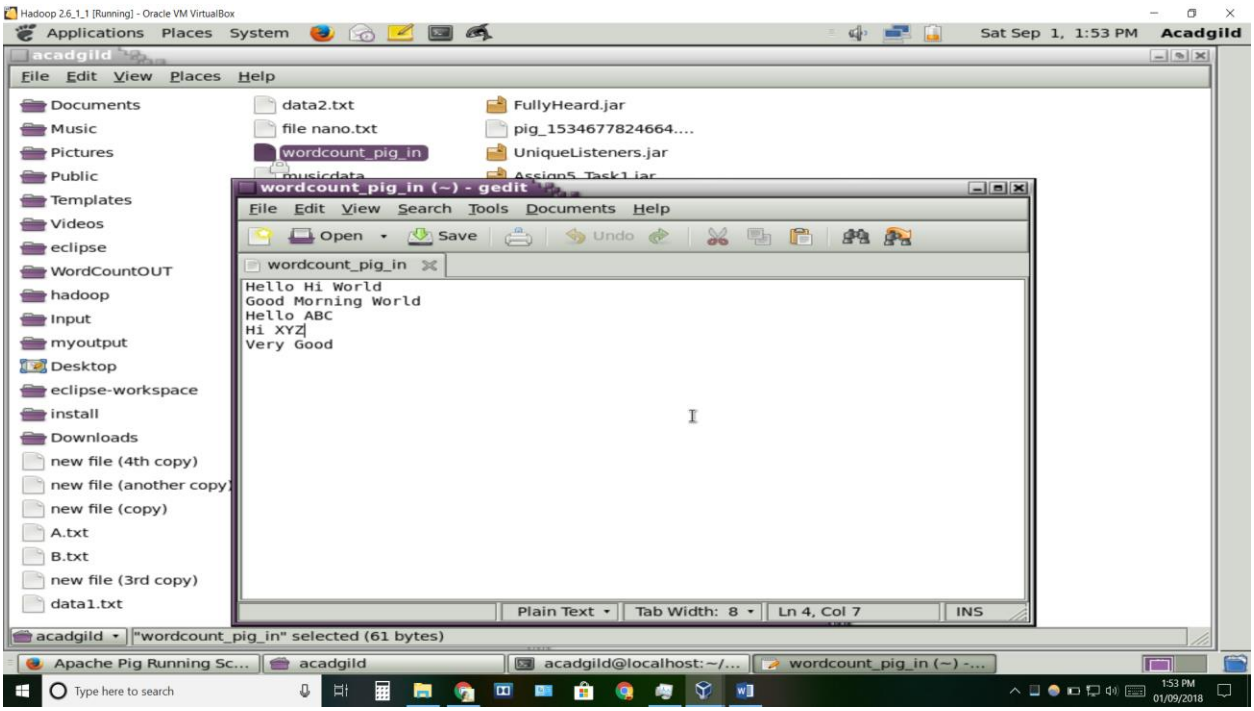
Big Data Hadoop 'Assignment Seven'

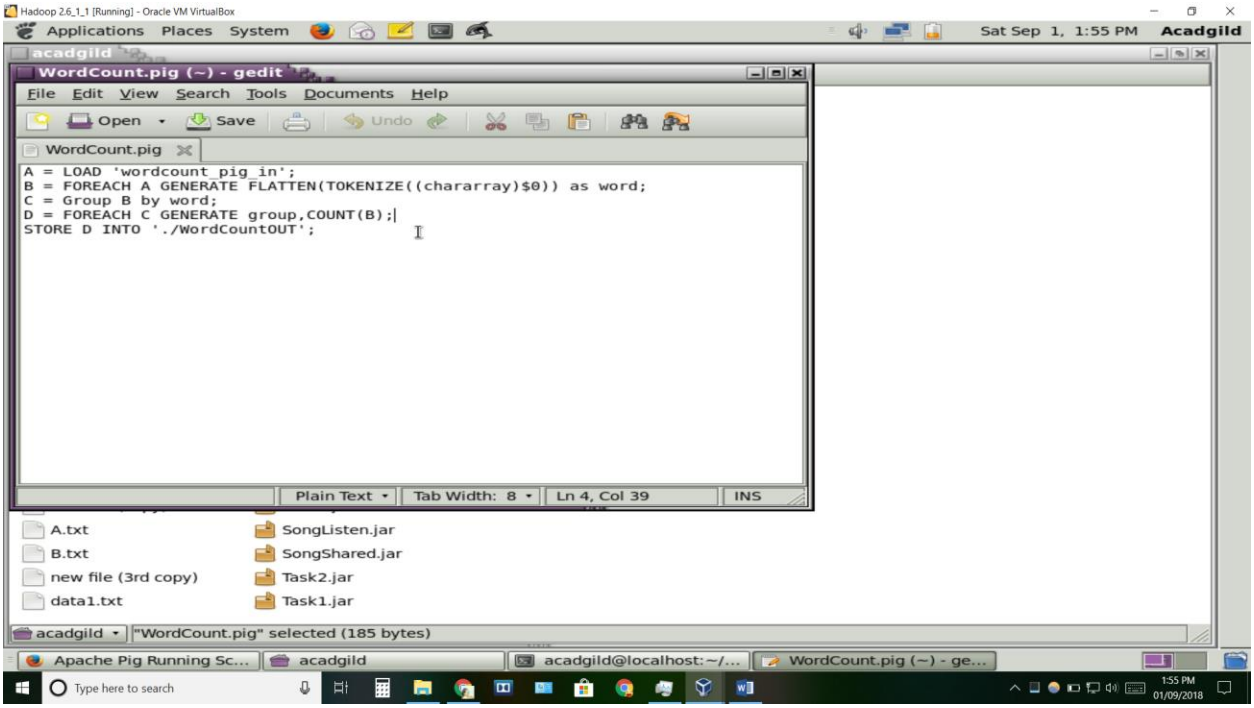
Write PIG scripts for following tasks.

