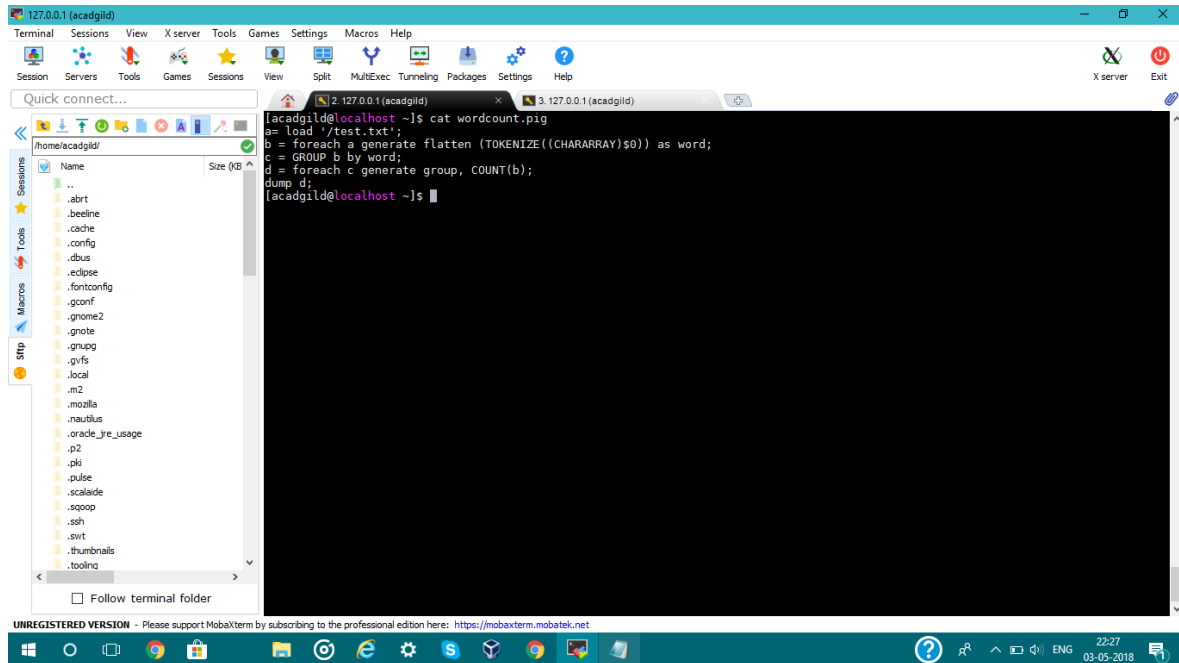


Assignment 7

Task1: Write a program to implement wordcount using Pig

Pig commands:



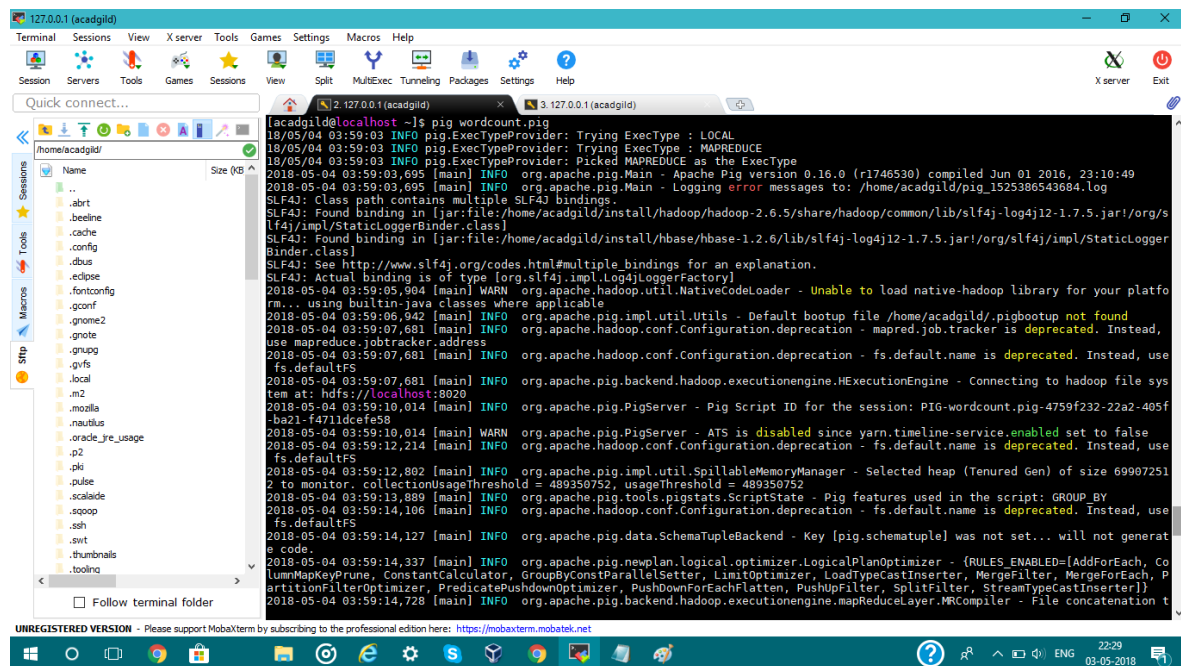
The screenshot shows a MobaXterm terminal window with a file explorer on the left. The terminal displays the following Pig script and its execution:

```
[acadgild@localhost ~]$ cat wordcount.pig
a= load '/test.txt';
b = foreach a generate flatten (TOKENIZE((CHARARRAY)$0)) as word;
c = GROUP b by word;
d = foreach c generate group, COUNT(b);
dump d;
[acadgild@localhost ~]$
```

The file explorer on the left shows the contents of the home directory, including various configuration files and folders.

