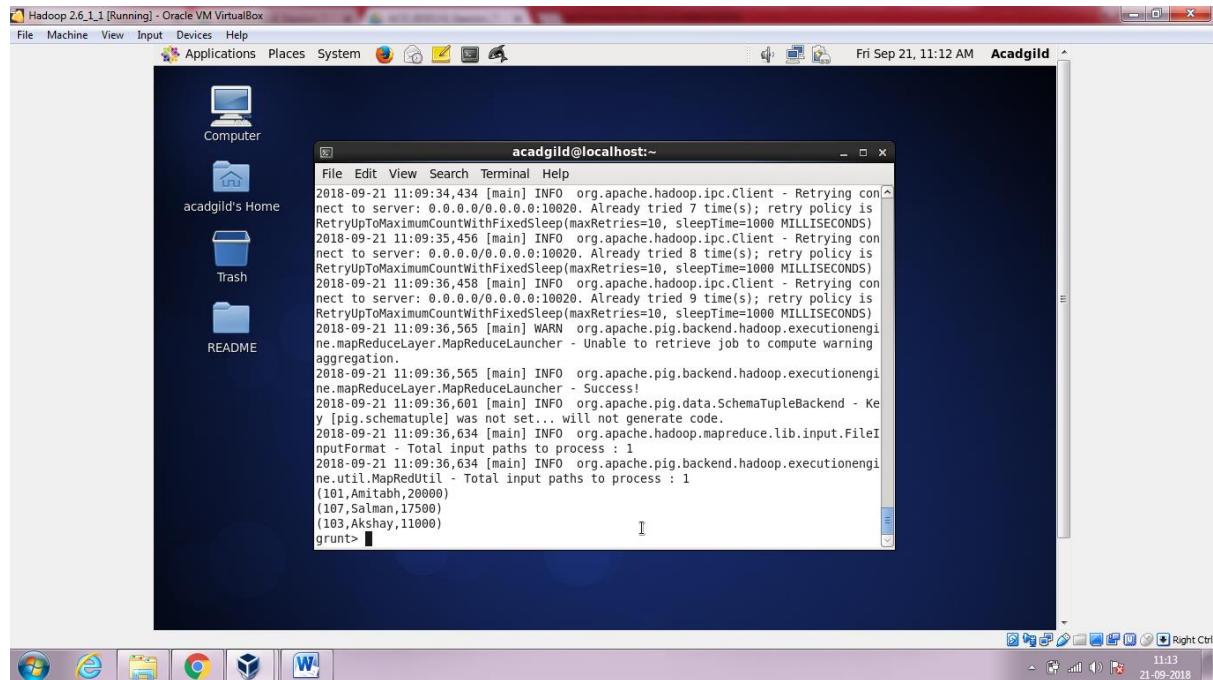
acadgild@localhost:~

```
File Edit View Search Terminal Help
2018-09-21 10:31:05,005 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 10:31:06,014 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 10:31:06,120 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-21 10:31:06,120 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-21 10:31:06,198 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-21 10:31:06,230 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-21 10:31:06,230 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(106,Aamir)
(101,Anitabh)
(113,Jubeen)
(111,Tushar)
(112,Ajay)
grunt> grunt>
```