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

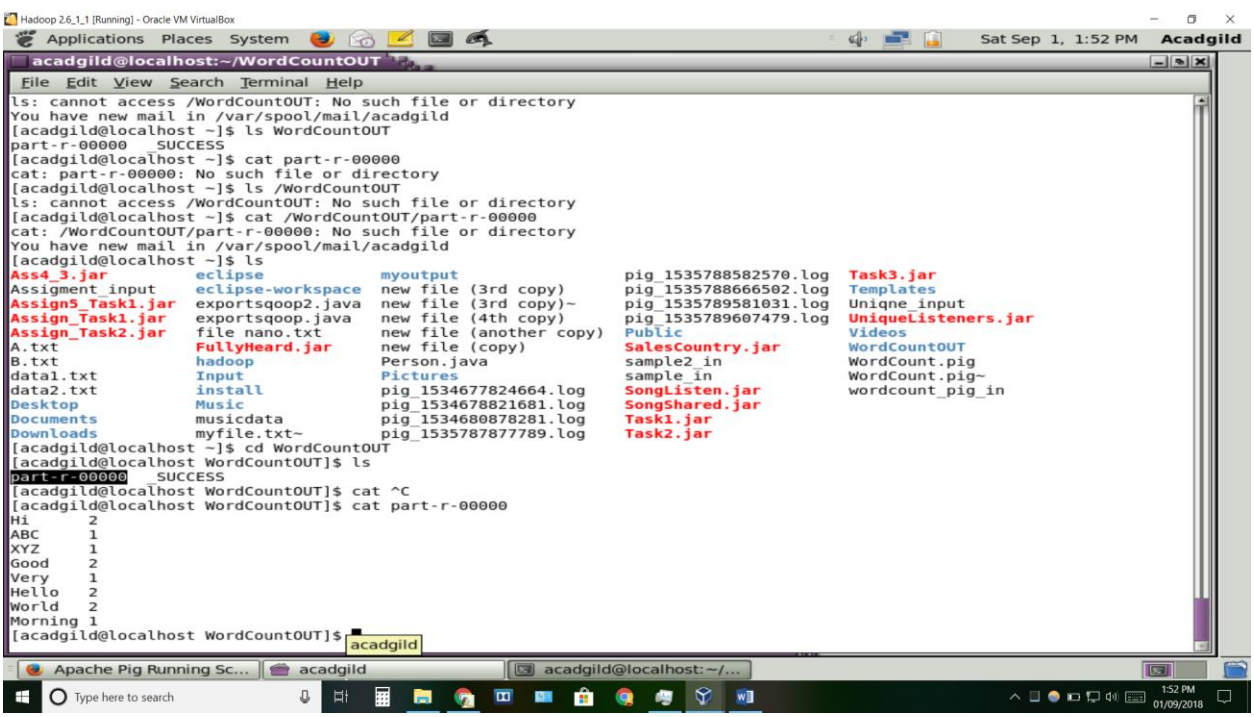
- Executed in pig LOCAL mode
- Input data: Wordcount_pig_in file