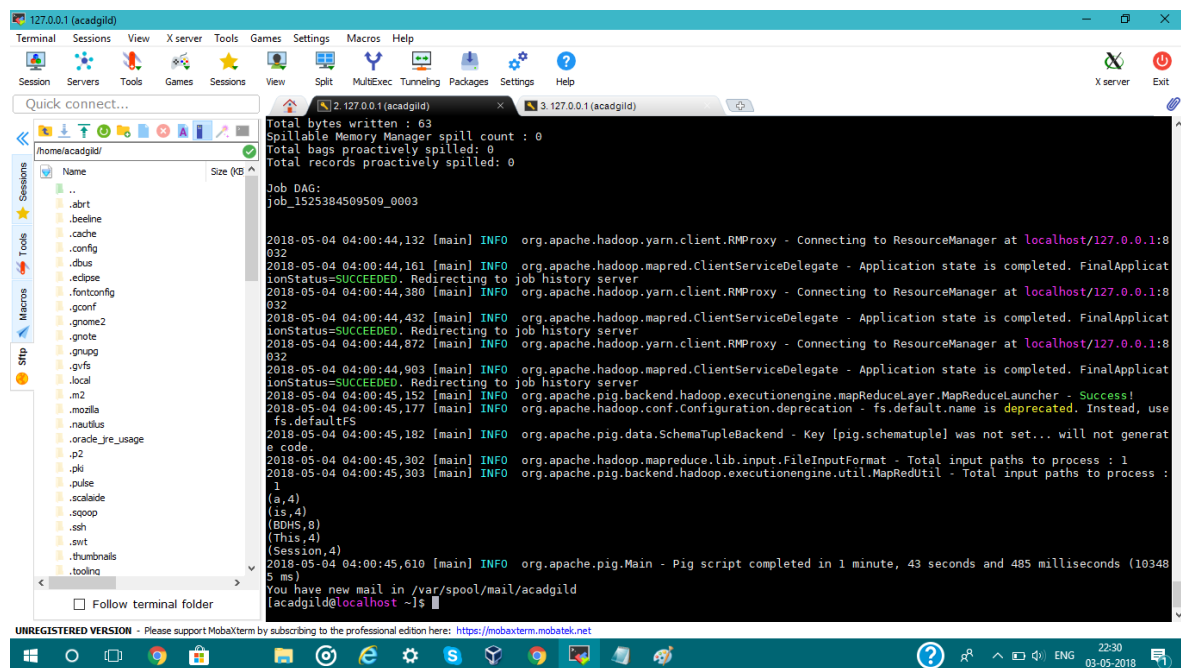
- a is loading the test.txt file, for which the count of the occurrence of the words is to be calculated.
- b is used to split every line of the text file into words.
- C is used to group same words together.
- D is used to calculate the count of words.

Command to execute Pig script: pig wordcount.pig



The screenshot shows a MobaXterm window with a terminal session. The command `127.0.0.1 (acadgild) ~$ pig wordcount.pig` has been executed. The output is a log of Pig and Hadoop operations. Key messages include: 'Picked MAPREDUCE as the ExecType', 'Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49', 'Logging error messages to: /home/acadgild/pig_1525386543084.log', 'SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]', 'SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.', 'SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]', 'Unable to load native-hadoop library for your platform... using builtin-java classes where applicable', 'org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use mapreduce.job.tracker.address', 'org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS', 'org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020', 'org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordcount.pig-4759f232-22a2-405f-ba21-f471ldcfe58', 'org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false', 'org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS', 'org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752', 'org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY', 'org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS', 'org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.', 'org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, StreamTypeCastInserter]}', 'org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - File concatenation t...

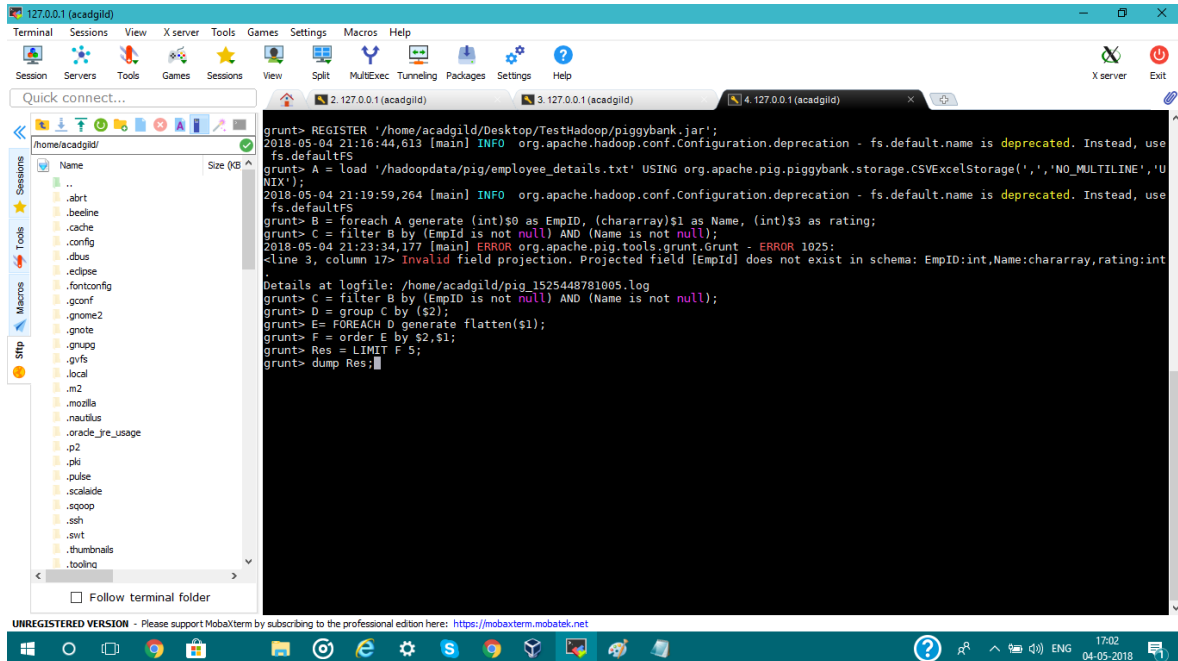
Output



The screenshot shows the continuation of the MobaXterm terminal session. The output displays the results of the Pig script execution. Key messages include: 'Total bytes written : 63', 'Spillable Memory Manager spill count : 0', 'Total bags proactively spilled: 0', 'Total records proactively spilled: 0', 'Job DAG: job_1525384509509_0003', 'org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032', 'org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server', 'org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032', 'org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server', 'org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032', 'org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server', 'org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!', 'org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS', 'org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.', 'org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1', 'org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1', 'org.apache.pig.Main - Pig script completed in 1 minute, 43 seconds and 485 milliseconds (103485 ms)', 'You have new mail in /var/spool/mail/acadgild', 'acadgild@localhost ~\$

Task2

- a. Display top 5 employees (employee_id and employee_name) with highest rating



```
grunt> REGISTER '/home/acadgild/Desktop/TestHadoop/piggybank.jar';
2018-05-04 21:16:44,613 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = load '/hadoopdata/pig/employee_details.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX');
2018-05-04 21:19:59,264 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$0 as EmpID, (chararray)$1 as Name, (int)$3 as rating;
grunt> C = filter B by (EmpID is not null) AND (Name is not null);
2018-05-04 21:23:34,177 [main] ERROR org.apache.pig.tools.grunt.grunt - ERROR 1025:
<line 3, column 17> Invalid field projection. Projected field [EmpID] does not exist in schema: EmpID:int,Name:chararray,rating:int
Details at logfile: /home/acadgild/pig_1525448781005.log
grunt> C = filter B by (EmpID is not null) AND (Name is not null);
grunt> D = group C by ($2);
grunt> E= FOREACH D generate flatten($1);
grunt> F = order E by $2,$1;
grunt> Res = LIMIT F 5;
grunt> dump Res;
```

- In first step, we need to register the jar file (piggybank.jar) for executing pig commands.
- Then we are loading the employee_details.txt file.
- In B, we are displaying the EmpID, Name and rating of the employees, by specifying the column position.
- In C, we are filtering out those columns in which the EmpID and the Name is null.
- In D, we are grouping the output of C by EmpID.
- In E, we are un-nesting the second column obtained from D.
- In F, we are sorting E by third column (rating) and then by second column (name).
- In Res, we are limiting the result of F by 5 rows.