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



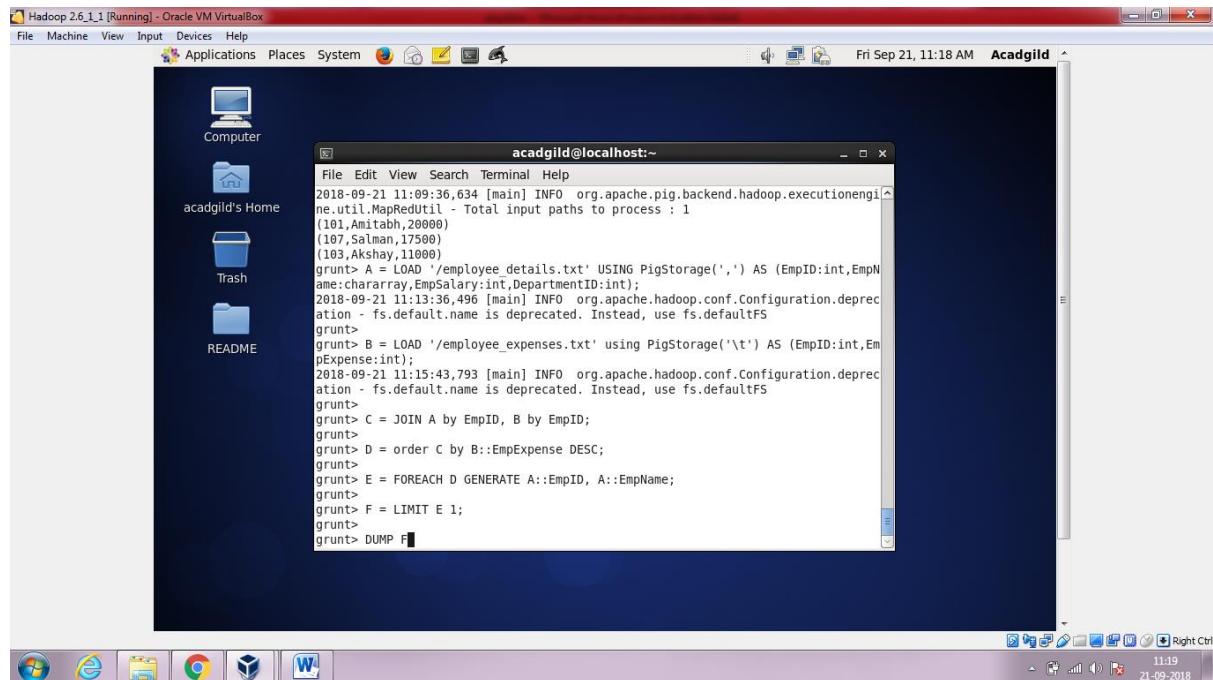
Output:



Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox

```
File Edit View Input Devices Help
Applications Places System Fri Sep 21, 11:12 AM Acadgild
Computer
acadgild's Home
Trash
README
File Edit View Search Terminal Help
2018-09-21 11:09:34,434 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 11:09:35,456 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 11:09:36,458 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 11:09:36,563 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job compute warning aggregation.
2018-09-21 11:09:36,563 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-21 11:09:36,601 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-21 11:09:36,634 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-21 11:09:36,634 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh,20000)
(107,Salman,17500)
(103,Akshay,11000)
grunt> 
```

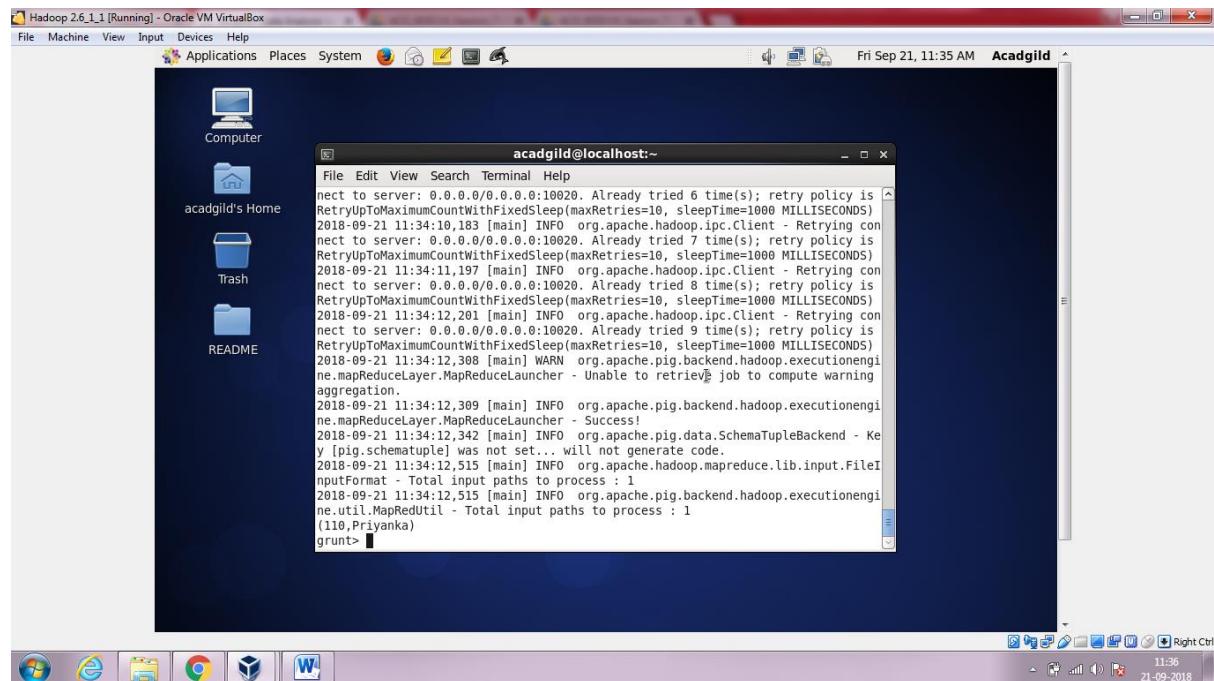
c) Employee(employee id and employee name) with maximum expense (In case two employee have same expense,employee with name coming first in dictionary should get preference)



Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox

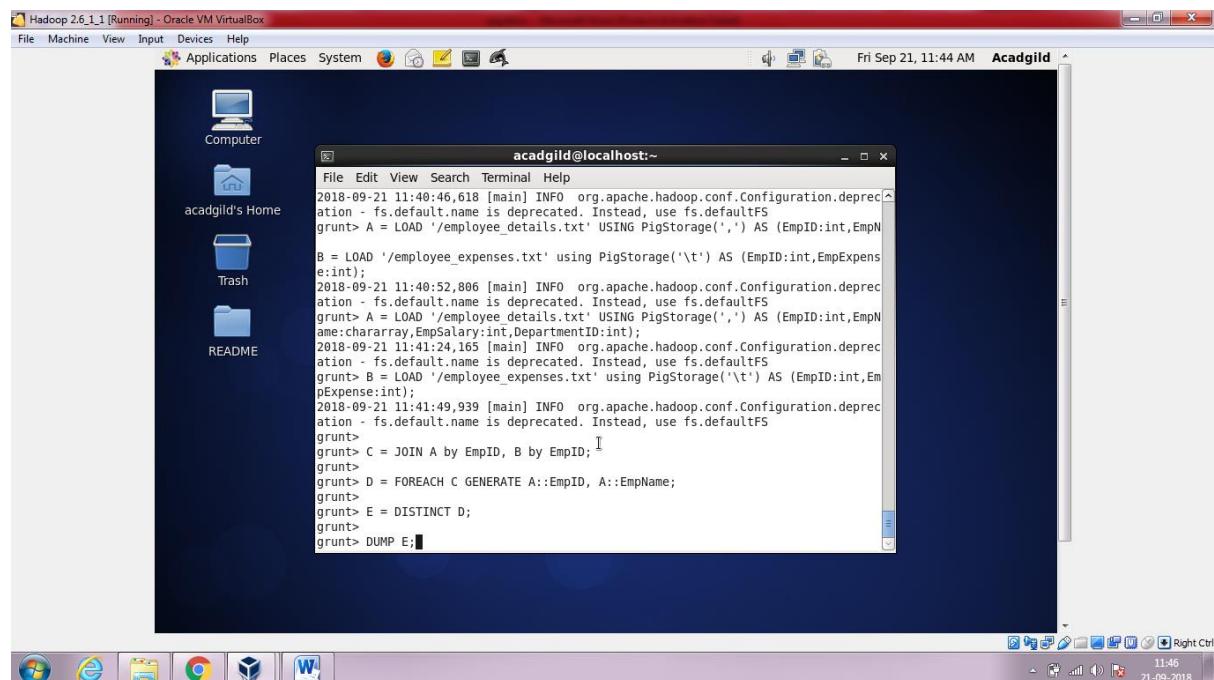
```
File Edit View Input Devices Help
Applications Places System Fri Sep 21, 11:18 AM Acadgild
Computer
acadgild's Home
Trash
README
File Edit View Search Terminal Help
2018-09-21 11:09:36,634 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh,20000)
(107,Salman,17500)
(103,Akshay,11000)
grunt> A = LOAD '/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpName:chararray,EmpSalary:int,DepartmentID:int);
2018-09-21 11:13:36,496 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - Ts.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> B = LOAD '/employee_expenses.txt' using PigStorage('\t') AS (EmpID:int,EmpExpense:int);
2018-09-21 11:15:43,793 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
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
11:19 21-09-2018 Right Ctrl
```

Output:



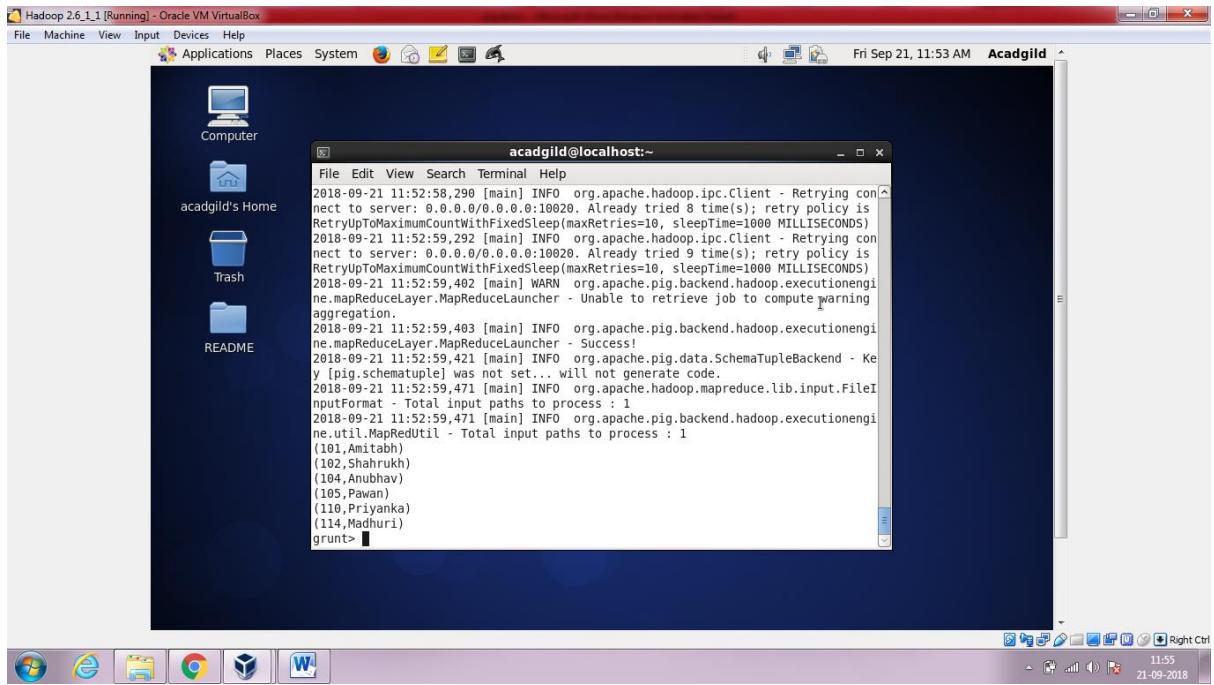
```
File Edit View Search Terminal Help
nect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-21 11:34:10,183 [main] INFO org.apache.hadoop.ipc.Client - Retrying con
nect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-21 11:34:11,197 [main] INFO org.apache.hadoop.ipc.Client - Retrying con
nect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-21 11:34:12,201 [main] INFO org.apache.hadoop.ipc.Client - Retrying con
nect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-21 11:34:12,309 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning
aggregation.
2018-09-21 11:34:12,309 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-09-21 11:34:12,341 [main] INFO org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-09-21 11:34:12,515 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2018-09-21 11:34:12,515 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(110, Priyanka)
grunt> 
```

d) List of employees (employee id and employee name) having entries i9n employee_expenses file.



```
File Edit View Search Terminal Help
2018-09-21 11:40:46,618 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = LOAD '/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpN
ame:chararray);
2018-09-21 11:40:52,806 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = LOAD '/employee_expenses.txt' using PigStorage('\t') AS (EmpID:int,Em
pExpense:int);
2018-09-21 11:40:52,806 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> C = JOIN A by EmpID, B by EmpID;
2018-09-21 11:41:24,165 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> D = FOREACH C GENERATE A::EmpID, A::EmpName;
2018-09-21 11:41:49,939 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> E = DISTINCT D;
2018-09-21 11:41:49,939 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DUMP E; 
```

Output:



```
File Edit View Search Terminal Help
Fri Sep 21, 11:53 AM Acadgild
acadgild@localhost:~
```