- WordCount pig script



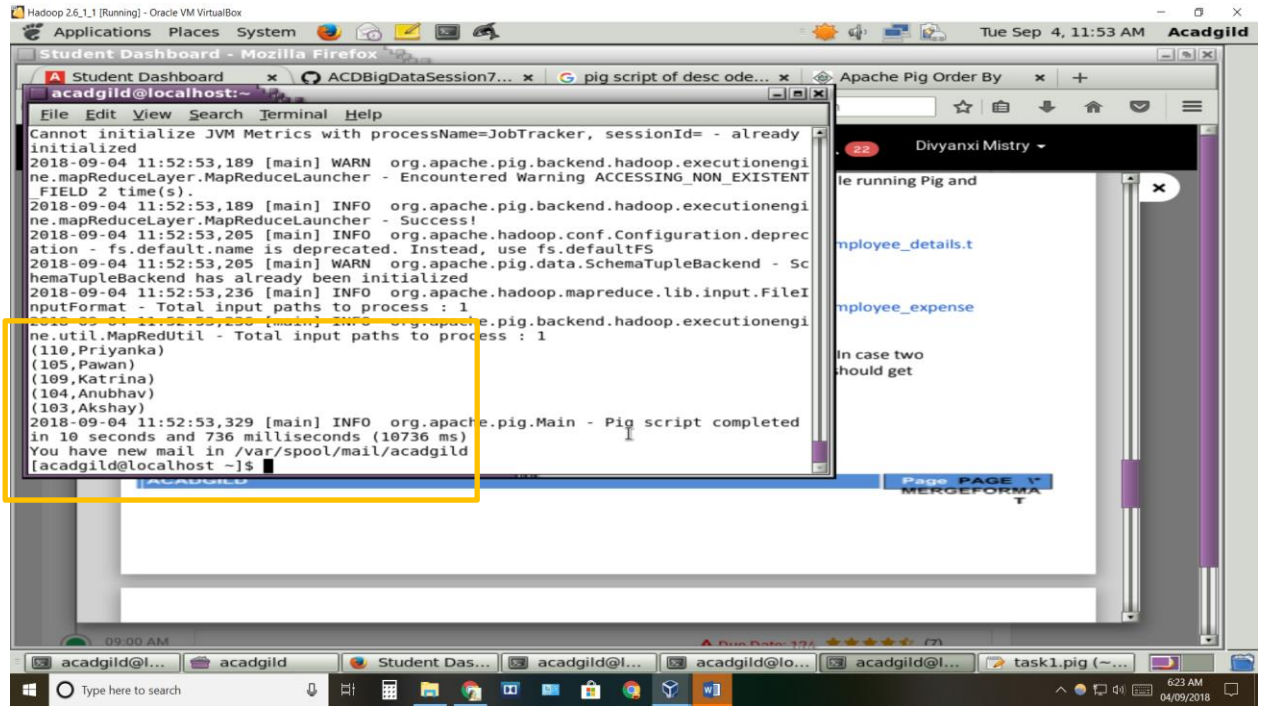
- Final Output



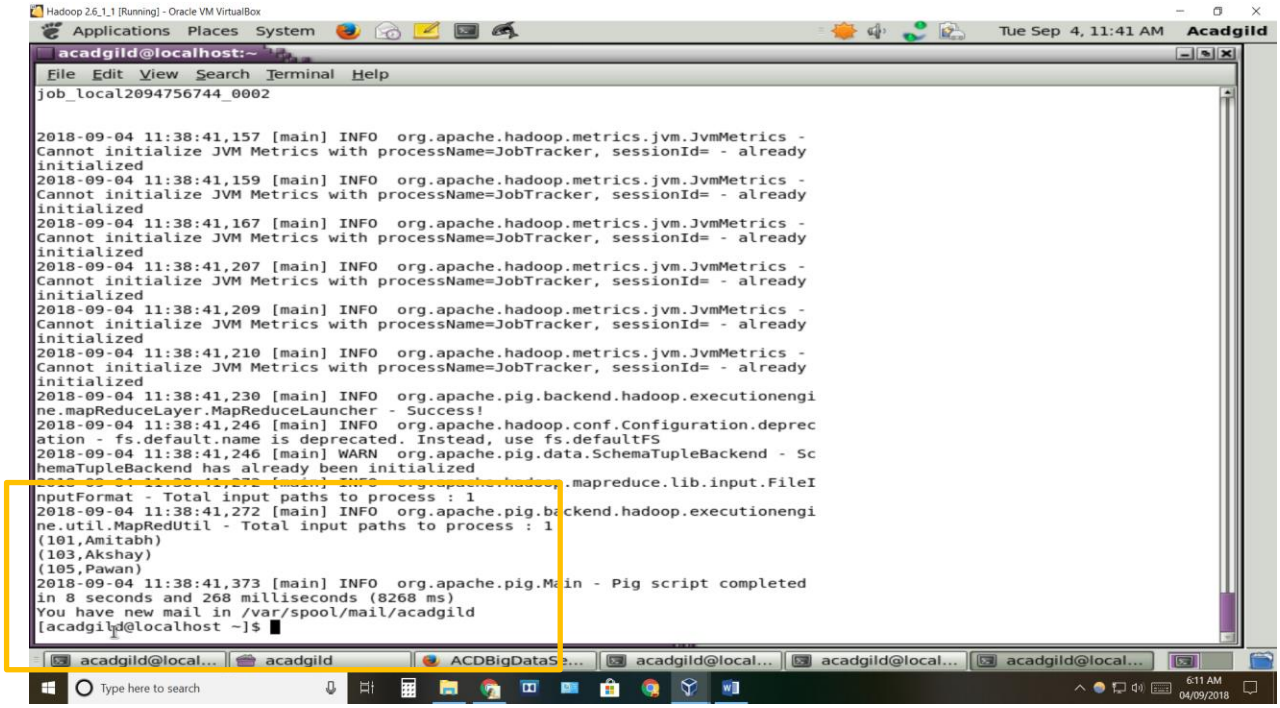
Task 2

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

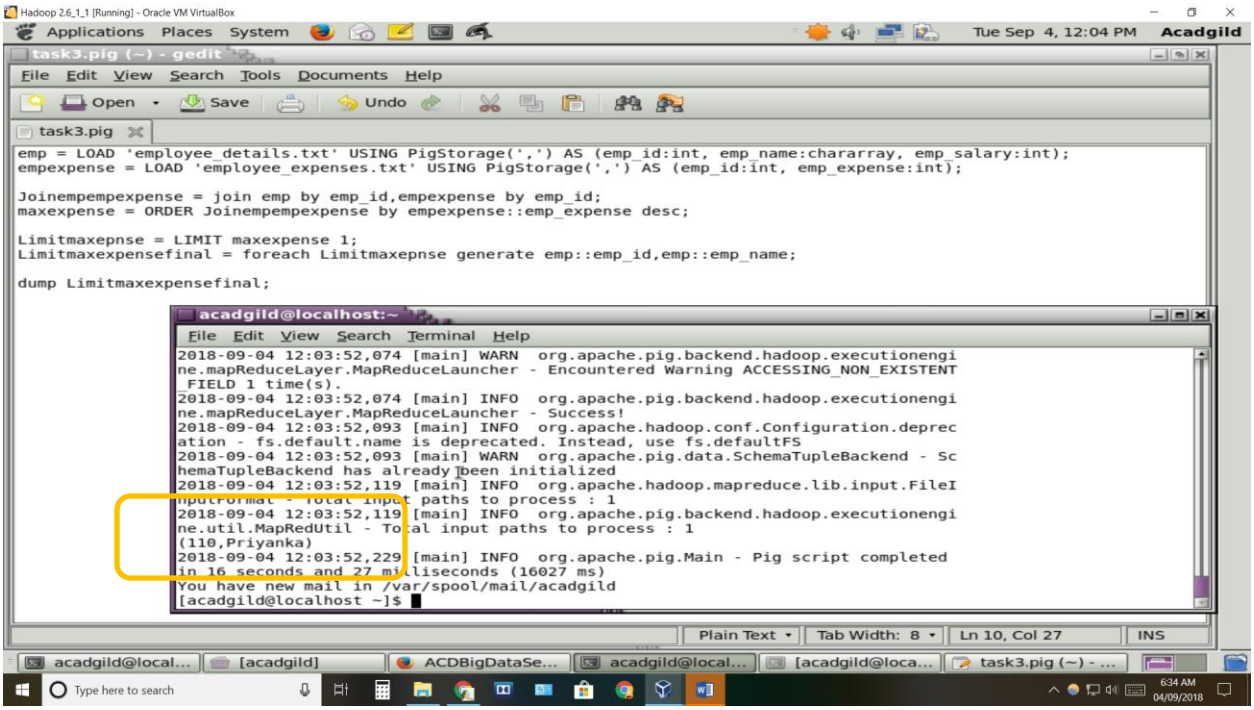
(a) Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)



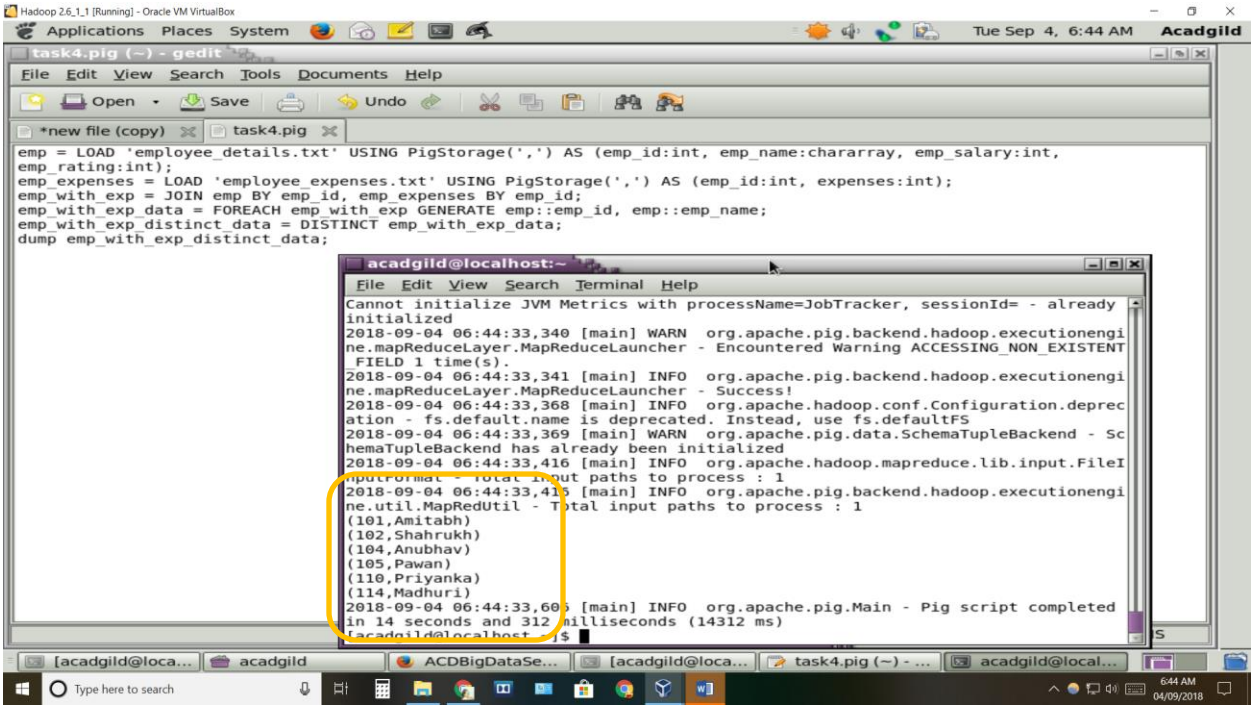
(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference)



