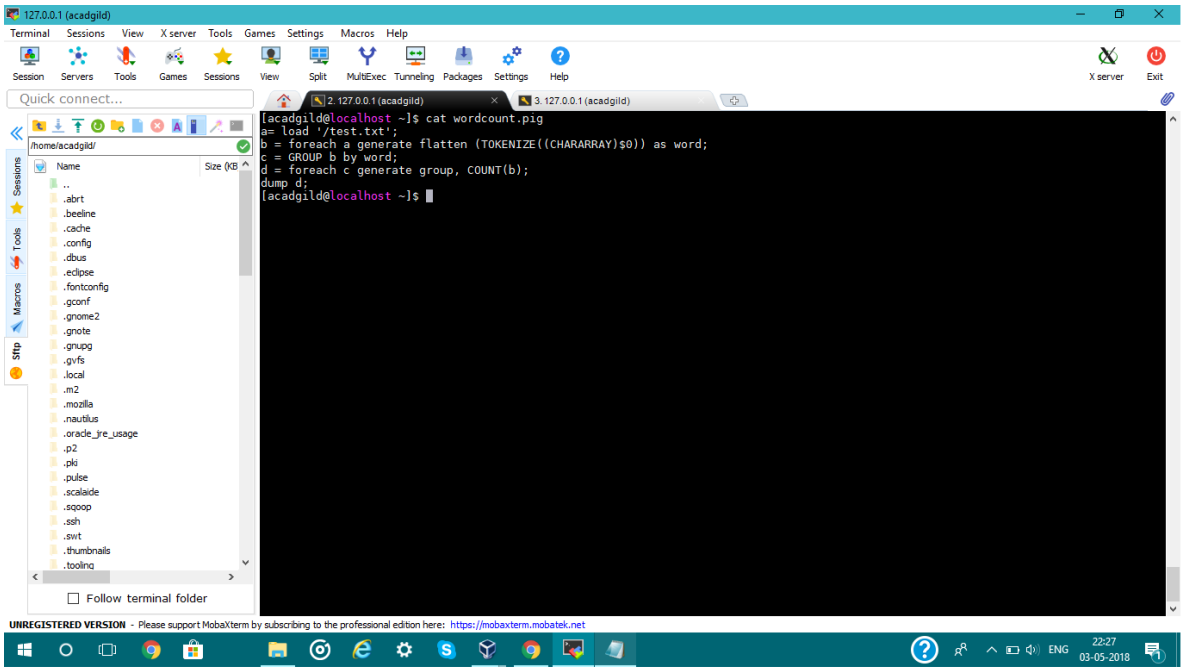
Assignment 7

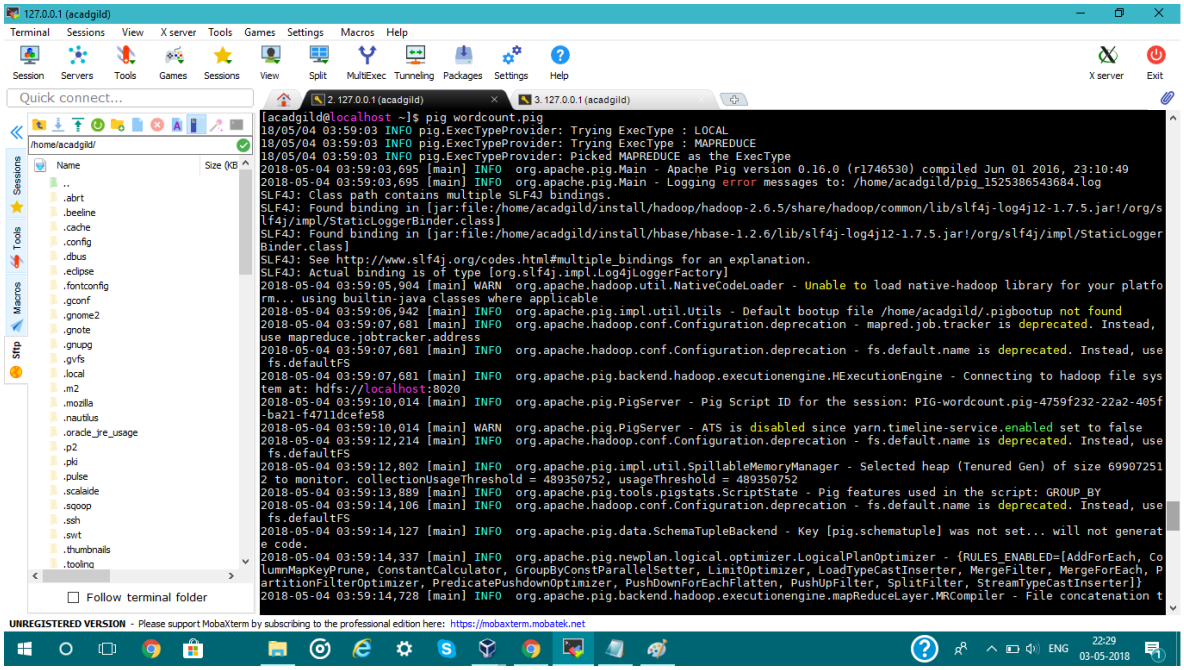
Task1: Write a program to implement wordcount using Pig

Pig commands:

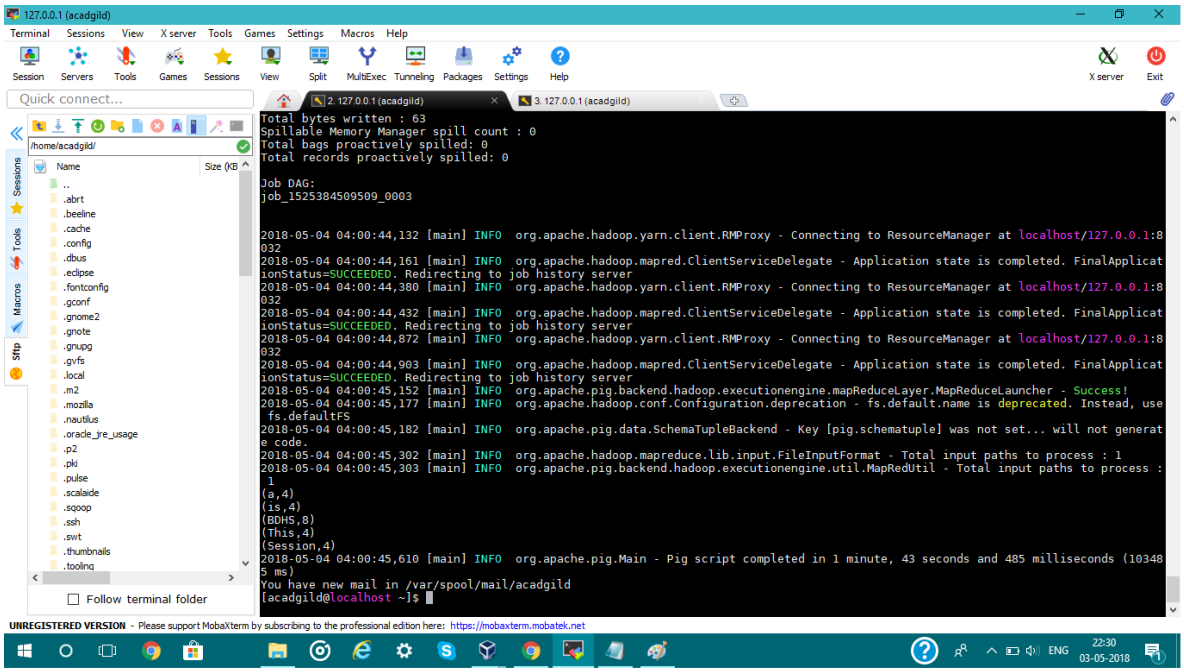


- a is loading the test.txt file, for which the count of the occurrence of the words is to be calculated.
- b is used to split every line of the text file into words.
- C is used to group same words together.
- D is used to calculate the count of words.

