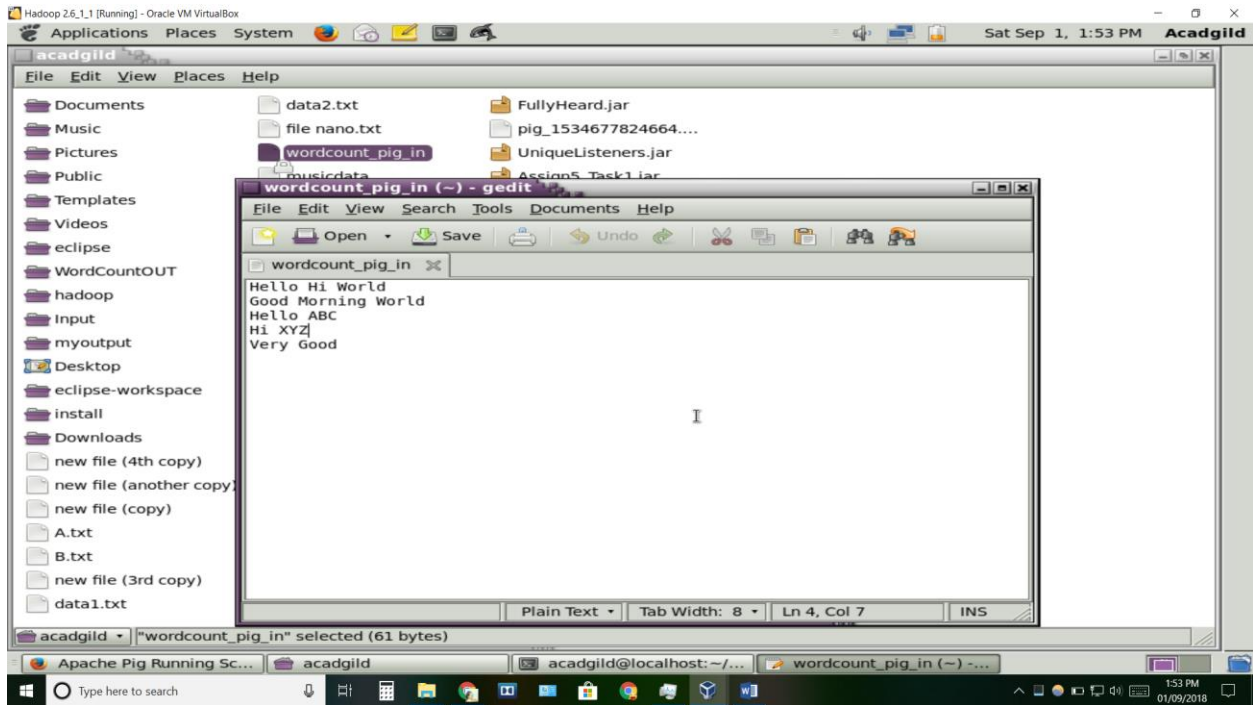


Big Data Hadoop 'Assignment Seven'

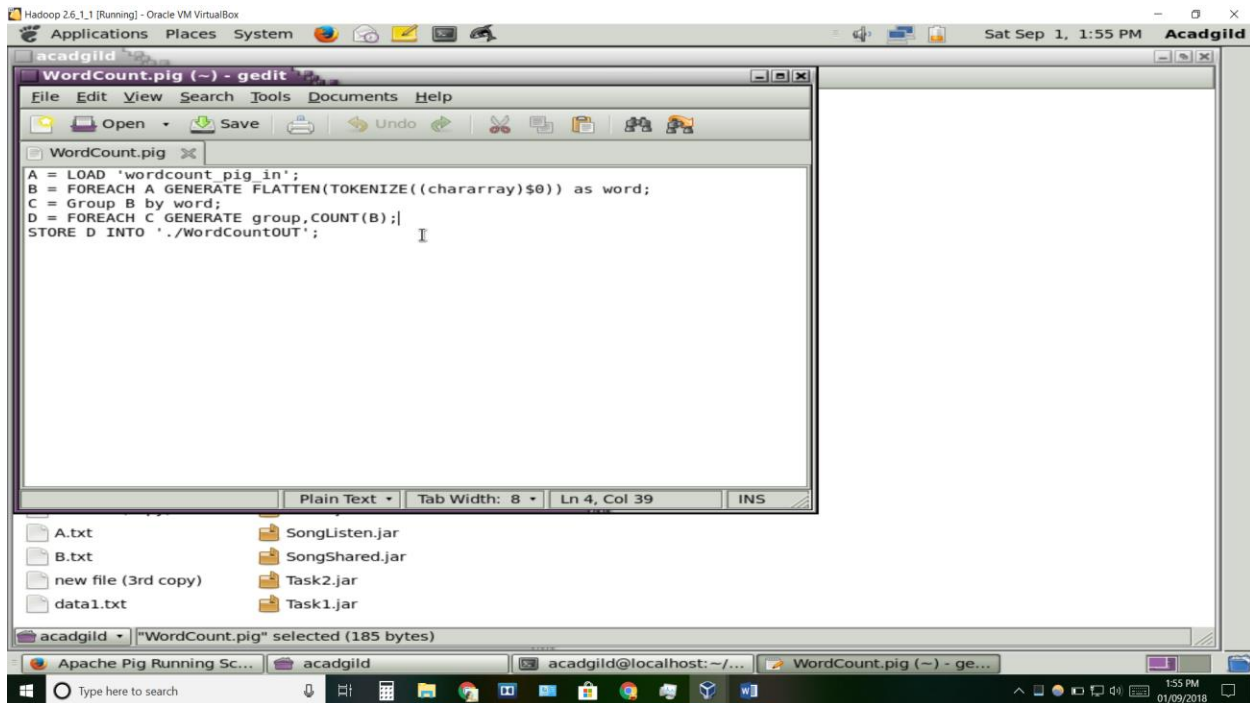
Write PIG scripts for following tasks.