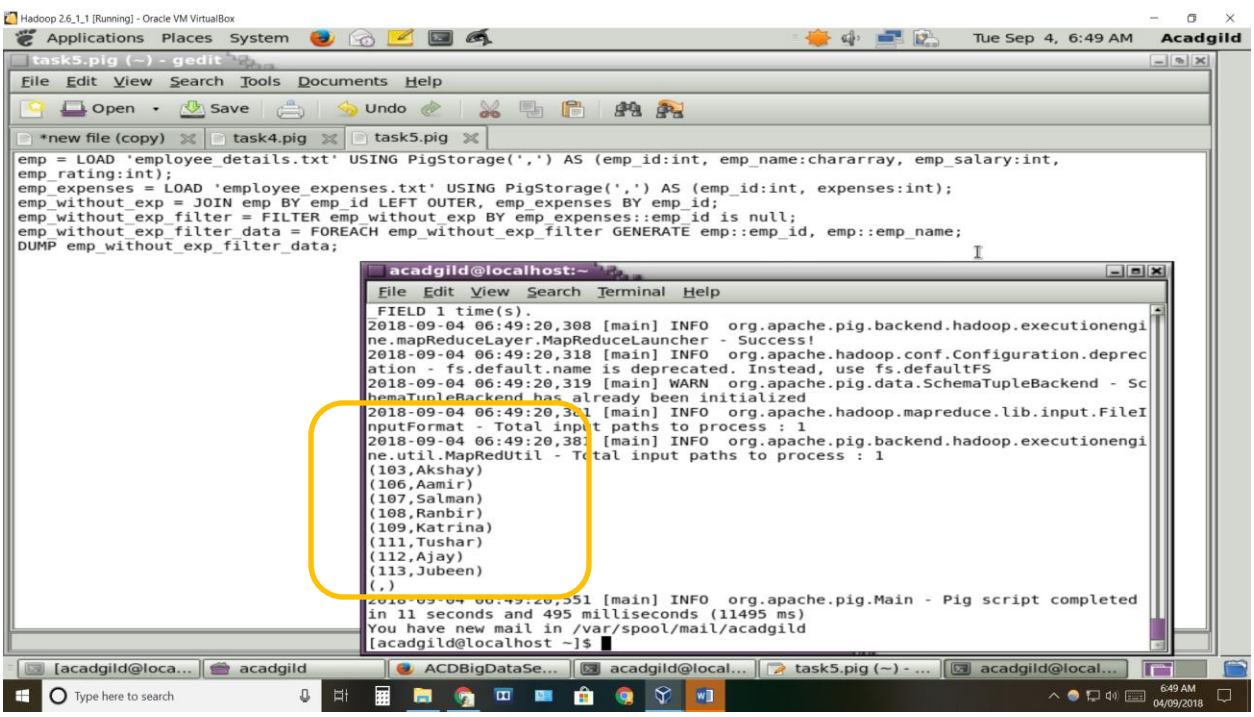
(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)