Command to execute Pig script: pig wordcount.pig

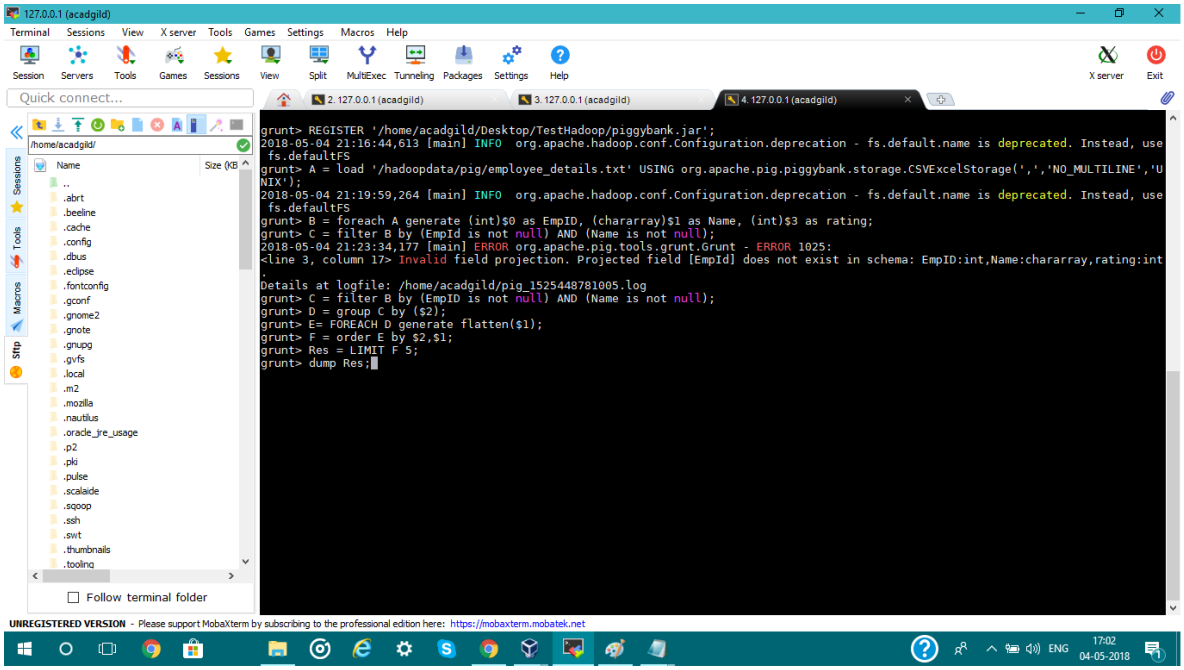


Output



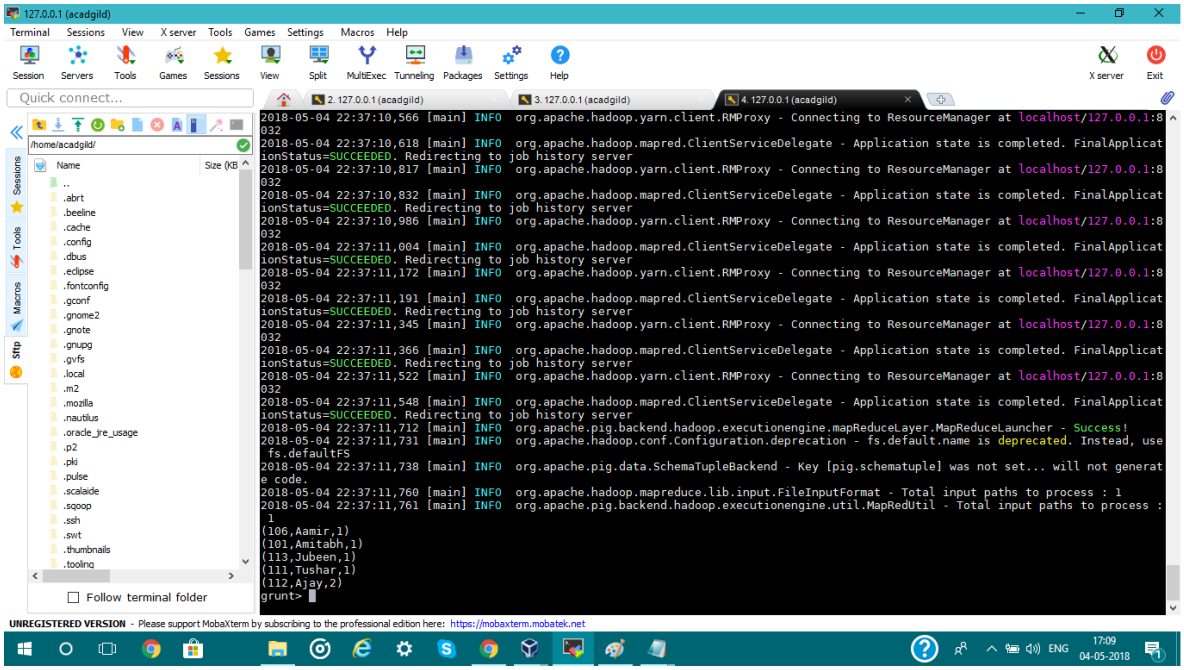
Task2

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



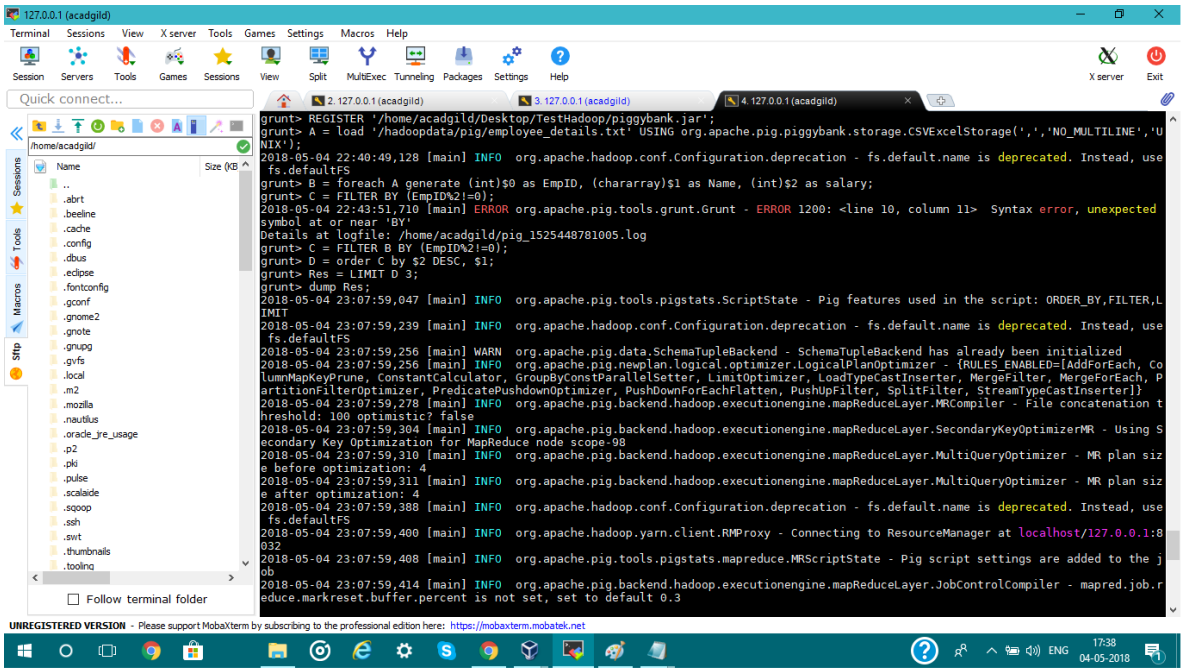
- In first step, we need to register the jar file (piggybank.jar) for executing pig commands.
- Then we are loading the employee_details.txt file.
- In B, we are displaying the EmpID, Name and rating of the employees, by specifying the column position.