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

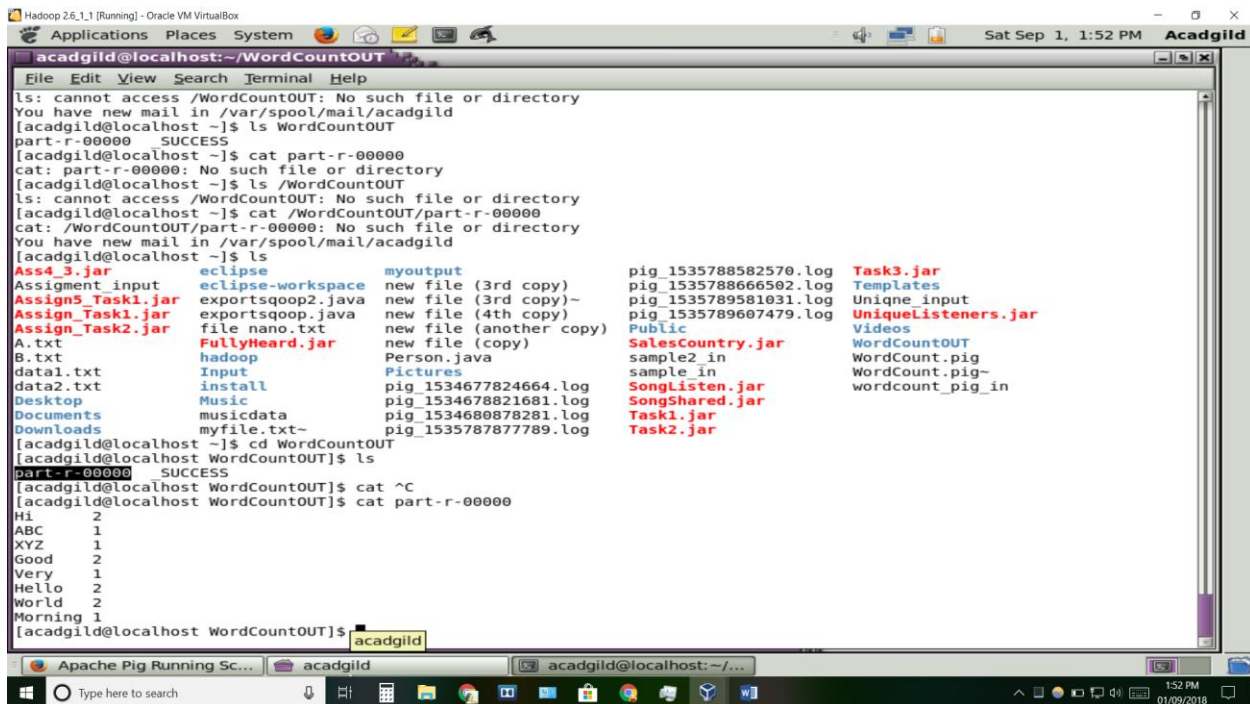
- Executed in pig LOCAL mode
- Input data: Wordcount_pig_in file