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



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

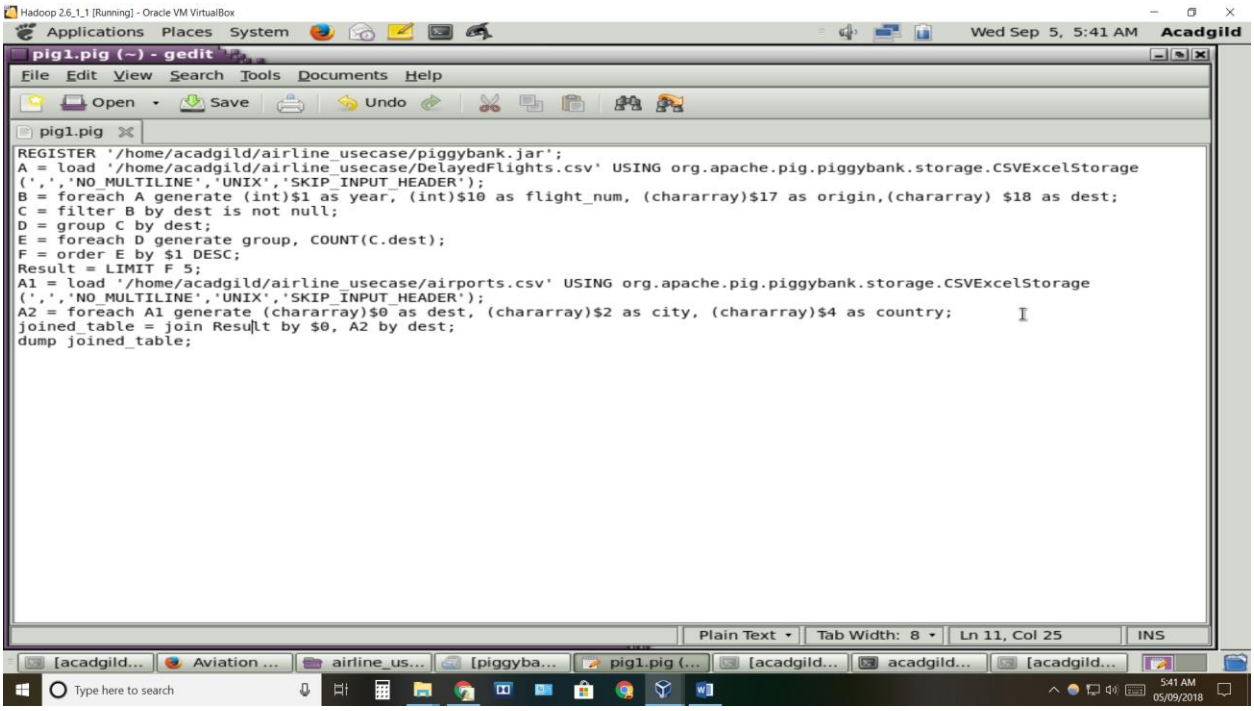


Task 3

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

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

1. Wrote Pig Script under and saved under pig1.pig

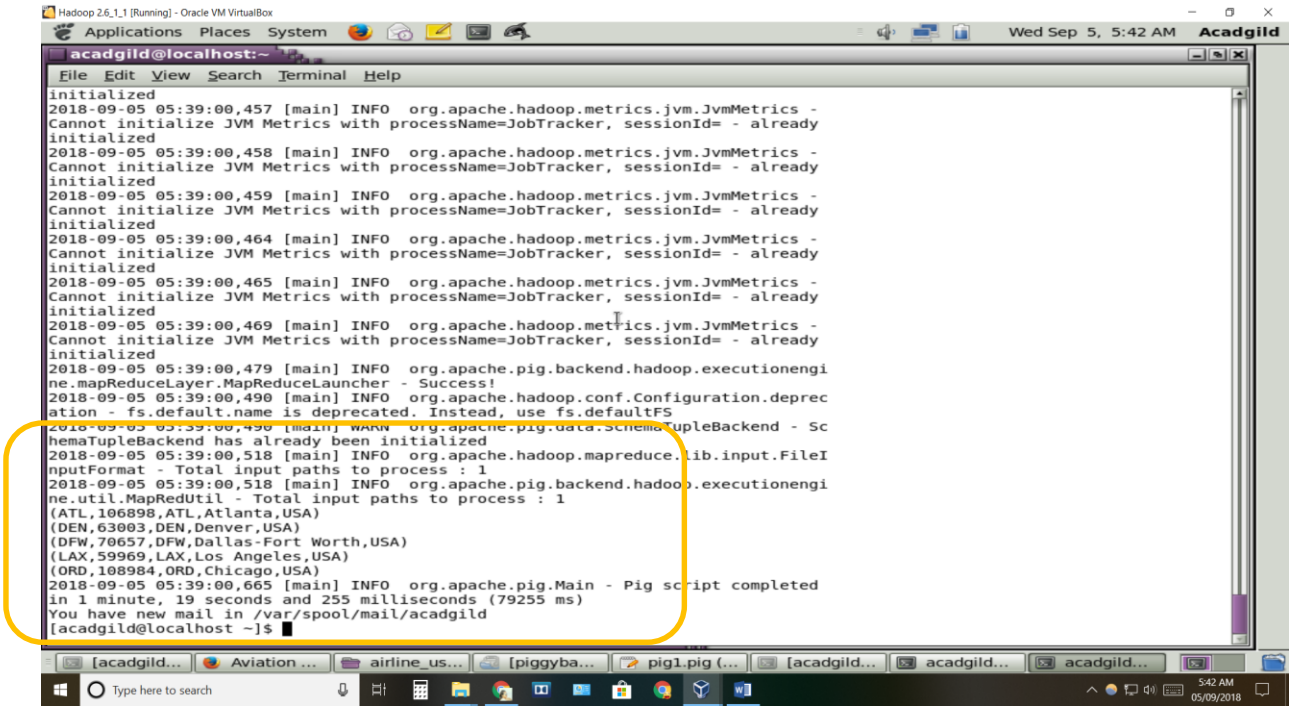


Steps

- **In First Line**: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.
- In relation **A**, we are loading the dataset using CSVExcelStorage.
- In relation **B**, we are generating the columns that are required for processing and typecasting each of them like int,chararray.
- In relation **C**, we are filtering the null values from the "dest" column.
- In relation **D**, we are grouping relation C by "dest."
- In relation **E**, we are generating the grouped column and the count of each.
- Relation **F** is used order in DESC;
- **Result** is used limit the result to top 5.
- In relation **A1**, we are loading another table to find the city as well as the country.
- In relation **A2**, we are generating dest, city, and country from the previous relation.
- In relation **joined_table**, we are joining Result and A2 based on a common column, i.e., "dest"
- Finally, using dump, we are printing the result.