- In C, we are filtering out those columns in which 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



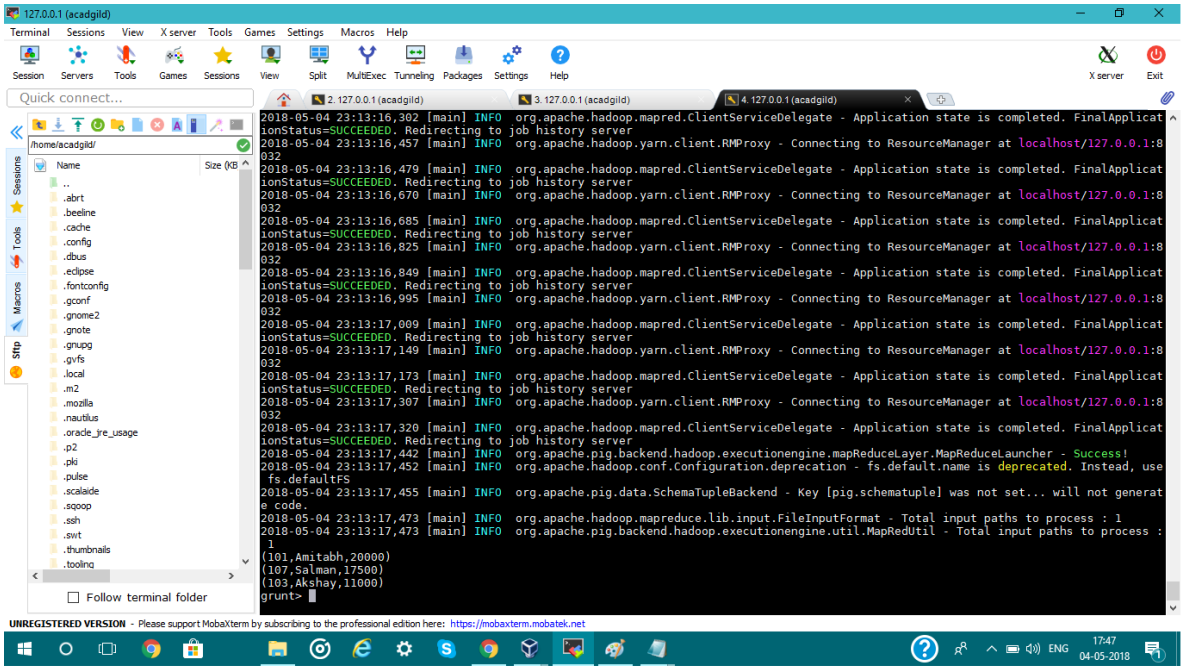
---

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



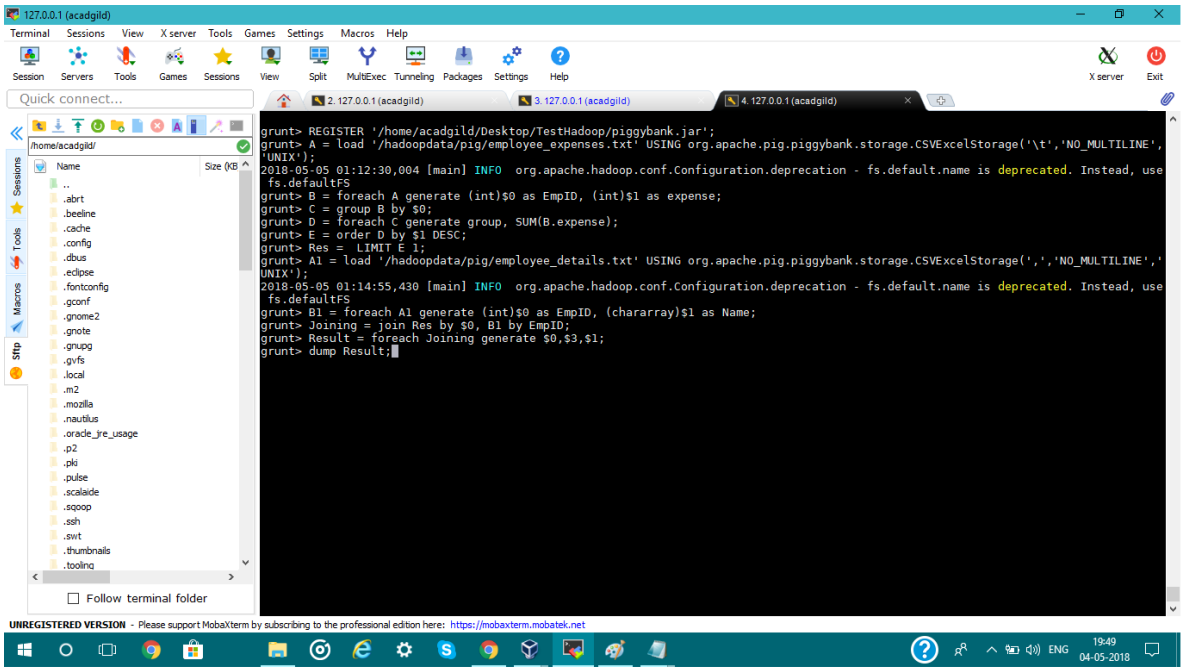
- The first two steps are the same as that of task a.
- In C, we are filtering those records where Emp_ID is odd.
- In D, we are sorting C on the basis of salary in descending order and then on the basis of emp_id.
- In Res, we are limiting the output of D to top 3 rows.