- WordCount pig script



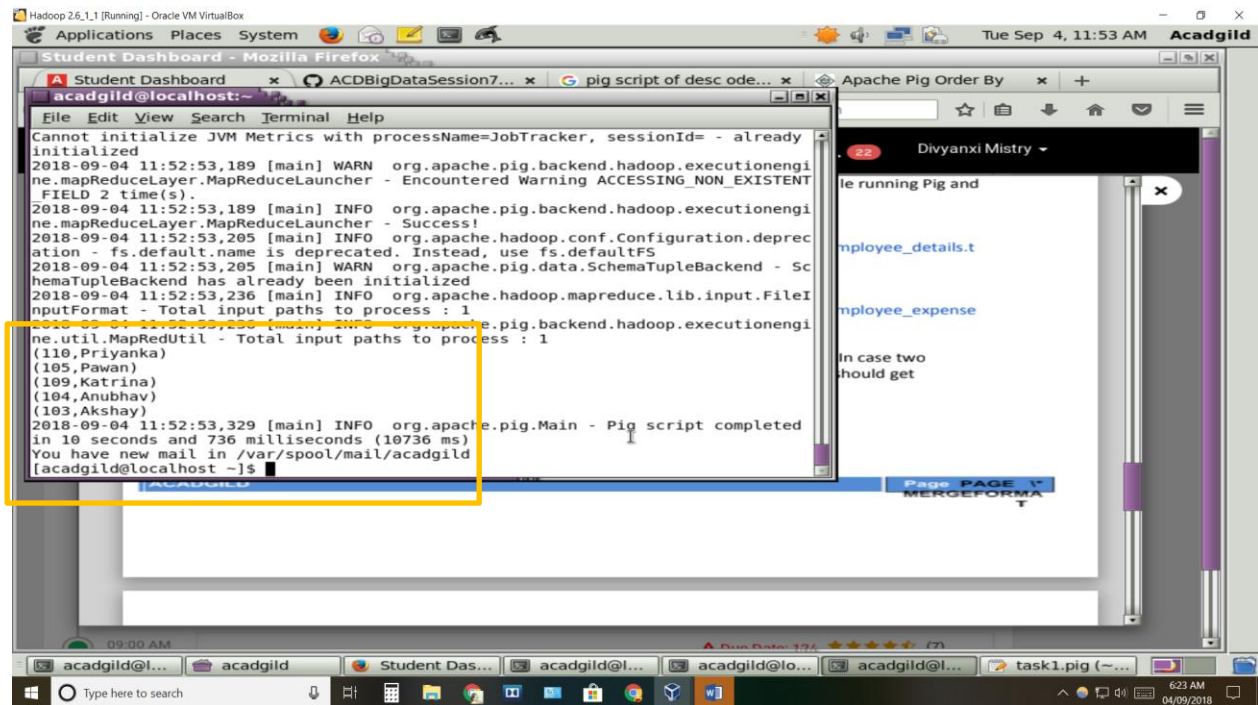
- Final Output



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)



```
acacgild@localhost:~$ pig script of desc ode...
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-04 11:52:53,189 [main] WARN org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT
FIELD 2 time(s).
2018-09-04 11:52:53,189 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-09-04 11:52:53,205 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-04 11:52:53,205 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2018-09-04 11:52:53,236 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2018-09-04 11:52:53,236 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
(105,Pawan)
(109,Katrina)
(104,Anubhav)
(103,Akshay)
2018-09-04 11:52:53,329 [main] INFO org.apache.pig.Main - Pig script completed
in 10 seconds and 736 milliseconds (10736 ms)
You have new mail in /var/spool/mail/acacgild
[acacgild@localhost ~]$
```

(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)

```
2018-09-04 11:38:41,157 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 11:38:41,159 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 11:38:41,167 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 11:38:41,207 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 11:38:41,209 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 11:38:41,210 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 11:38:41,230 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-04 11:38:41,246 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-04 11:38:41,246 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 11:38:41,272 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(103,Akshay)
(105,Pawan)
2018-09-04 11:38:41,373 [main] INFO org.apache.pig.Main - Pig script completed
in 8 seconds and 268 milliseconds (8268 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

(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)

```
emp = LOAD 'employee_details.txt' USING PigStorage(',') AS (emp_id:int, emp_name:chararray, emp_salary:int);
empexpense = LOAD 'employee_expenses.txt' USING PigStorage(',') AS (emp_id:int, emp_expense:int);