Output

```

2018-05-04 22:37:10,566 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-04 22:37:10,618 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 22:37:10,817 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-04 22:37:10,832 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 22:37:10,986 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-04 22:37:11,084 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 22:37:11,172 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-04 22:37:11,191 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 22:37:11,349 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-04 22:37:11,366 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 22:37:11,712 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-04 22:37:11,731 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-04 22:37:11,738 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-05-04 22:37:11,760 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-04 22:37:11,761 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(106,Aamir,1)
(101,Amitabh,1)
(113,Jubeen,1)
(111,Tushar,1)
(112,Ajay,2)
grunt>

```

b. Display top 3 employees with highest salary, with odd employee_id

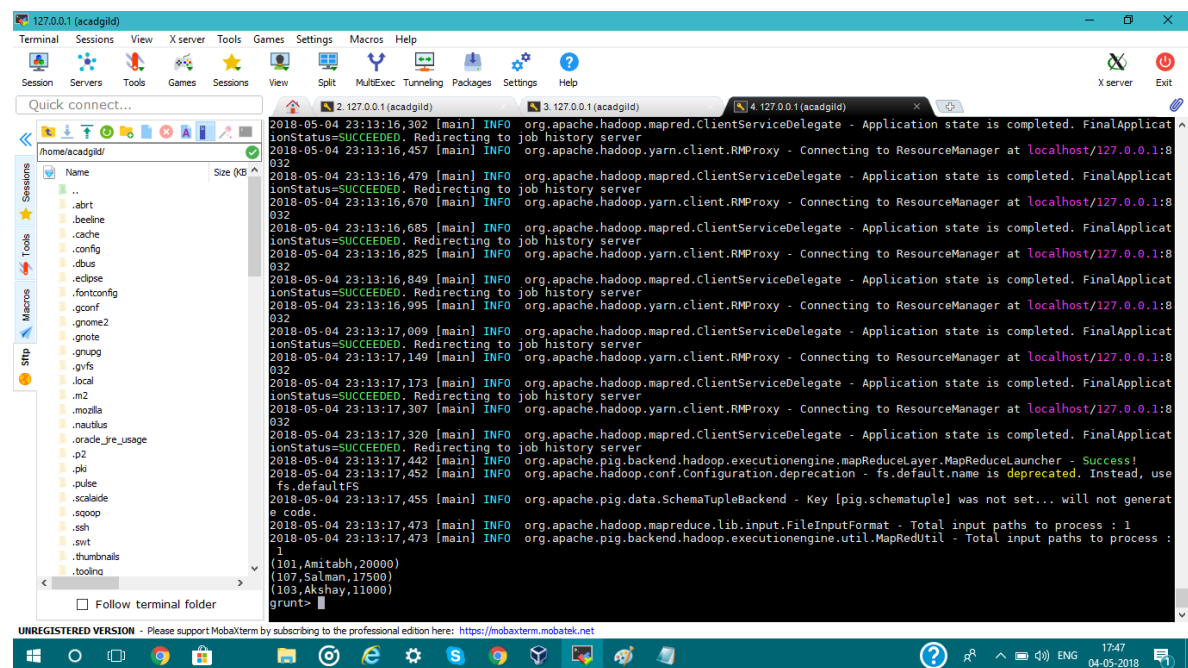
```

grunt> REGISTER '/home/acadgild/Desktop/TestHadoop/piggybank.jar';
grunt> A = load '/hadoopdata/pig/employee_details.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX');
2018-05-04 22:40:49,128 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$0 as EmpID, (chararray)$1 as Name, (int)$2 as salary;
2018-05-04 22:43:51,710 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 10, column 11> Syntax error, unexpected symbol at or near 'BY'
Details at logfile: /home/acadgild/pig_1525448781005.log
grunt> C = FILTER B BY (EmpID%2)=0;
2018-05-04 23:07:59,239 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> D = order C by $2 DESC, $1;
2018-05-04 23:07:59,256 [main] INFO org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-05-04 23:07:59,278 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-05-04 23:07:59,304 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope=98
2018-05-04 23:07:59,311 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2018-05-04 23:07:59,388 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2018-05-04 23:07:59,400 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-04 23:07:59,408 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-04 23:07:59,414 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2018-05-04 23:07:59,414 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3

```

- The first two steps are the same as that of task a.
- In C, we are filtering those records where Emp_ID is odd.
- In D, we are sorting C on the basis of salary in descending order and then on the basis of emp_id.
- In Res, we are limiting the output of D to top 3 rows.