2018-09-21 11:52:58,290 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10620. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 11:52:59,292 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10620. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-21 11:52:59,402 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-21 11:52:59,403 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-21 11:52:59,421 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-21 11:52:59,471 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-21 11:52:59,471 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Anitabh)
(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.

```

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Fri Sep 21, 11:58 AM Acadgild
Computer
acadgild's Home
Trash
README
File Edit View Search Terminal Help
acadgild@localhost:~
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt> B = LOAD '/employee_details.txt' using PigStorage('\t') AS (EmpID:int,Em
A = LOAD '/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpName:cha
array,EmpSalary:int,DepartmentID:int);
2018-09-21 11:54:25,798 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = LOAD '/employee_details.txt' USING PigStorage(',') AS (EmpID:int,EmpN
B = LOAD '/employee_expenses.txt' using PigStorage('\t') AS (EmpID:int,EmpExpens
e:int);
2018-09-21 11:54:31,977 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
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;

```

Output:

```

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Fri Sep 21, 12:06 PM Acadgild
Computer
acadgild's Home
Trash
README
File Edit View Search Terminal Help
acadgild@localhost:~
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-21 12:05:11,573 [main] INFO org.apache.hadoop.ipc.Client - Retrying con
nect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2018-09-21 12:05:11,676 [main] WARN org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning
aggregation.
2018-09-21 12:05:11,676 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-09-21 12:05:11,684 [main] INFO org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-09-21 12:05:11,704 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2018-09-21 12:05:11,704 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubleen)
grunt>

```

Task3:Implementing the use case aviation-data-analysis-using-apache-pig.

Problem 1:Top 5 most viewed destinations.

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox

```

File Machine View Input Devices Help
2018-09-25 16:15:24,729 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020
2018-09-25 16:15:33,298 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG default-859d2307-e5eb-40af-bd21-a1997efc1cae
2018-09-25 16:15:33,299 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> REGISTER '/home/acadgild/Downloads/piggybank.jar';
2018-09-25 16:15:49,733 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-09-25 16:16:12,609 [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 '/home/acadgild/Downloads/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-09-25 16:17:44,480 [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:

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox

```

File Machine View Input Devices Help
File Edit View Search Terminal Help
File Open Save Undo Redo Cut Copy Paste Select All Delete All Find Replace
*pig.txt (~) - gedit
File Edit View Search Terminal Help
*pig.txt X acadgild@localhost:-
REGISTER '/home/acadgild/Downloads/piggybank.jar'
A = load '/home/acadgild/Downloads/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
C = filter B by dest is not null;
D = group C by dest;
E = foreach D generate group, COUNT(C.dest);
F = order E by $1 DESC;
Result = LIMIT F 5;
A1 = load '/home/acadgild/Downloads/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
dump joined_table;
A = load '/DelayedFlightData.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
Result = LIMIT F 5;
A1 = load '/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;

```

Problem 2: Which month has seen the most number of cancellations due to whether.

```

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Wed Sep 26, 4:08 PM Acadgild
Computer
acadgild's Home
Trash
README
File Edit View Search Terminal Help
2018-09-26 15:42:29,998 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-26 15:42:30,051 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-26 15:42:30,051 [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> REGISTER '/home/acadgild/Downloads/jar_files/piggybank-0.17.0.jar';
grunt> A = load '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(' ','NO_MULTILINE','UNIX','SKIP INPUT HEADER');
2018-09-26 16:07:44,584 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt> C = filter B by cancelled == 1 AND cancel_code == 'B';
grunt> D = group C by month;
grunt> E = foreach D generate group, COUNT(C.cancelled);
grunt> F= order E by $1 DESC;
grunt> Result = limit F 1;
grunt> dump Result;

```

Output:

```

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Wed Sep 26, 4:36 PM Acadgild
Computer
acadgild's Home
Trash
README
File Edit View Search Terminal Help
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-26 16:35:48,703 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-26 16:35:49,704 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-26 16:35:49,706 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-26 16:35:50,819 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-26 16:35:50,819 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-26 16:35:50,857 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-26 16:35:50,911 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-26 16:35:50,911 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt> 

```

Problem 3:Top ten origins with the highest AVG departure delay.

Hadoop 2.6.1 [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Wed Sep 26, 4:40 PM Acadgild

Computer

acadgild's Home

Trash

README

acadgild@localhost:~