Output



---

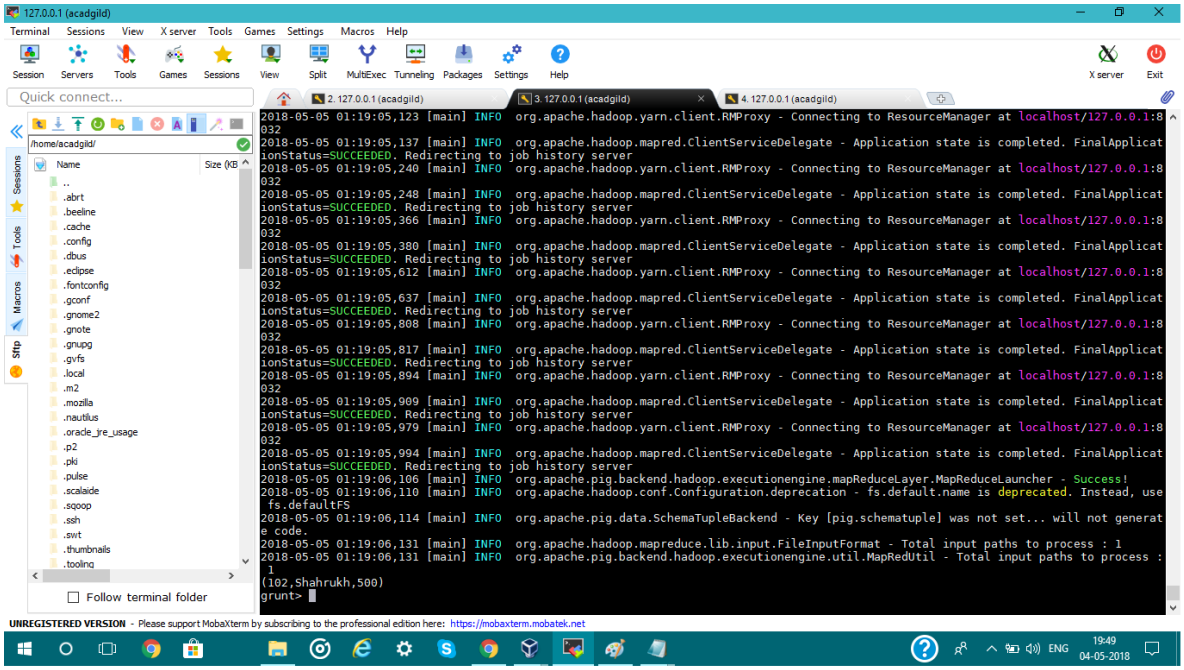
c. Display top 5 employees (employee_id and employee_name) with maximum expense



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

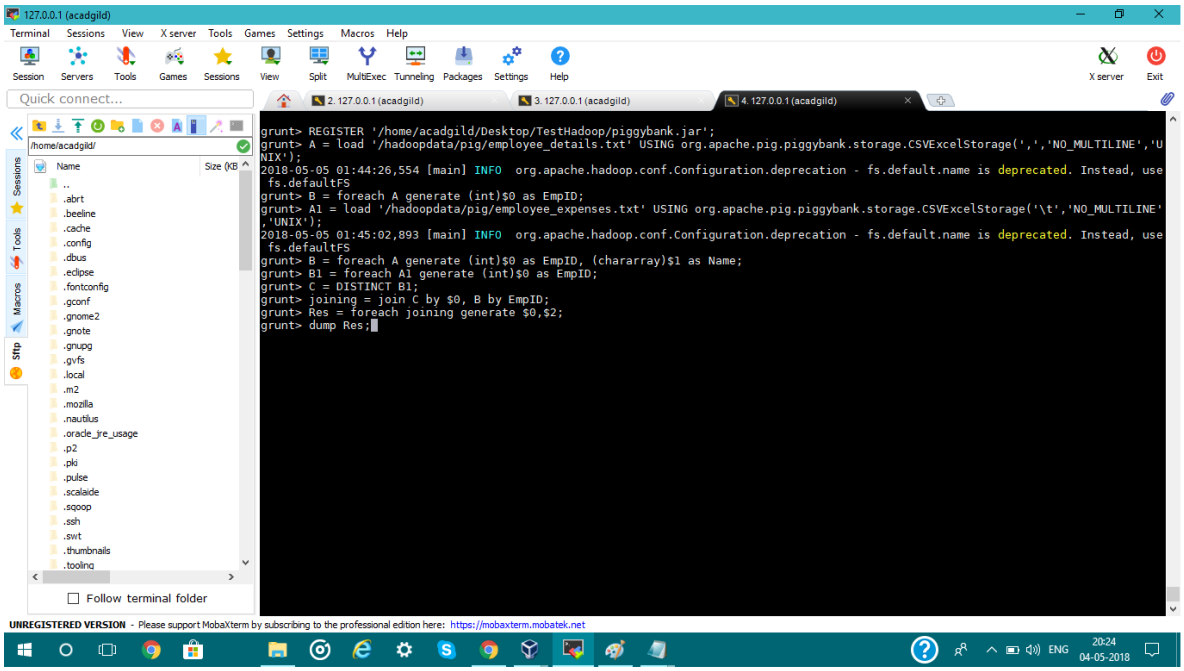
- In Joining, we are joining the two relations on the basis of employee id and then for final result, we are displaying the employee_id, employee name and his expense.