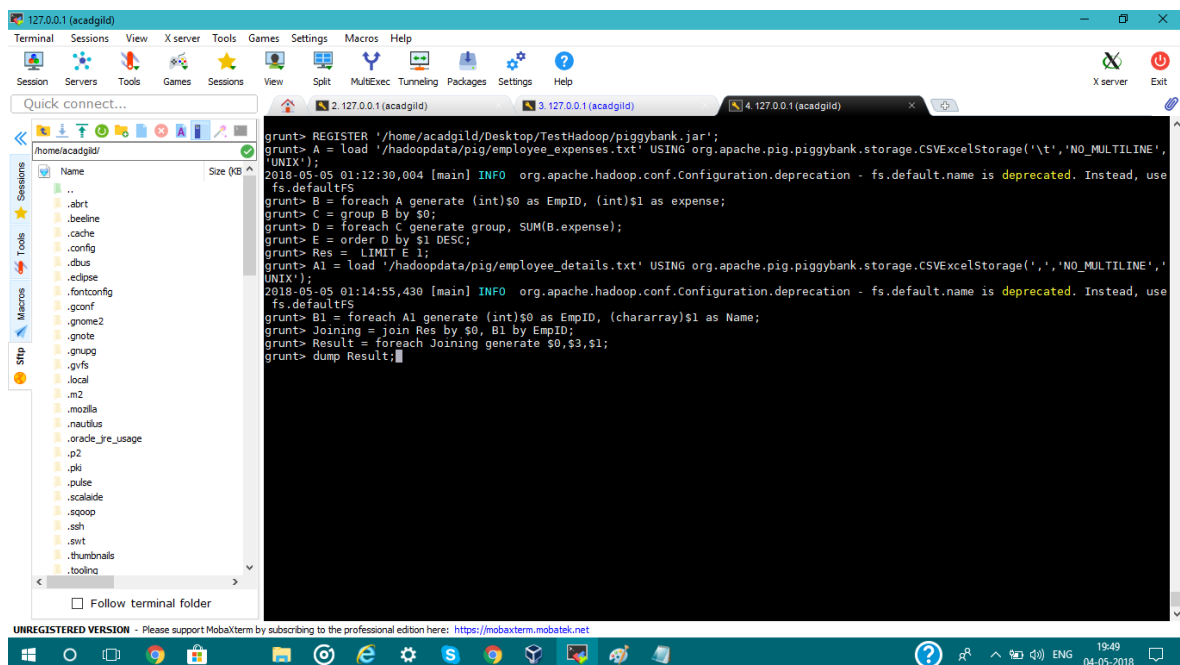
Output



The screenshot shows a MobaXterm terminal window with three tabs. The active tab displays the output of a Hadoop MapReduce job. The output consists of multiple log entries from the main and INFO loggers, showing the application state, redirection to the job history server, and connection to the Resource Manager. The final output is a list of employee IDs and names: (101,Amitabh,20000), (107,Salman,17500), and (103,Akshay,11000). The terminal window also shows a file explorer on the left and a taskbar at the bottom.

```
2018-05-04 23:13:16,302 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:16,457 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 23:13:16,479 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:16,670 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 23:13:16,685 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:16,825 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 23:13:16,849 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:16,995 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 23:13:17,009 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:17,149 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 23:13:17,173 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:17,307 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 23:13:17,320 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 23:13:17,442 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-04 23:13:17,452 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2018-05-04 23:13:17,455 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generat
e code.
2018-05-04 23:13:17,472 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-04 23:13:17,473 [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. Display top 5 employees (employee_id and employee_name) with maximum expense



The screenshot shows a MobaXterm terminal window with three tabs. The active tab displays the output of a Pig Latin script. The script registers a jar file, loads a CSV file of employee expenses, and performs a series of operations: generating a list of employee IDs and expenses, grouping by employee ID, and then joining with a list of employee details. The final output is a list of employee IDs and names: (101,Amitabh,20000), (107,Salman,17500), and (103,Akshay,11000). The terminal window also shows a file explorer on the left and a taskbar at the bottom.

```
grunt> REGISTER '/home/acadgild/Desktop/TestHadoop/piggybank.jar';
grunt> A = load '/hadoopdata/pig/employee_expenses.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage('t','NO_MULTILINE',
'UNIX');
2018-05-05 01:12:30,004 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
grunt> B = foreach A generate (int)$0 as EmpID, (int)$1 as expense;
grunt> C = group B by $0;
grunt> D = foreach C generate group, SUM(B.expense);
grunt> E = order D by $1 DESC;
grunt> Res = LIMIT E 1;
grunt> A1 = load '/hadoopdata/pig/employee_details.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage('t','NO_MULTILINE','
UNIX');
2018-05-05 01:14:55,430 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
grunt> B1 = foreach A1 generate (int)$0 as EmpID, (chararray)$1 as Name;
grunt> Joining = join Res by $0, B1 by EmpID;
grunt> Result = foreach Joining generate $0,$3,$1;
grunt> dump Result;
```

- In D, we are calculating the sum of expenses grouped by employee_id.

- In Joining, we are joining the two relations on the basis of employee id and then for final result, we are displaying the employee_id, employee name and his expense.

Output

```

2018-05-05 01:19:05,123 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-05 01:19:05,137 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-05 01:19:05,240 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-05 01:19:05,248 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-05 01:19:05,366 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-05 01:19:05,380 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-05 01:19:05,612 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-05 01:19:05,637 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-05 01:19:05,808 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-05 01:19:05,817 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-05 01:19:05,894 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-05-05 01:19:05,909 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-05 01:19:06,106 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-05 01:19:06,110 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-05 01:19:06,114 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-05-05 01:19:06,131 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-05 01:19:06,131 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(102,Shahrukh,500)
grunt>
  
```

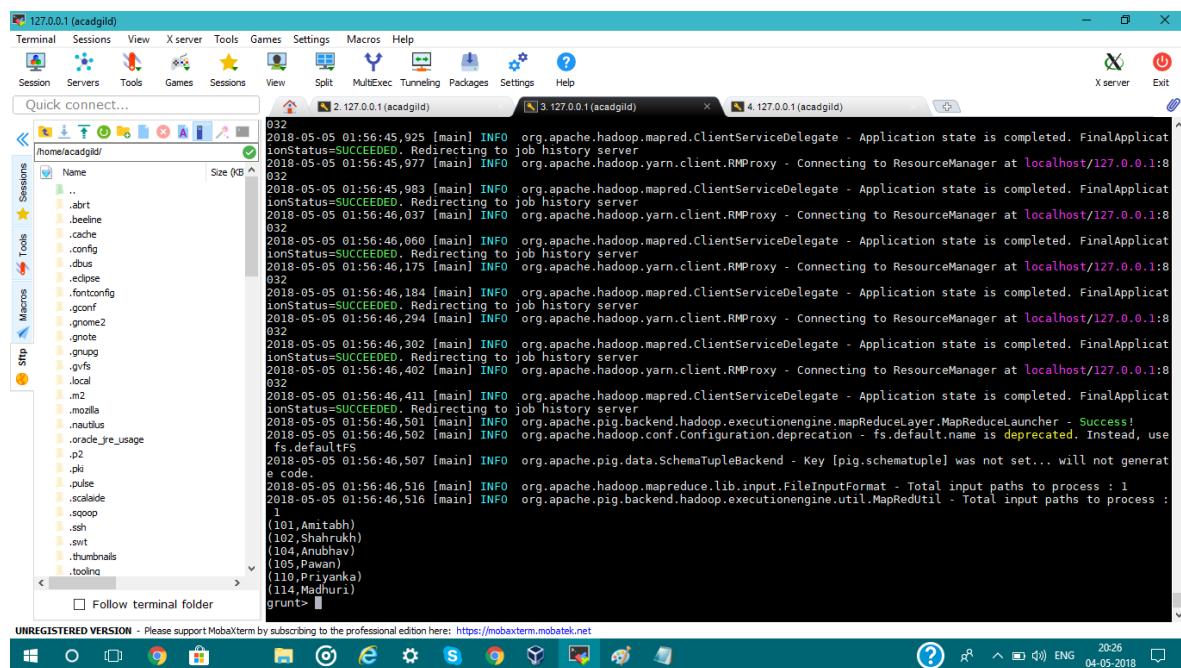
d. List of employees having entries in employee_expenses file