2. Ran PigScript in local Mode

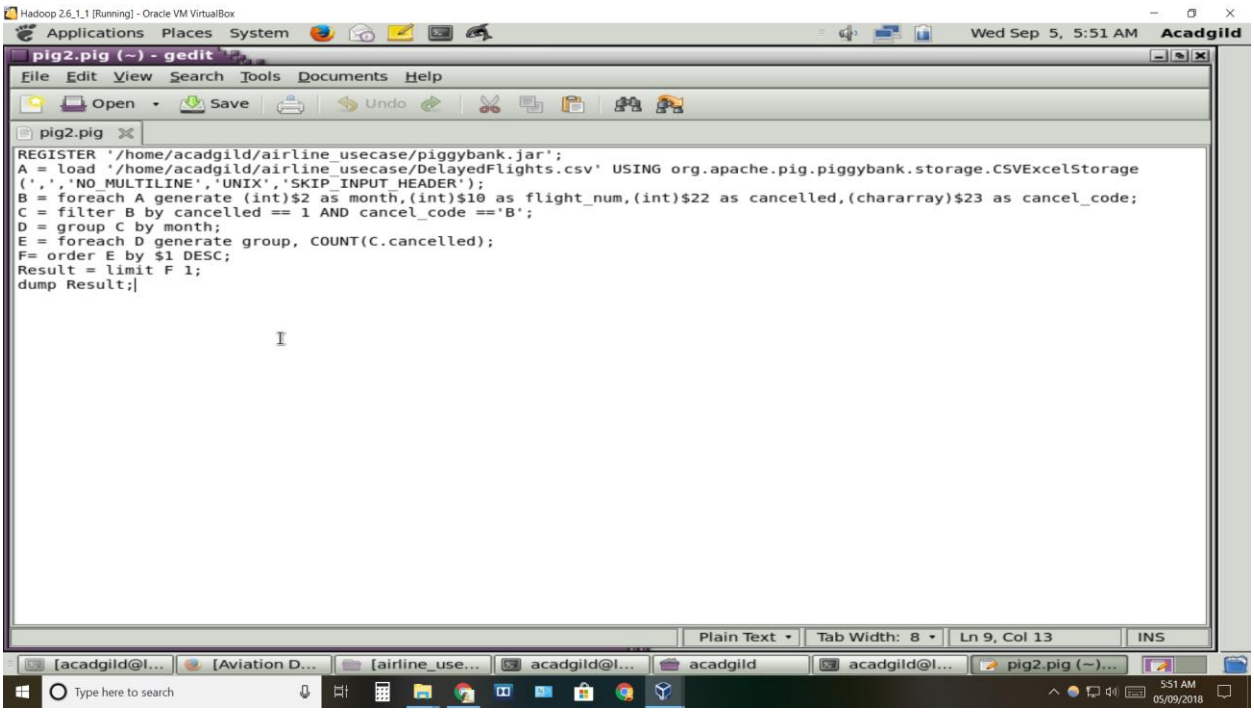
   Pig –x local pig1.pig



Problem Statement 2

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

1. Wrote Pig Script under and saved under pig2.pig

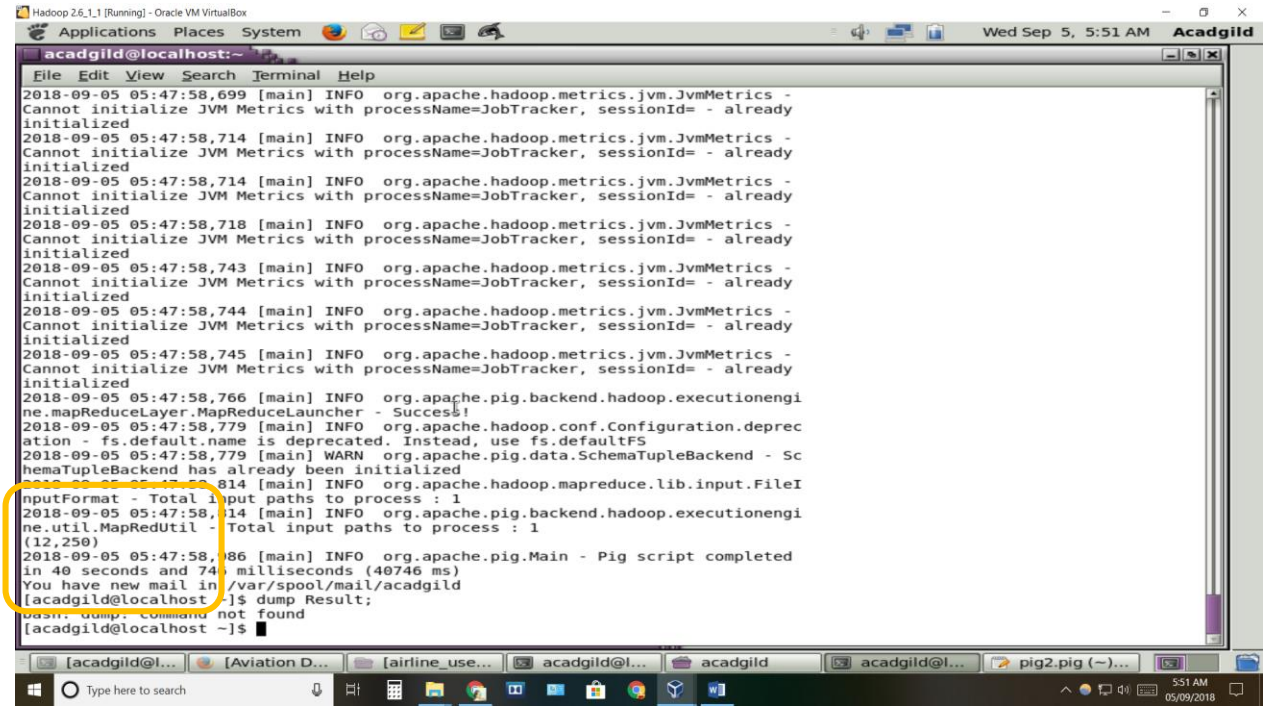


Steps

- **In First Line**: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.
- In relation **A**, we are loading the dataset using CSVExcelStorage.
- In relation **B**, we are generating the columns that are required for processing and typecasting each of them like int,chararray.
- In relation **C**, we are filtering the data based on cancellation and cancellation code, i.e., canceled = 1 means flight have been canceled and cancel_code = 'B' means the reason for cancellation is "weather." So relation C will point to the data which consists of canceled flights due to bad weather.
- In relation **D**, we are grouping the relation C based on every month.
- In relation **E**, we are finding the count of canceled flights every month.
- Relation **F** is for orderin desc.
- **Result** finding the top month based on cancellation with suing limit 1;
- **Dump Result** print results as shown below;