```
File Edit View Search Terminal Help
2018-09-26 16:35:50,911 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt> REGISTER '/home/acadgild/Downloads/jar_files/piggybank-0.17.0.jar';
grunt> A = load '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');
2018-09-26 16:38:29,234 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
grunt> C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
grunt> D1 = group C1 by origin;
grunt> E1 = foreach D1 generate group, AVG(C1.dep_delay);
grunt> Result = order E1 by $1 DESC;
grunt> Top_ten = limit Result 10;
grunt> Lookup = load '/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');
2018-09-26 16:39:32,155 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, 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, Top_ten by $0;
grunt> Final = foreach Joined generate $0,$1,$2,$4;
grunt> Final Result = ORDER Final by $3 DESC;
grunt> dump Final_Result;
```

16:40
26-09-2018

Output:

Hadoop 2.6.1 [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Wed Sep 26, 5:18 PM Acadgild

Computer

acadgild's Home

Trash

README

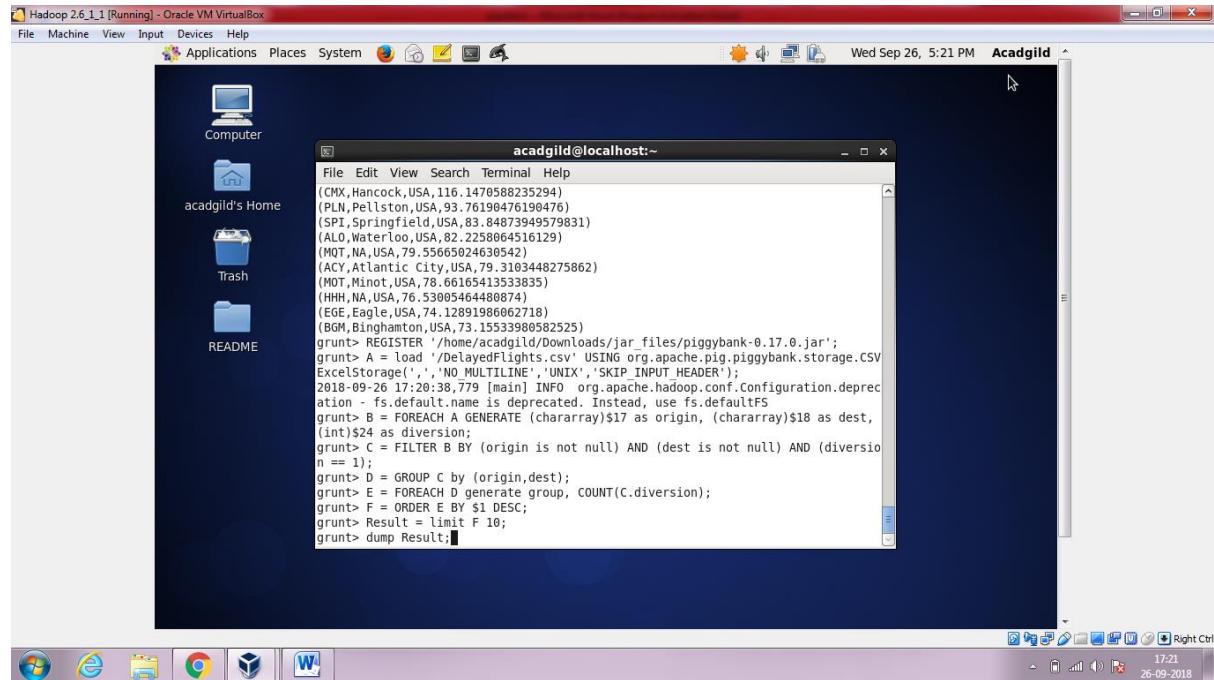
acadgild@localhost:~