Output



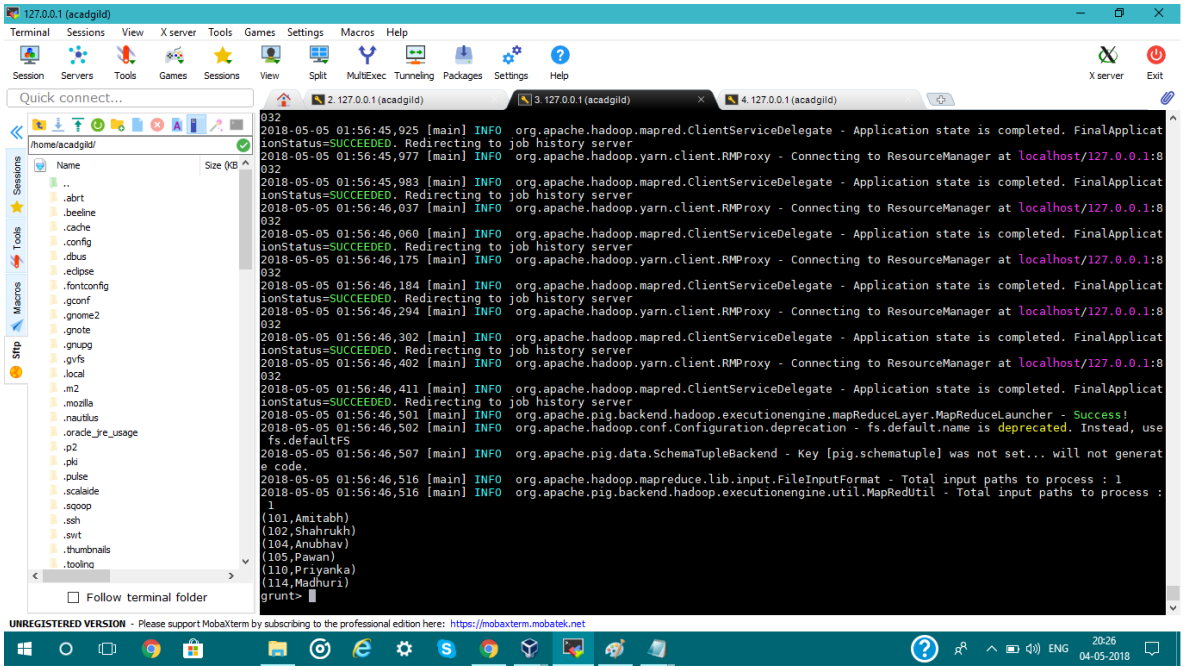
---

d. List of employees having entries in employee_expenses file

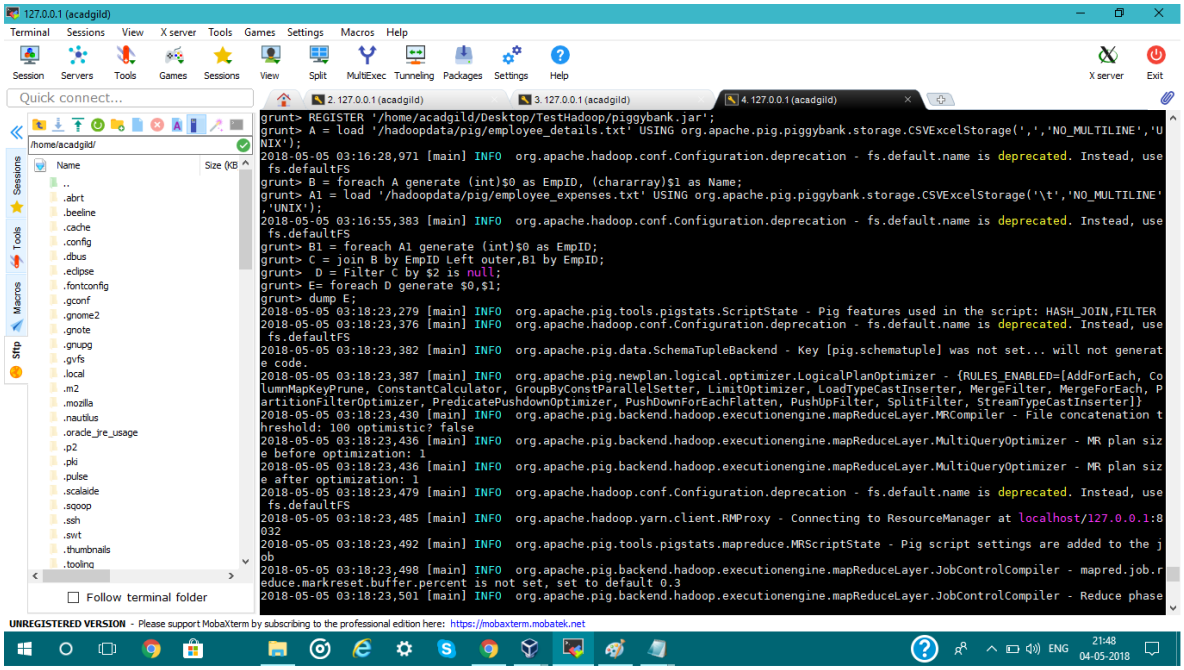


Here, we are joining both the relations on the basis of employee_id and then displaying the result.