```

grunt> REGISTER '/home/acadgild/Desktop/TestHadoop/piggybank.jar';
grunt> A = load '/hadoopdata/pig/employee_details.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX');
2018-05-05 01:44:26,554 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$0 as EmpID;
grunt> A1 = load '/hadoopdata/pig/employee_expenses.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage('\t', 'NO_MULTILINE', 'UNIX');
2018-05-05 01:45:02,893 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$0 as EmpID, (chararray)$1 as Name;
grunt> B1 = foreach A1 generate (int)$0 as EmpID;
grunt> C = DISTINCT B1;
grunt> joining = join C by $0, B by EmpID;
grunt> Res = foreach joining generate $0,$2;
grunt> dump Res;
  
```

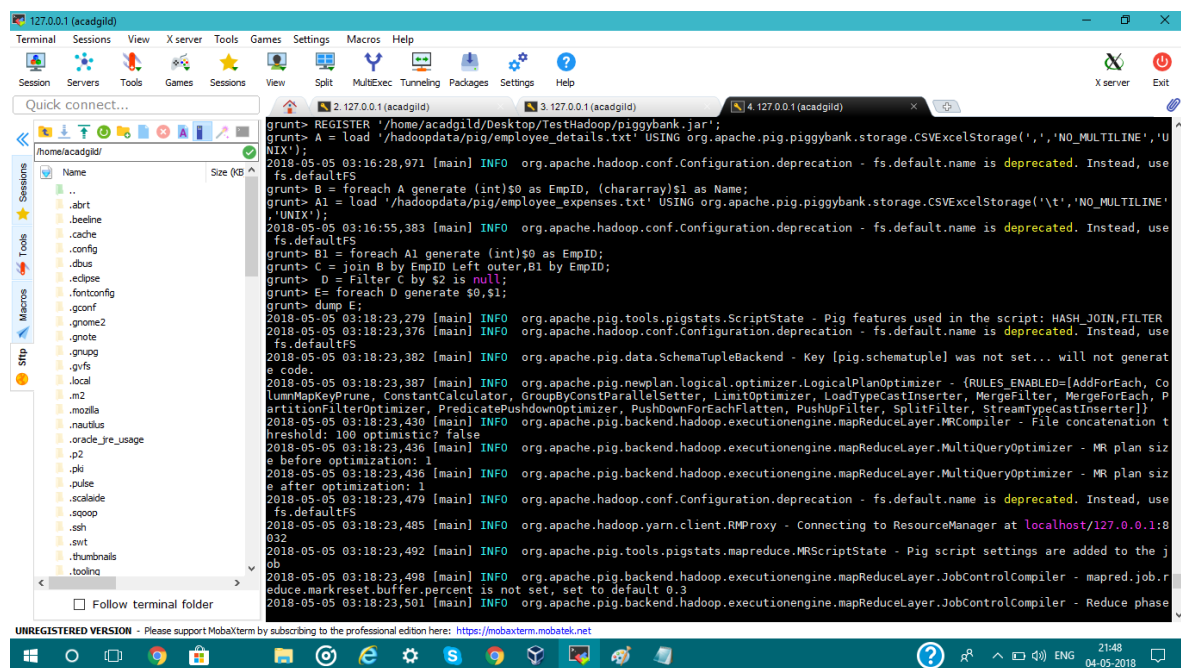
Here, we are joining both the relations on the basis of employee_id and then displaying the result.

Output



```
2018-05-05 01:56:45,925 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to
2018-05-05 01:56:45,977 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-05 01:56:45,983 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to
2018-05-05 01:56:46,037 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-05 01:56:46,060 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to
2018-05-05 01:56:46,175 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-05 01:56:46,184 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to
2018-05-05 01:56:46,294 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-05 01:56:46,302 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to
2018-05-05 01:56:46,402 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-05 01:56:46,411 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to
2018-05-05 01:56:46,501 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-05 01:56:46,502 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2018-05-05 01:56:46,507 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generat
e code.
2018-05-05 01:56:46,516 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-05-05 01:56:46,516 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process :
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Medhuri)
grunt>
```

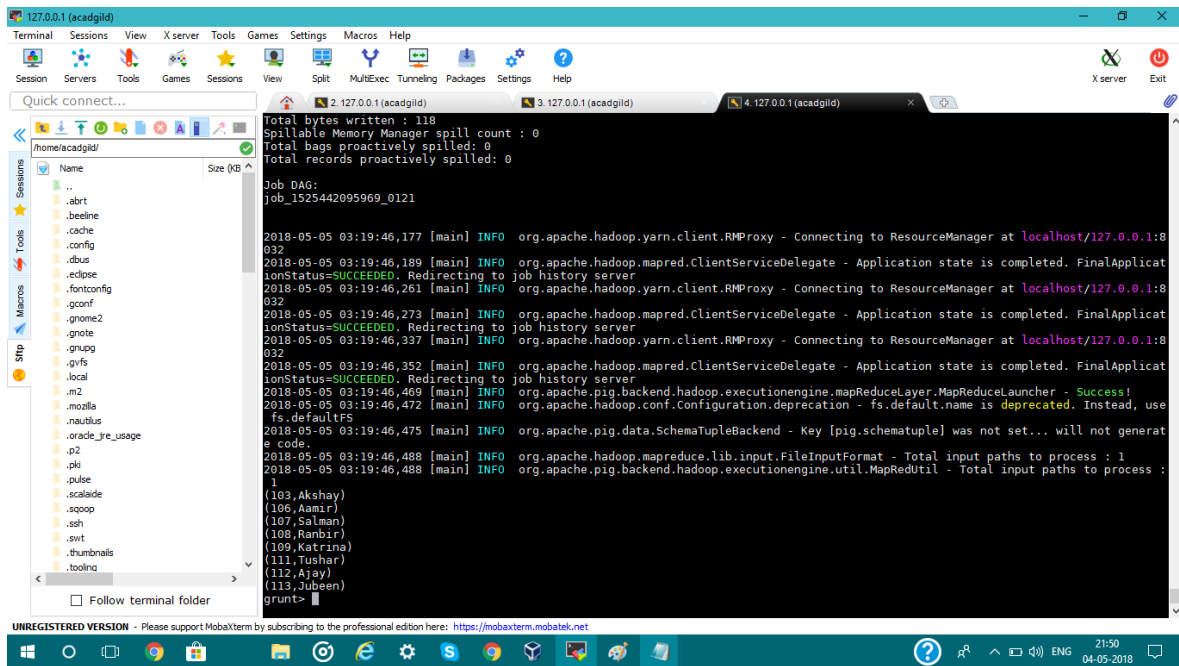
e. List of employees having no entries in employee_expenses file