```
File Edit View Search Terminal Help
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISSECONDS)
2018-09-26 17:18:32,965 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-26 17:18:32,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-26 17:18:33,004 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig schematuple] was not set... will not generate code.
2018-09-26 17:18:33,061 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-26 17:18:33,062 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(PLN,Pelston,USA,93.76199476199476)
(ALO,Waterloo,USA,82.2258064516129)
(MOT,NA,USA,79.55665024630542)
(ACY,Atlantic_City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.53005464488074)
(EGE,Eagle,USA,74.12891986062718)
(BGM,Binghamton,USA,73.15533980582525)
grunt>
```

17:18
26-09-2018

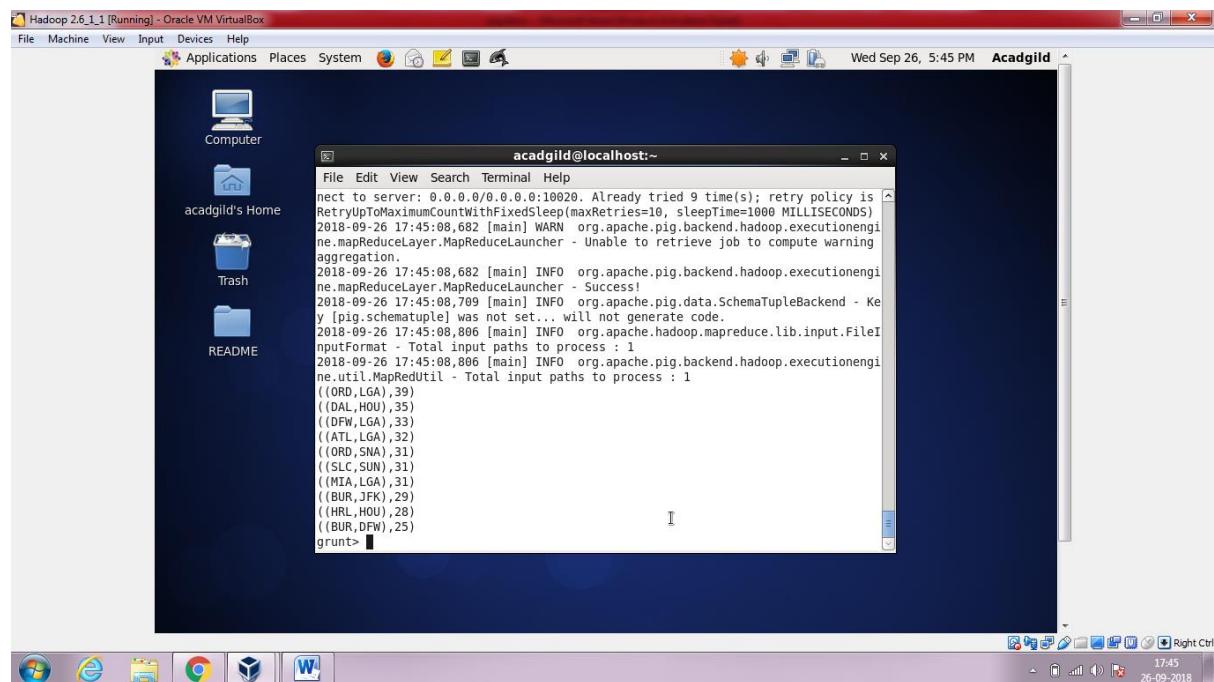
Problem 4:

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



```
File Edit View Search Terminal Help
(acMX,Hancock,USA,116.1470588235294)
(PLN,Pelston,USA,93.7619047619047)
(SPI,Springfield,USA,83.84873949579831)
(ALO,Waterloo,USA,82.2258664516129)
(MOT,NA,USA,79.55665024630542)
(ACY,Atlantic City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.538054644880874)
(EGE,Eagle,USA,74.12891986062718)
(BGM,Binghamton,USA,73.15533980582525)
grunt> REGISTER '/home/acadgild/Downloads/jar_files/piggybank-0.17.0.jar';
grunt> A = load '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','');
grunt> B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest,
      (int)$24 as diversion;
grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
grunt> D = GROUP C by (origin,dest);
grunt> E = FOREACH D generate group, COUNT(C.diversion);
grunt> F = ORDER E BY $1 DESC;
grunt> Result = limit F 10;
grunt> dump Result;
```

Output:



```
File Edit View Search Terminal Help
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISCONDS)
2018-09-26 17:45:08,682 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2018-09-26 17:45:08,682 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-26 17:45:08,709 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-09-26 17:45:08,806 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-26 17:45:08,806 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process :
((ORD,LGA),39)
((DAL,HOU),35)
((DFW,LGA),33)
((ATL,LGA),32)
((ORD,SNA),31)
((SLC,SUN),31)
((MIA,LGA),31)
((BUR,JFK),29)
((HNL,HOU),28)
((BUR,DFW),25)
grunt>
```