Output

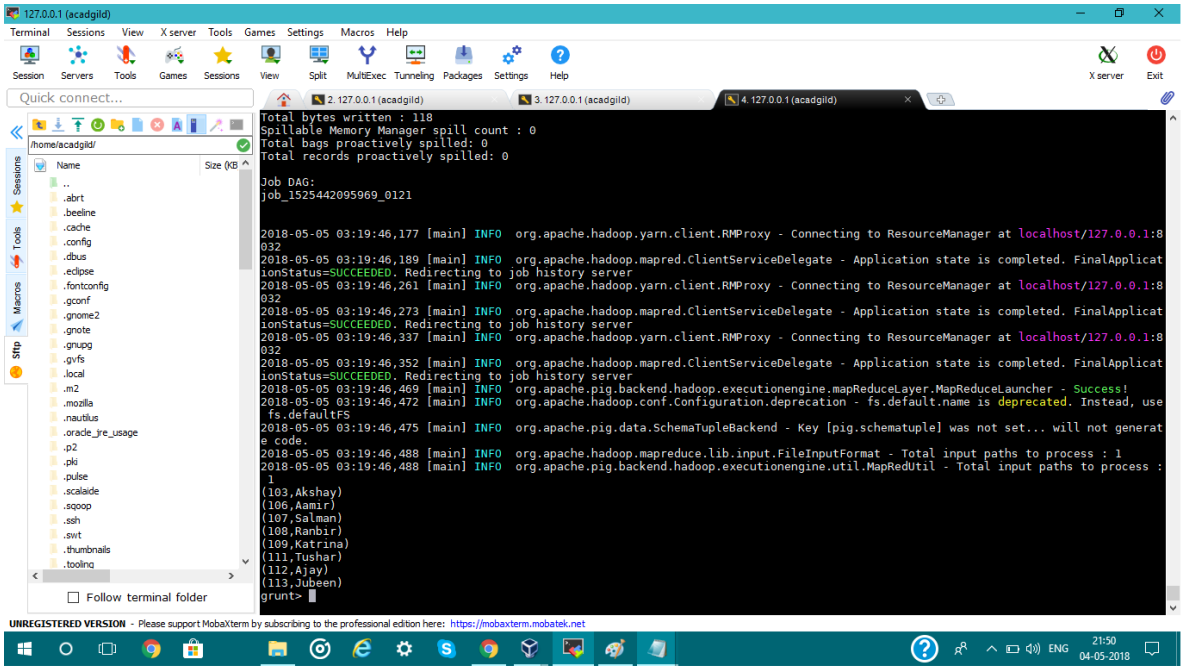


e. List of employees having no entries in employee_expenses file



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

Output

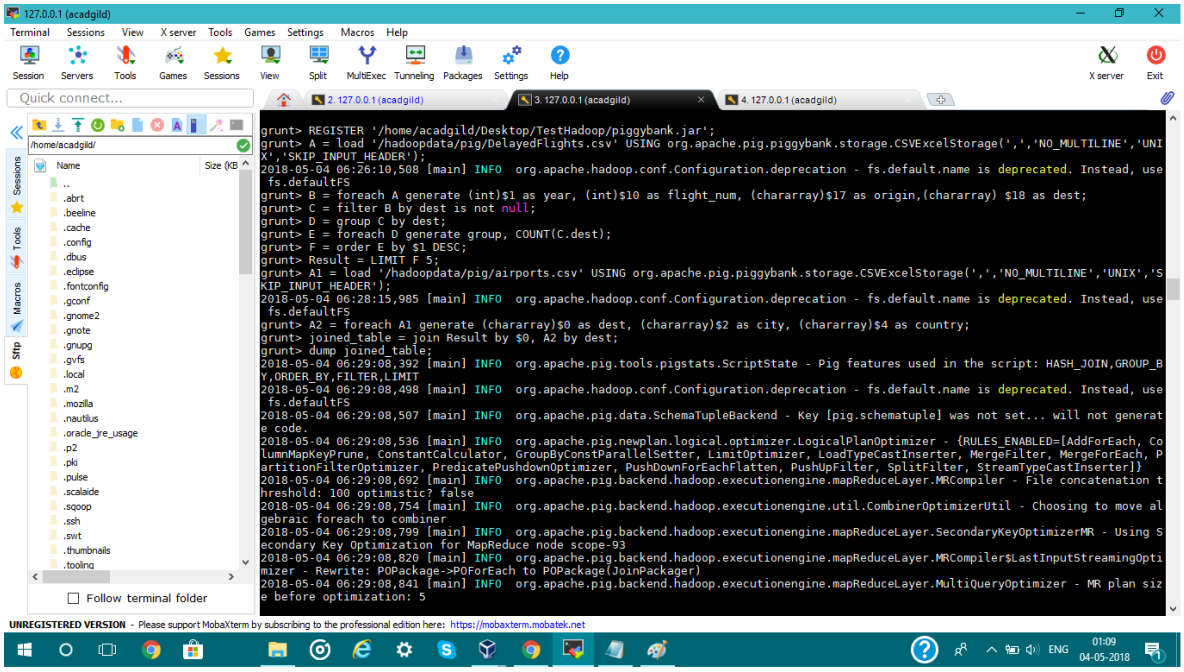


Task 3

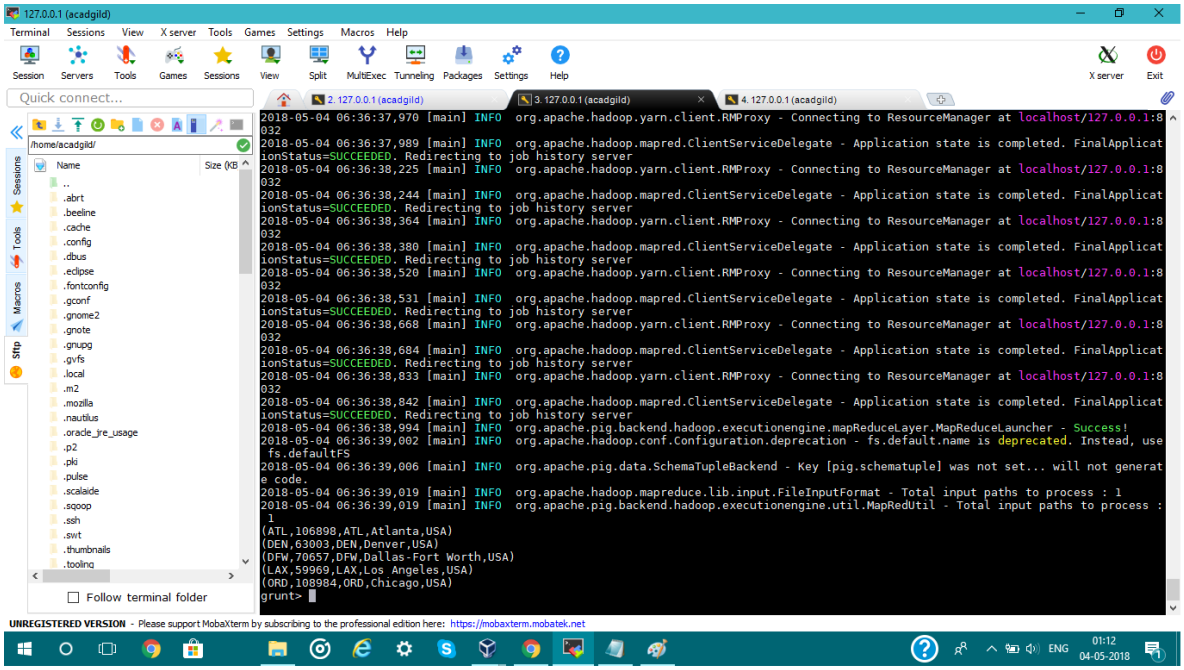
Implement the aviation data analysis use case

a. Find out the top 5 most visited destinations.