Joinempexpense = join emp by emp_id,empexpense by emp_id;
maxexpense = ORDER Joinempexpense by empexpense::emp_expense desc;

Limitmaxexpnse = LIMIT maxexpense 1;
Limitmaxexpensefinal = foreach Limitmaxexpnse generate emp::emp_id,emp::emp_name;

dump Limitmaxexpensefinal;
```

```
2018-09-04 12:03:52,074 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT FIELD 1 time(s).
2018-09-04 12:03:52,074 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-04 12:03:52,093 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-04 12:03:52,093 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 12:03:52,119 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
2018-09-04 12:03:52,229 [main] INFO org.apache.pig.Main - Pig script completed
in 16 seconds and 27 milliseconds (16027 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```


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

```
emp = LOAD 'employee_details.txt' USING PigStorage(',') AS (emp_id:int, emp_name:chararray, emp_salary:int, emp_rating:int);
emp_expenses = LOAD 'employee_expenses.txt' USING PigStorage(',') AS (emp_id:int, expenses:int);
emp_with_exp = JOIN emp BY emp_id, emp_expenses BY emp_id;
emp_with_exp_data = FOREACH emp_with_exp GENERATE emp::emp_id, emp::emp_name;
emp_with_exp_distinct_data = DISTINCT emp_with_exp_data;
dump emp_with_exp_distinct_data;
```

```
acadgild@localhost:~$
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-04 06:44:33,340 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 1 time(s).
2018-09-04 06:44:33,341 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-04 06:44:33,368 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-04 06:44:33,369 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 06:44:33,416 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 06:44:33,415 [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)
2018-09-04 06:44:33,605 [main] INFO org.apache.pig.Main - Pig script completed in 14 seconds and 312 milliseconds (14312 ms)
acacagild@localhost ~$
```

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

```
emp = LOAD 'employee_details.txt' USING PigStorage(',') AS (emp_id:int, emp_name:chararray, emp_salary:int, emp_rating:int);
emp_expenses = LOAD 'employee_expenses.txt' USING PigStorage(',') AS (emp_id:int, expenses:int);
emp_without_exp = JOIN emp BY emp_id LEFT OUTER, emp_expenses BY emp_id;
emp_without_exp_filter = FILTER emp_without_exp BY emp_expenses::emp_id is null;
emp_without_exp_filter_data = FOREACH emp_without_exp_filter GENERATE emp::emp_id, emp::emp_name;
DUMP emp_without_exp_filter_data;
```

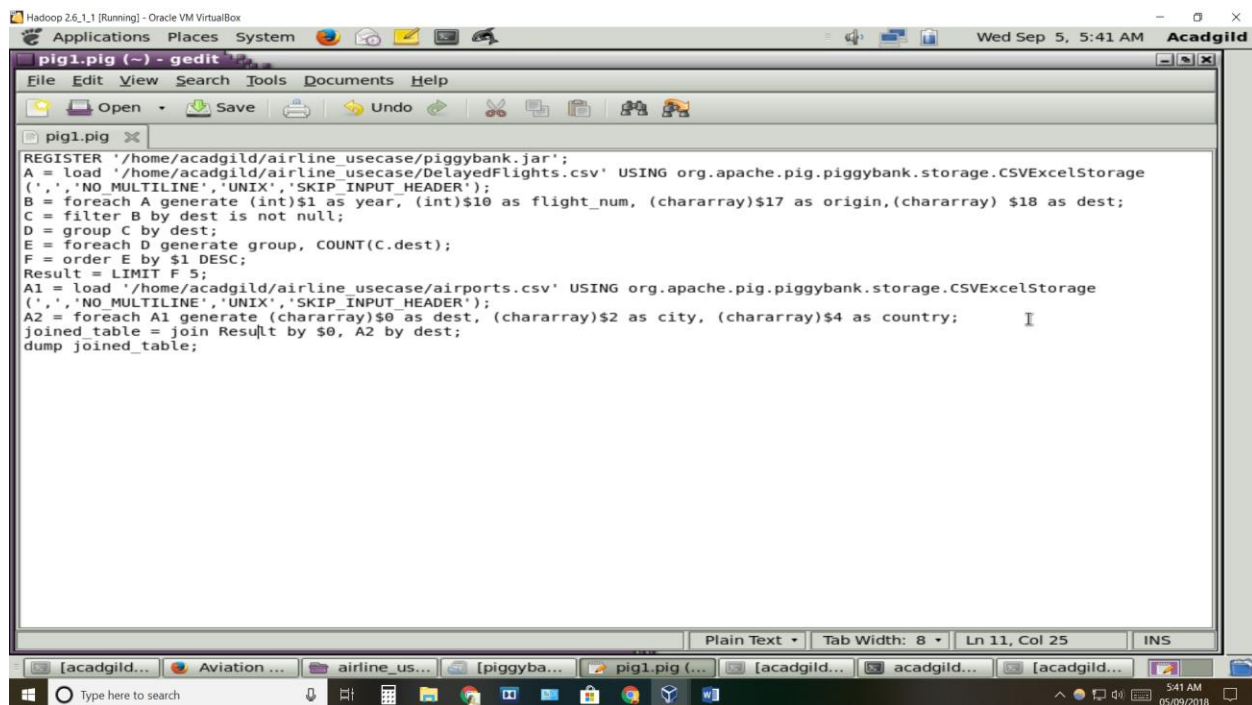
```
acacagild@localhost:~$
FIELD 1 time(s).
2018-09-04 06:49:20,308 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-04 06:49:20,318 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-04 06:49:20,319 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-04 06:49:20,361 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-04 06:49:20,381 [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)
(.)
2018-09-04 06:49:20,351 [main] INFO org.apache.pig.Main - Pig script completed in 11 seconds and 495 milliseconds (11495 ms)
You have new mail in /var/spool/mail/acacagild
acacagild@localhost ~$
```