2.Ran PigScript in local Mode

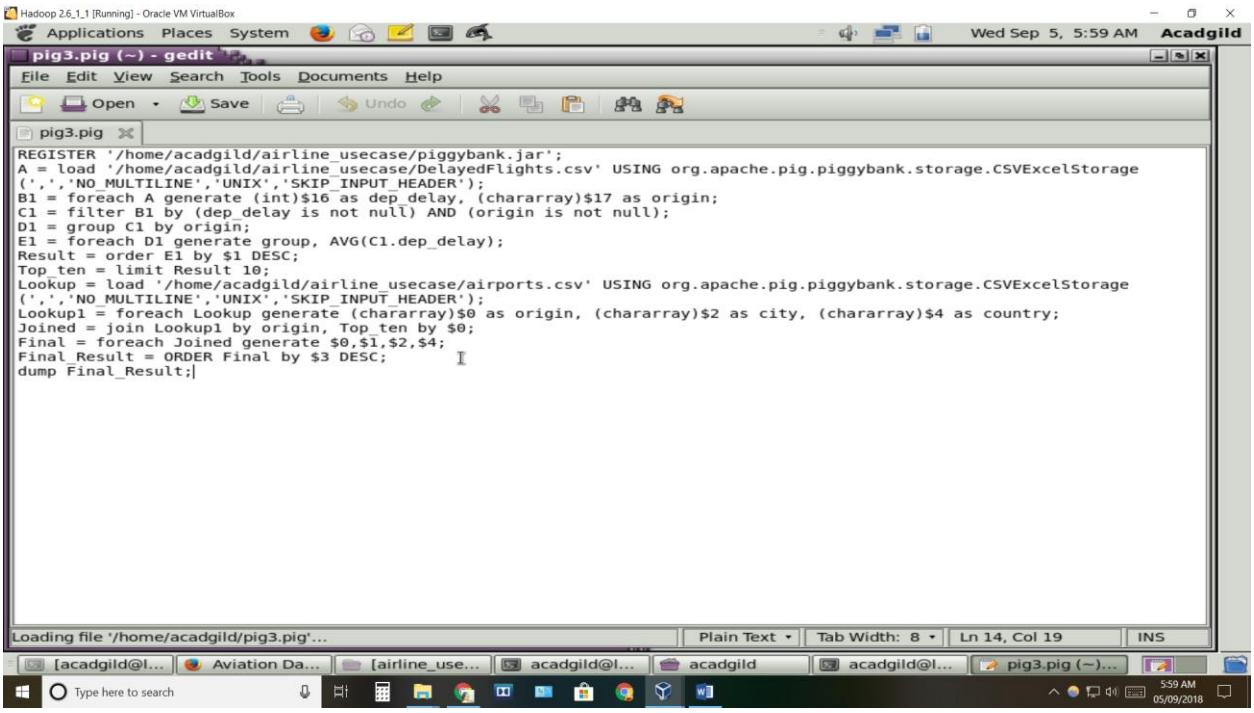
Pig –x local pig2.pig



Problem Statement 3

Top ten origins with the highest AVG departure delay

1. Wrote Pig Script under and saved under pig3.pig

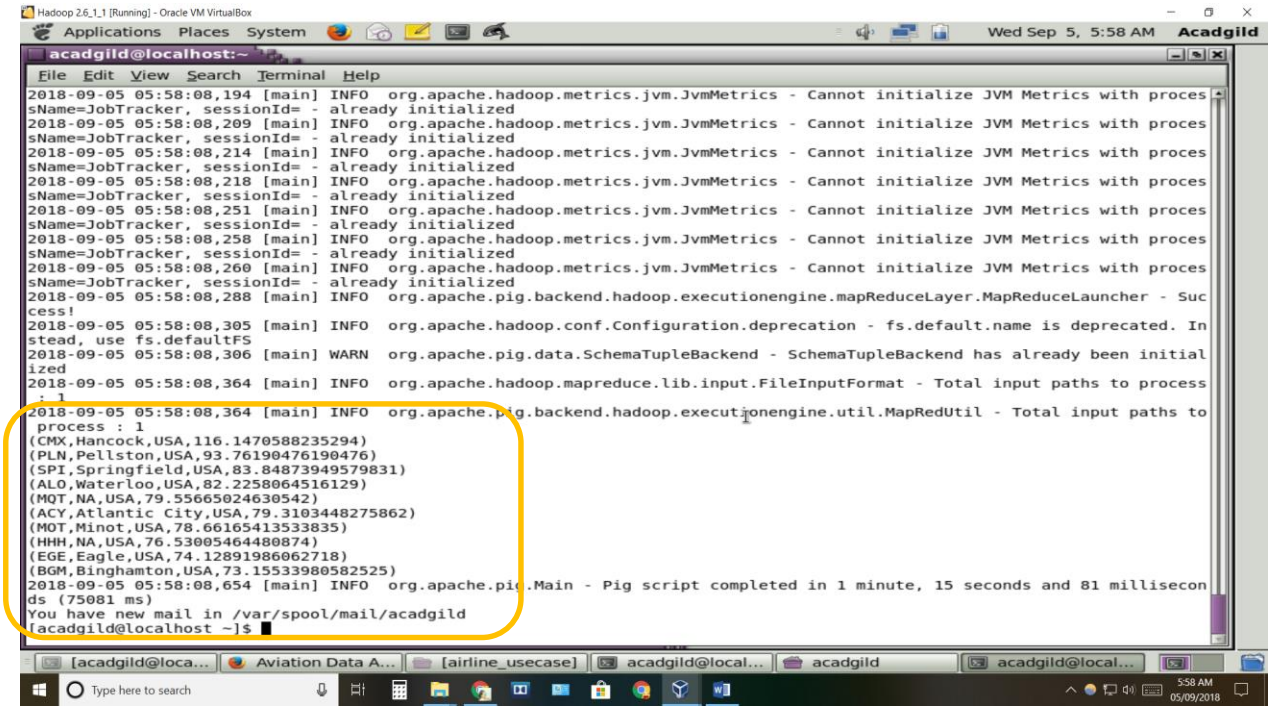


Steps

- **In First Line**: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.
- In relation **A**, we are loading the dataset using CSVExcelStorage.
- In relation **B**, we are generating the columns that are required for processing and typecasting each of them like int,chararray.
- In relation **C1**, we are removing the null values fields present if any.
- In relation **D1**, we are grouping the data based on column "origin."
- In relation **E1**, we are finding average delay from each unique origin.
- Relations named **Result** and **Top_ten** are ordering the results in descending order and printing the top ten values.
- In the relation **Lookup**, we are loading another table we will look up and find the city and country.
- In the relation L**ookup1,** we are generating the destination, city, and country from the previous relation.
- In the relation **Joined**, we are joining relation Top_ten and Lookup1 based on common a column, i.e., "origin."
- In the relation **Final,** we are generating required columns from the Joined table.
- Final_Result will order data by desc and dump will print results as shown below;