Pig script



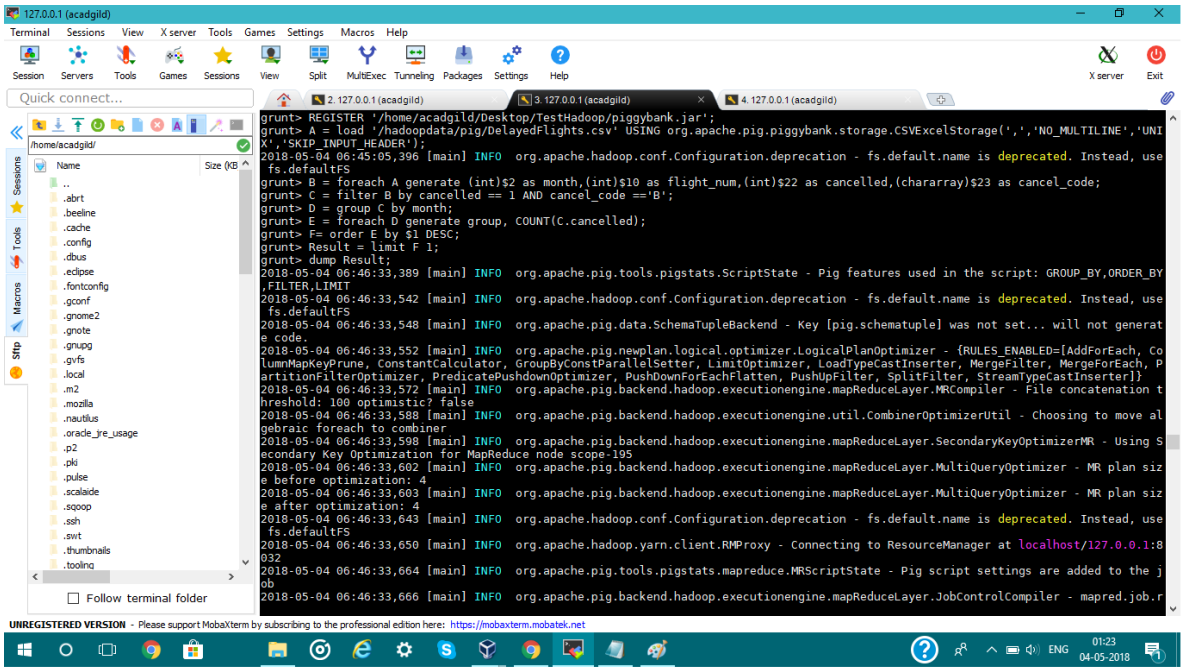
Output



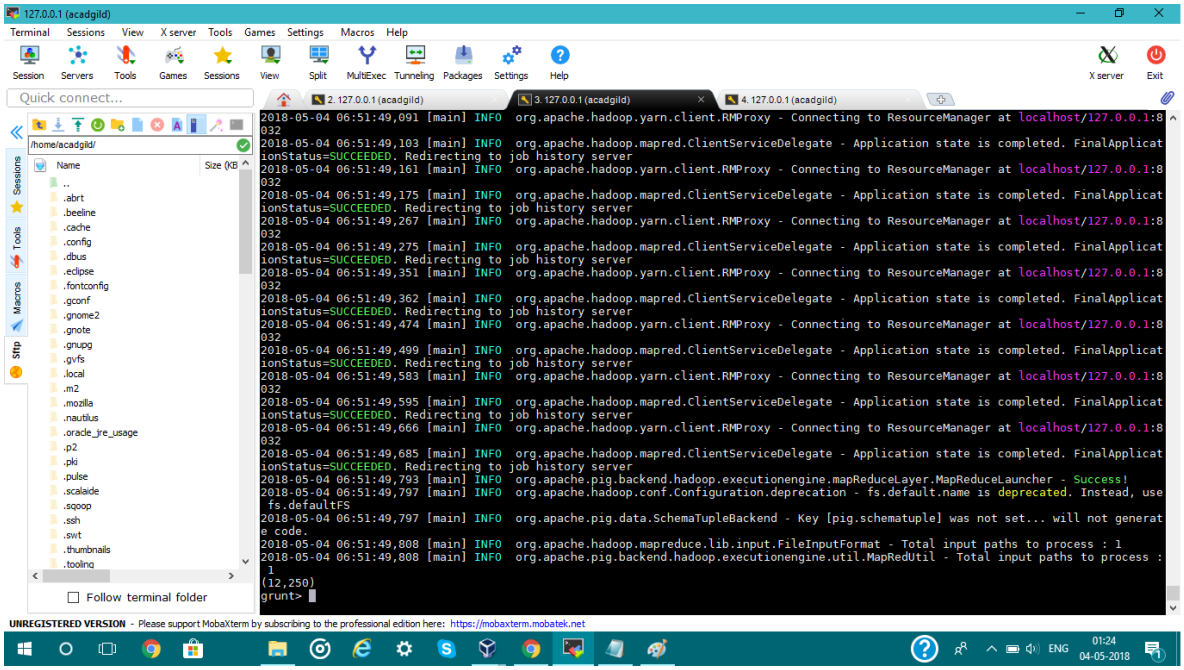
---

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

Pig Commands



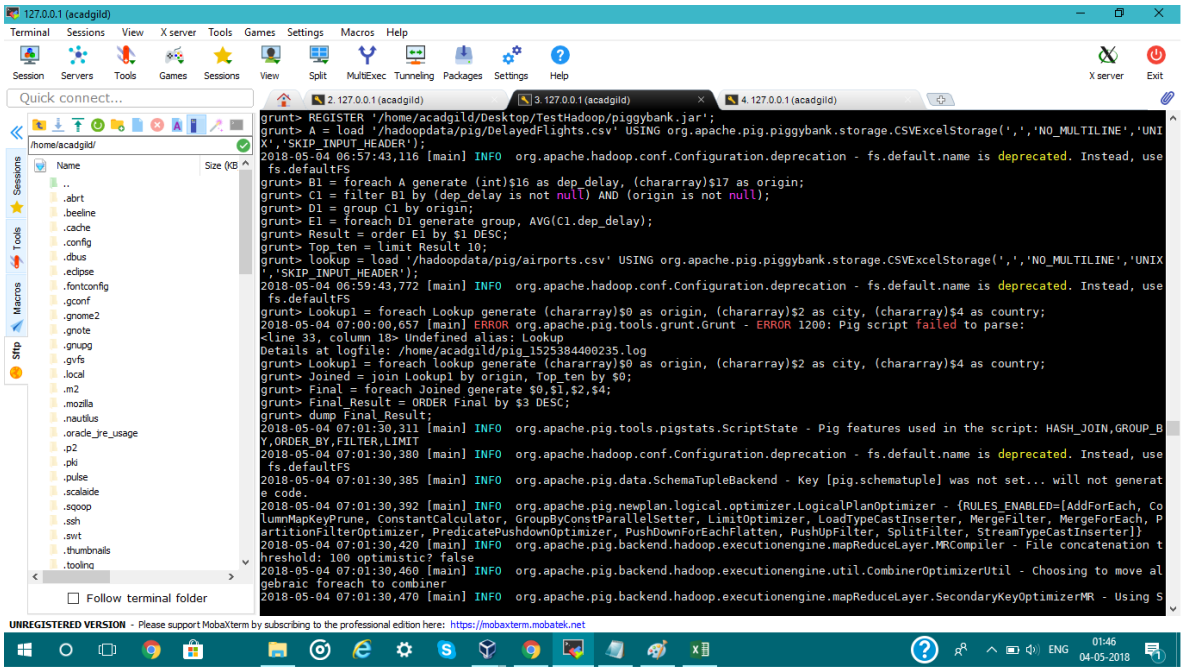
Output



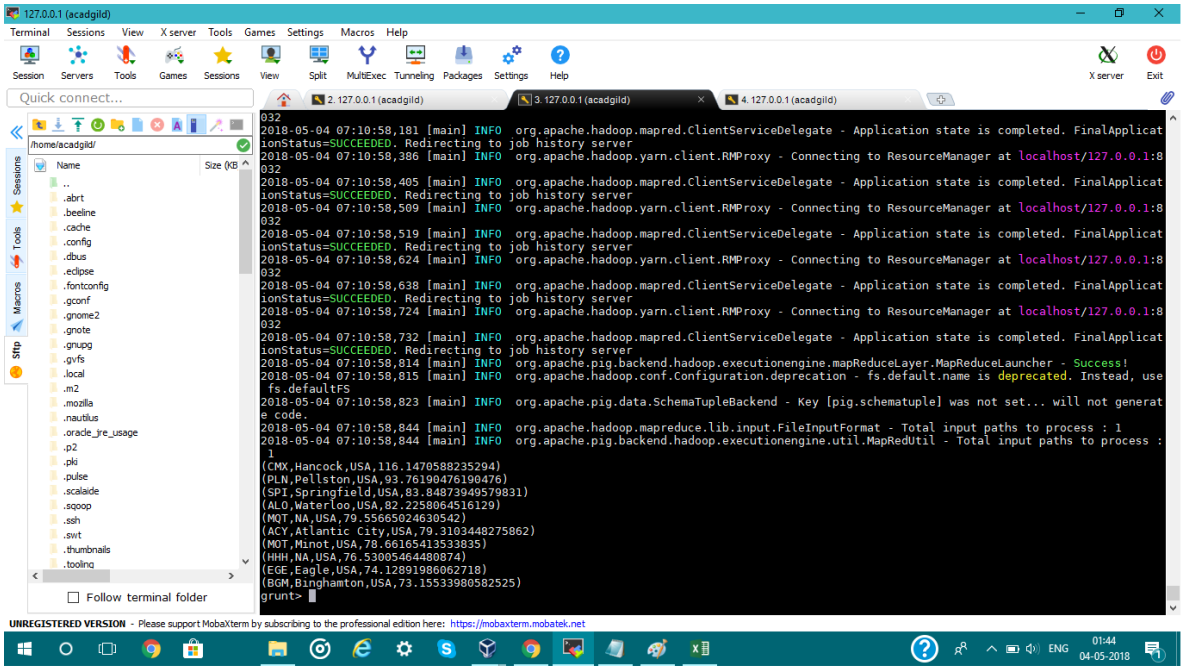
---

c. Top ten origins with the highest AVG departure delay

Pig commands



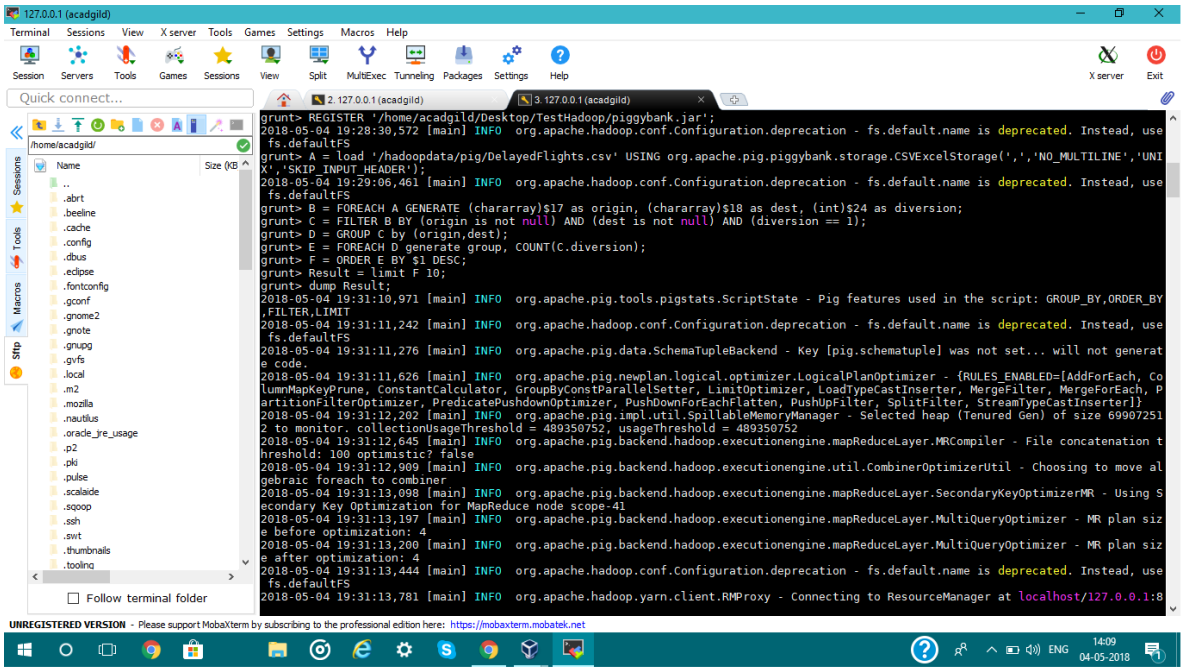
Output



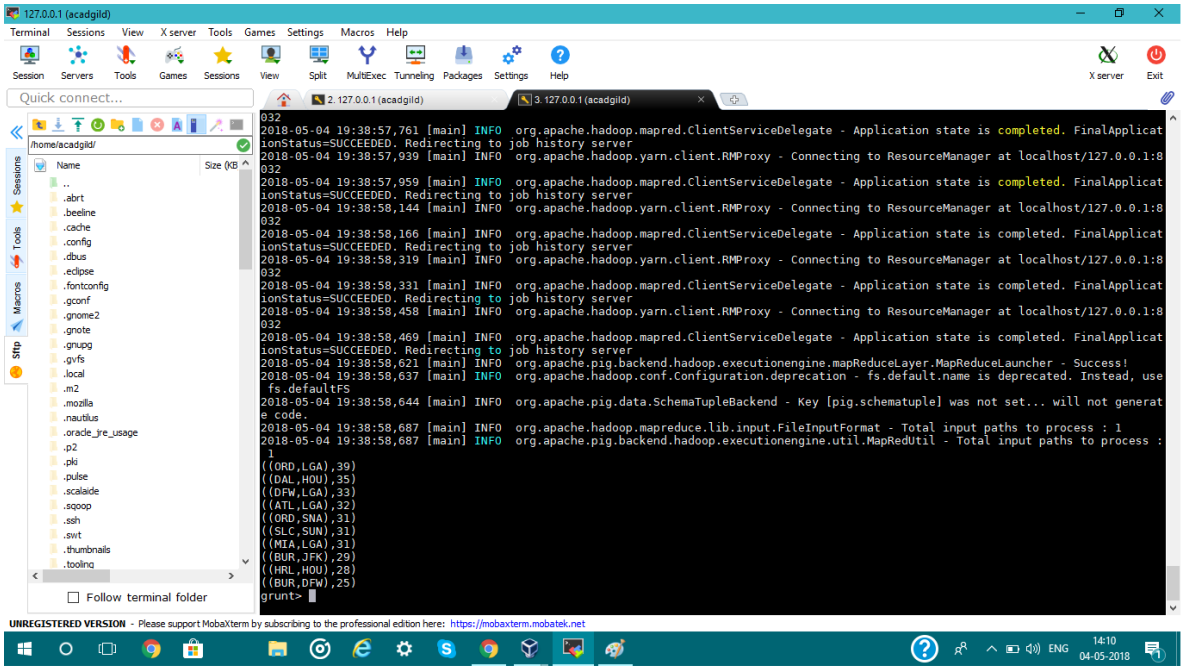
---

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

Pig commands



Output


127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Quick connect...
2. 127.0.0.1 (acadgild)
3. 127.0.0.1 (acadgild)
/home/acadgild/
Follow terminal folder
032
2018-05-04 19:38:57,761 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 19:38:57,939 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 19:38:57,959 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 19:38:58,144 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 19:38:58,166 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 19:38:58,319 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 19:38:58,331 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 19:38:58,458 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8
032
2018-05-04 19:38:58,469 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicat
ionStatus=SUCCEEDED. Redirecting to job history server
2018-05-04 19:38:58,621 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-05-04 19:38:58,637 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2018-05-04 19:38:58,644 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generat
e code.
2018-05-04 19:38:58,687 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-05-04 19:38:58,687 [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)
grunt>
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
ENG
14:10
04-05-2018