```
grunt> REGISTER '/home/acadgild/Desktop/TestHadoop/piggybank.jar';
grunt> A = load '/hadoopdata/pig/employee_details.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'U
NIX');
2018-05-05 03:16:28,971 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
grunt> B = foreach A generate (int)$0 as EmpID, (chararray)$1 as Name;
grunt> A1 = load '/hadoopdata/pig/employee_expenses.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage('\t', 'NO_MULTILINE'
, 'UNIX');
2018-05-05 03:16:55,383 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
grunt> B1 = foreach A1 generate (int)$0 as EmpID;
grunt> C = join B by EmpID Left outer, B1 by EmpID;
grunt> D = Filter C by $2 is null;
grunt> E = foreach D generate $0, $1;
grunt> dump E;
2018-05-05 03:18:23,279 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN, FILTER
2018-05-05 03:18:23,376 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2018-05-05 03:18:23,382 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generat
e code.
2018-05-05 03:18:23,387 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Co
lumnMapKeyPrune, ConstantCalculator, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-05-05 03:18:23,430 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation t
hreshold: 100 optimistic: false
2018-05-05 03:18:23,436 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan siz
e before optimization: 1
2018-05-05 03:18:23,436 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan siz
e after optimization: 1
2018-05-05 03:18:23,479 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2018-05-05 03:18:23,485 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-05 03:18:23,492 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the j
ob
2018-05-05 03:18:23,498 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.r