2.Ran PigScript in local Mode

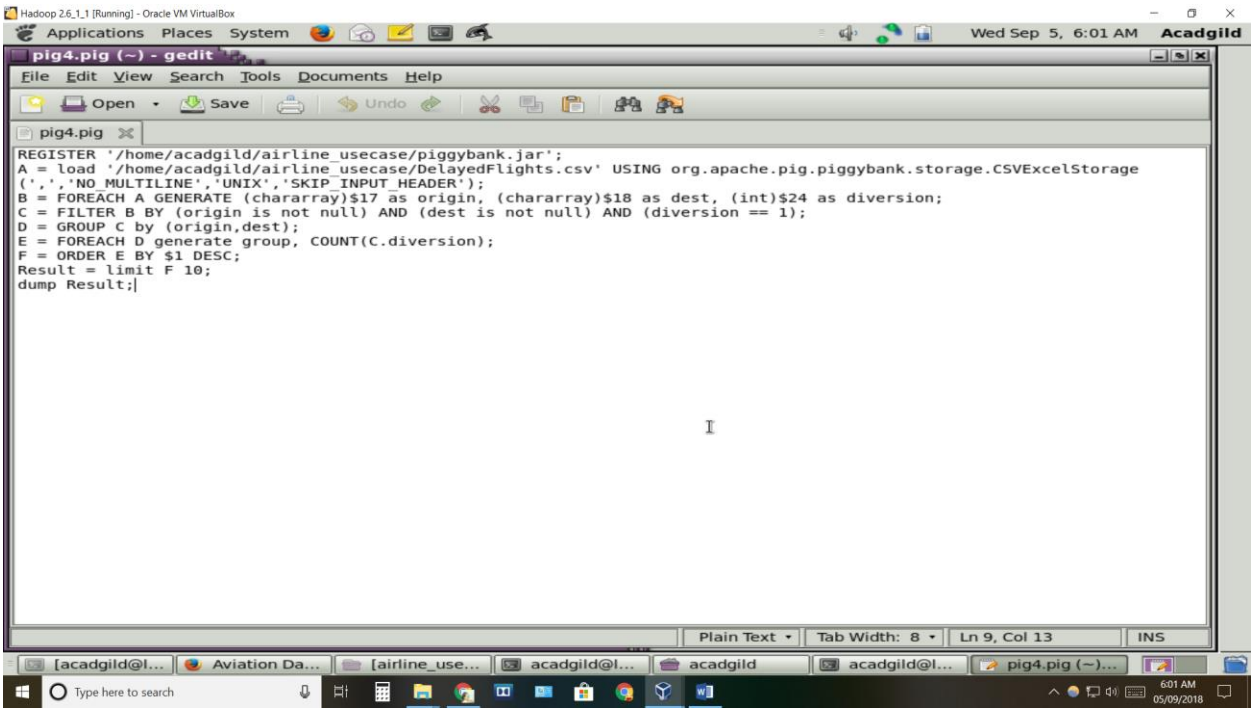
Pig –x local pig3.pig



Problem Statement 4

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

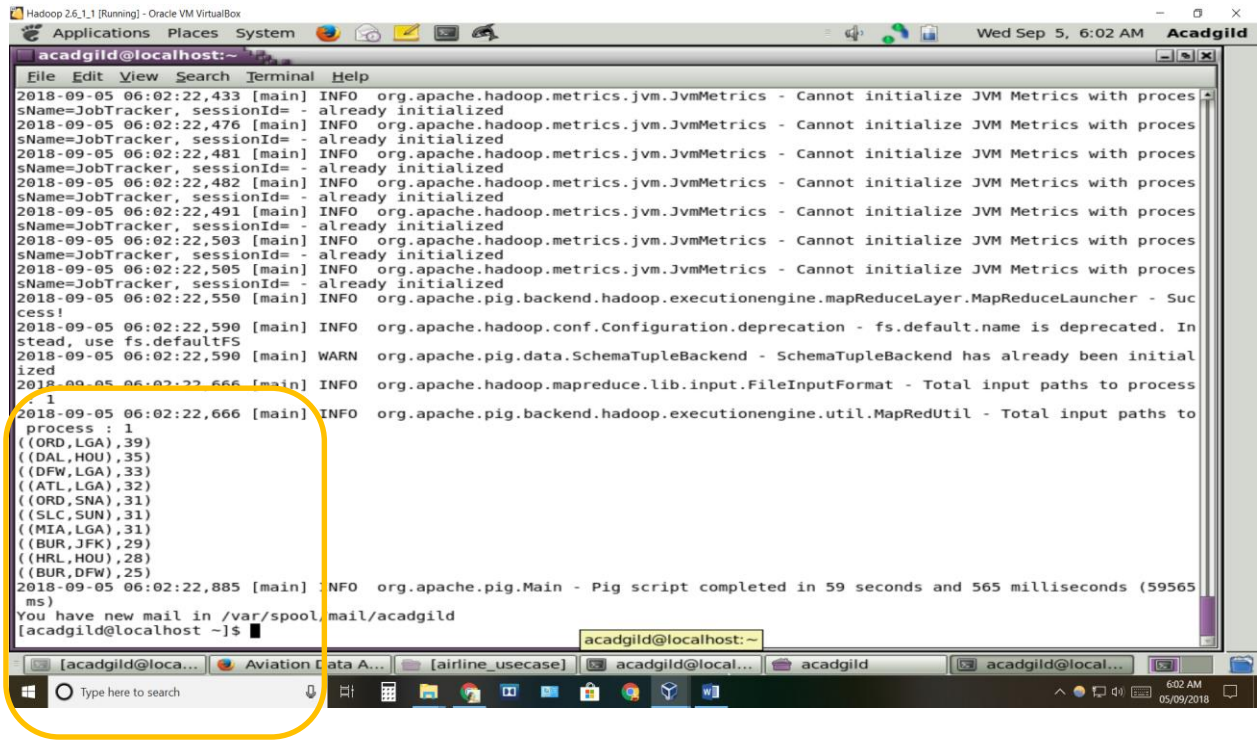
1. Wrote Pig Script under and saved under pig4.pig



Steps

- **In First Line**: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.
- In relation **A**, we are loading the dataset using CSVExcelStorage.
- In relation **B**, we are generating the columns that are required for processing and typecasting each of them like int,chararray.
- In relation **C**, we are filtering the data based on "not null" and diversion =1. This will remove the null records, if any, and give the data corresponding to the diversion taken.
- In relation **D**, we are grouping the data based on origin and destination.
- Relation **D** finds the count of diversion taken per unique origin and destination.
- Relations **F** and **Result** orders the result and produces top 10 results with LIMIT 10;
- DUMP Result will print results as shown below;

2. 2.Ran PigScript in local Mode
   Pig –x local pig4.pig


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Wed Sep 5, 6:02 AM
Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
2018-09-05 06:02:22,433 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,476 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,481 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,482 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,491 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,503 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,505 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 06:02:22,550 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Suc
cess!
2018-09-05 06:02:22,590 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. In
stead, use fs.defaultFS
2018-09-05 06:02:22,590 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initial
ized
2018-09-05 06:02:22,666 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2018-09-05 06:02:22,666 [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)
2018-09-05 06:02:22,885 [main] INFO org.apache.pig.Main - Pig script completed in 59 seconds and 565 milliseconds (59565
ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
acadgild@localhost:~
[acadgild@loca...
Aviation Data A...
[airline_usecase]
acadgild@local...
acadgild
acadgild@local...
Type here to search
6:02 AM
05/09/2018