educe.markreset.buffer.percent is not set, set to default 0.3
2018-05-05 03:18:23,501 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase
```

Here, we are performing left outer join on the two relations (employee_details, employee_expenses).

Output

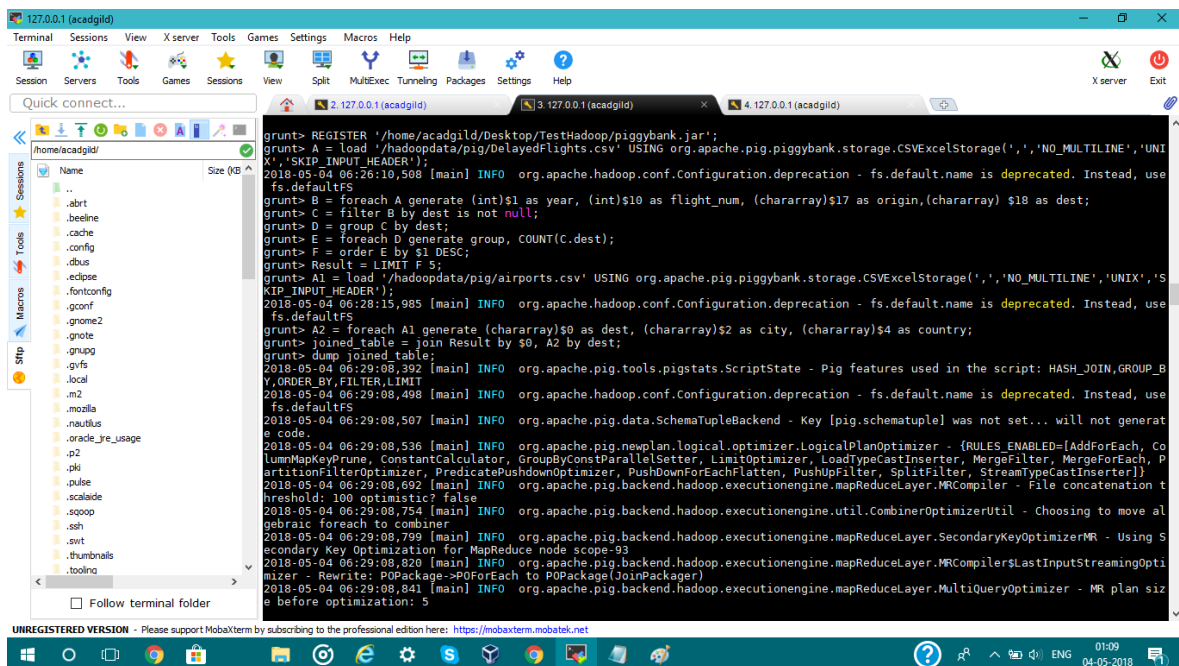


Task 3

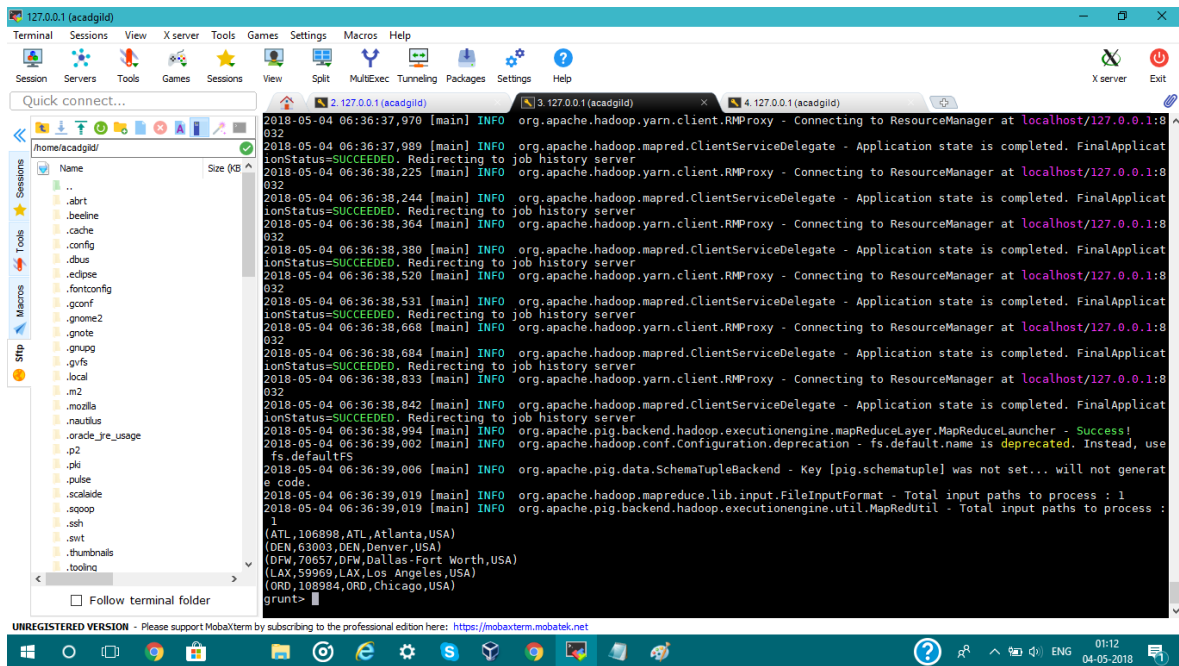
Implement the aviation data analysis use case

- Find out the top 5 most visited destinations.

Pig script

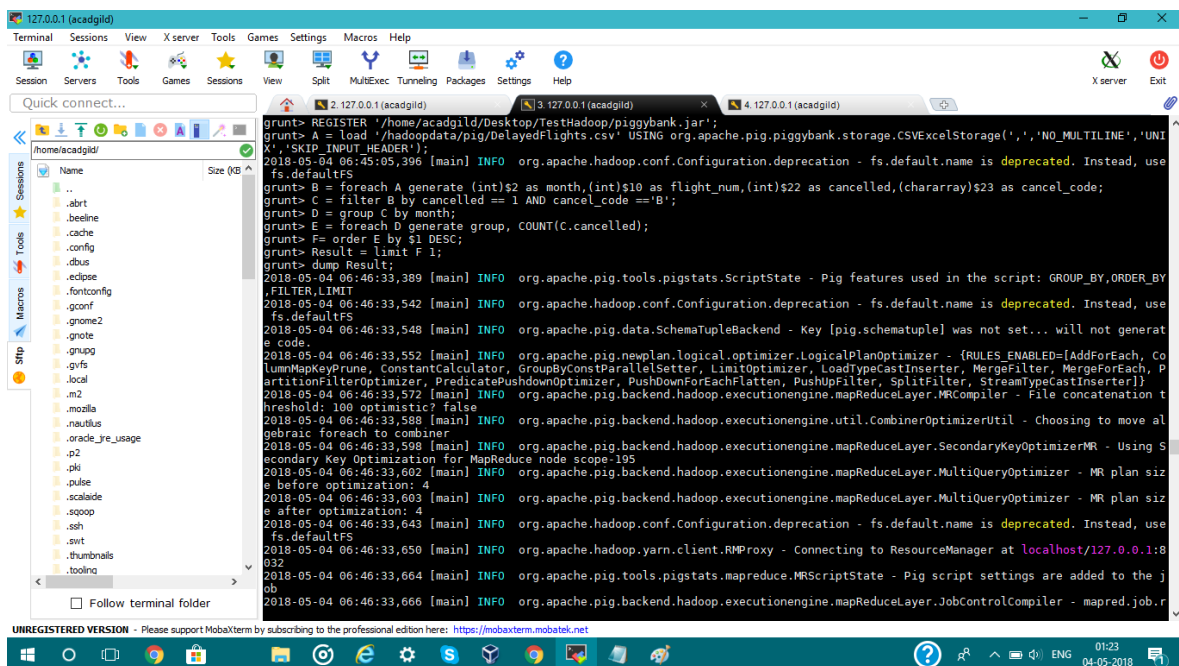


Output

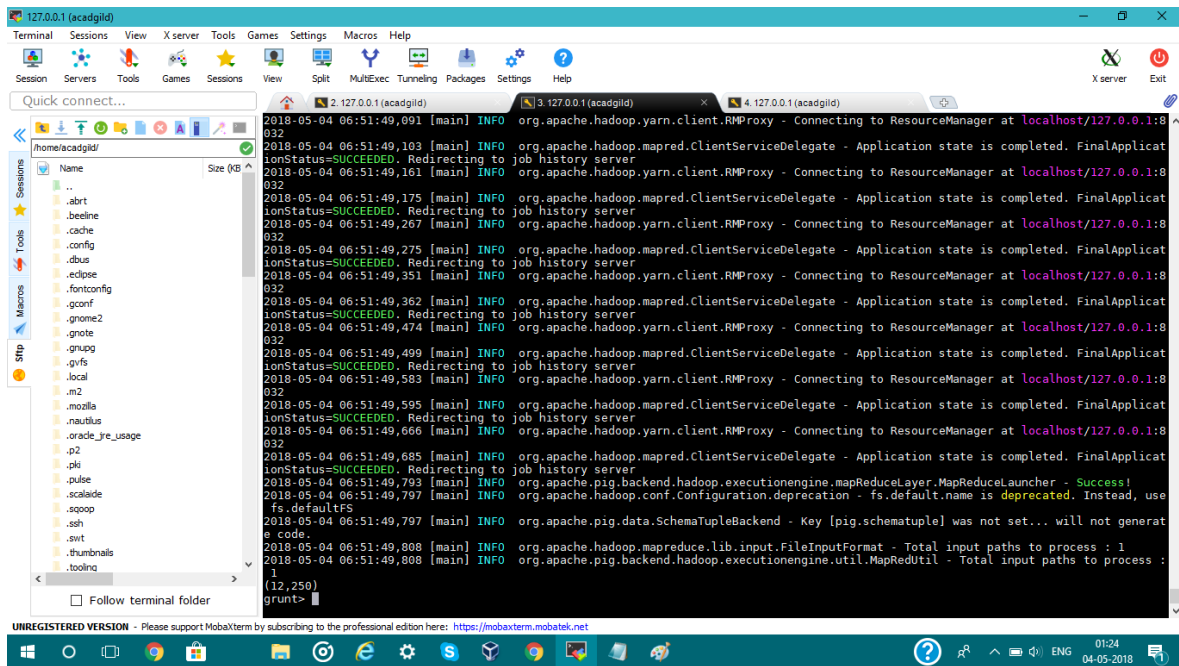


b. Which month has seen the most number of cancellations due to bad weather?

Pig Commands

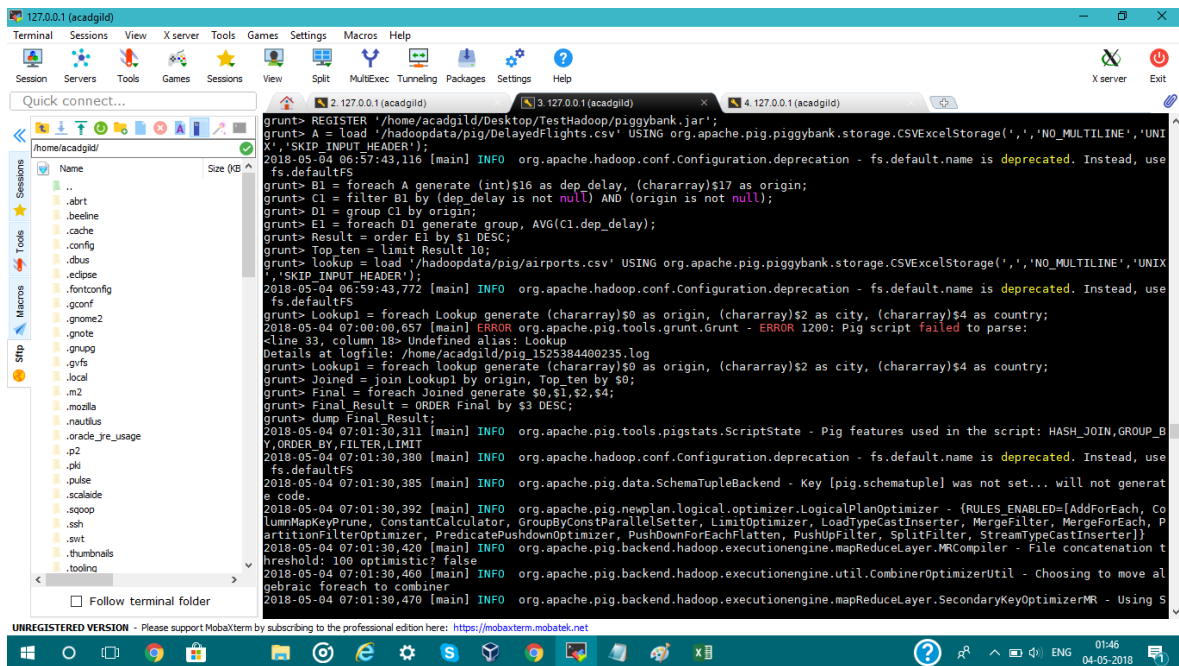


Output

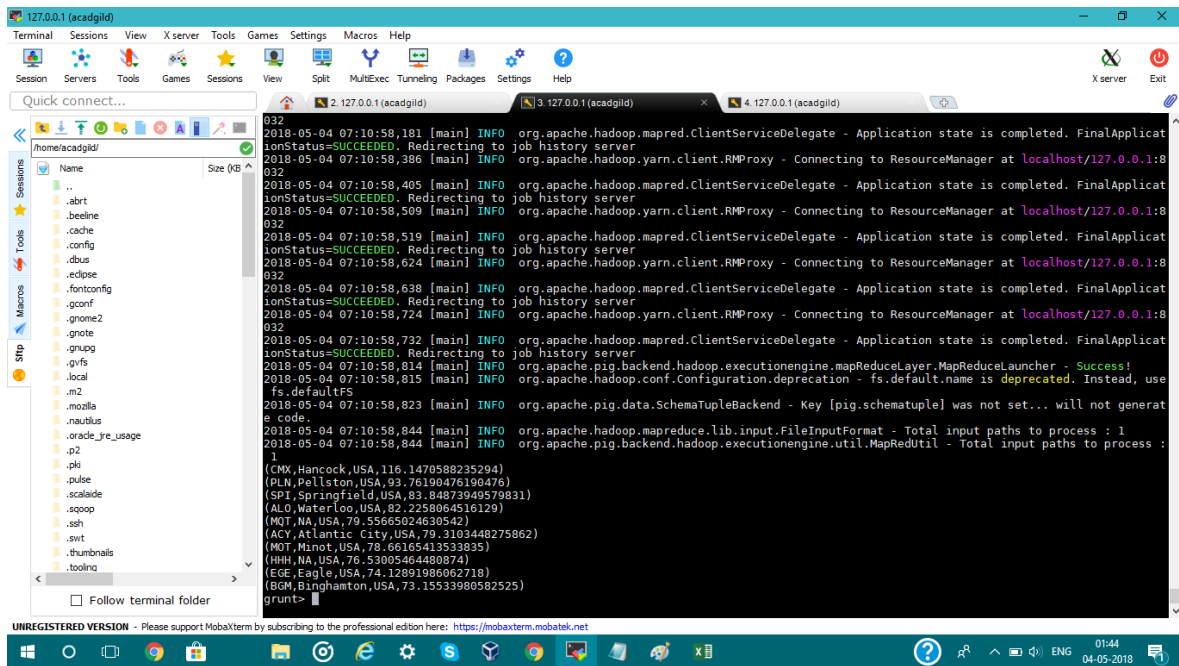


c. Top ten origins with the highest AVG departure delay

Pig commands

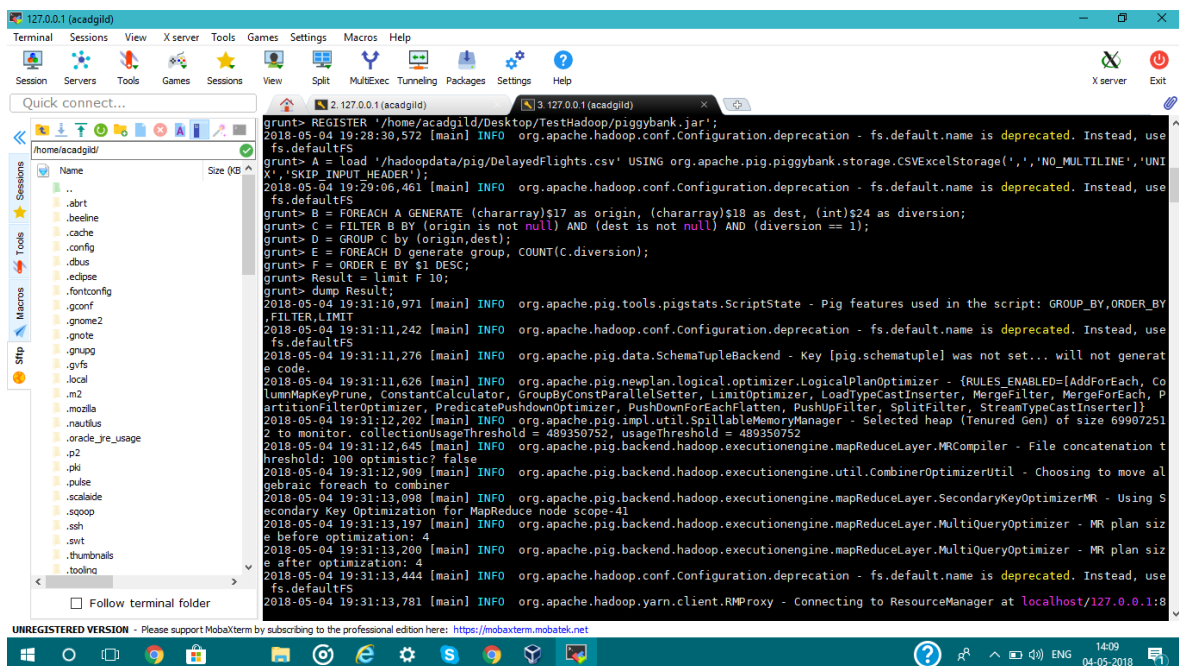


Output



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

Pig commands



Output