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



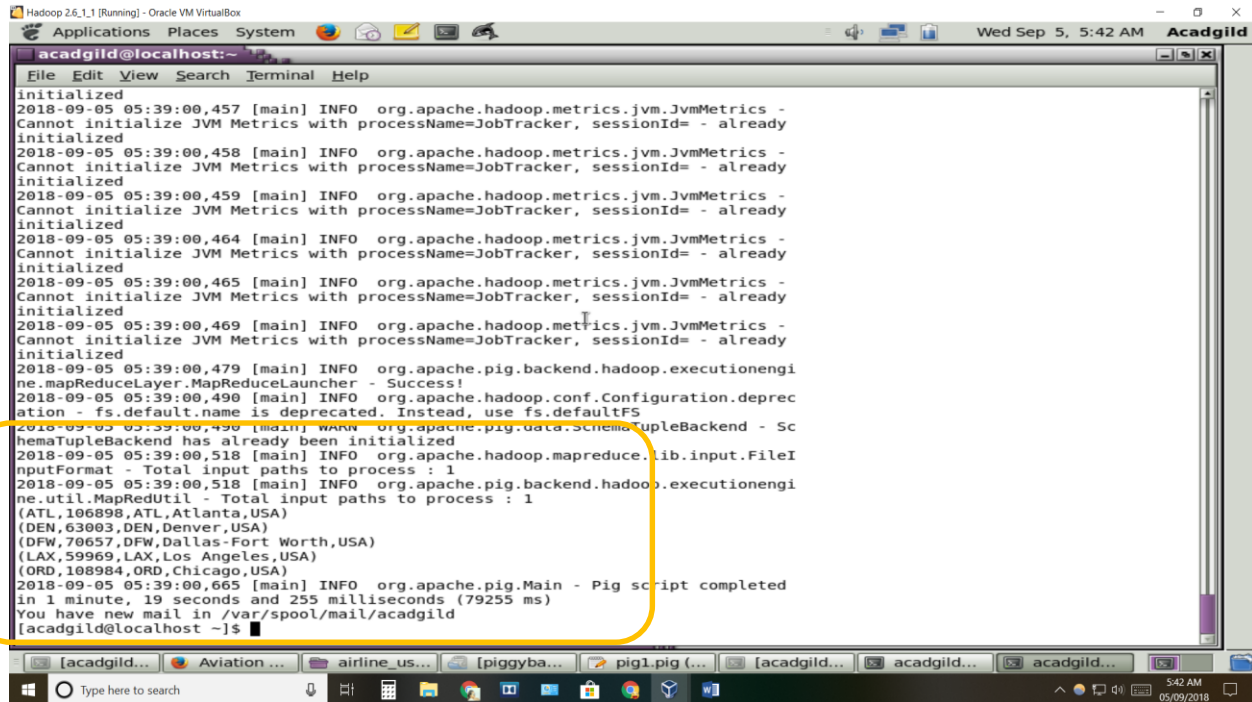
```
REGISTER '/home/acadgild/airline_usecase/piggybank.jar';
A = load '/home/acadgild/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 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/airline_usecase/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;
```

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

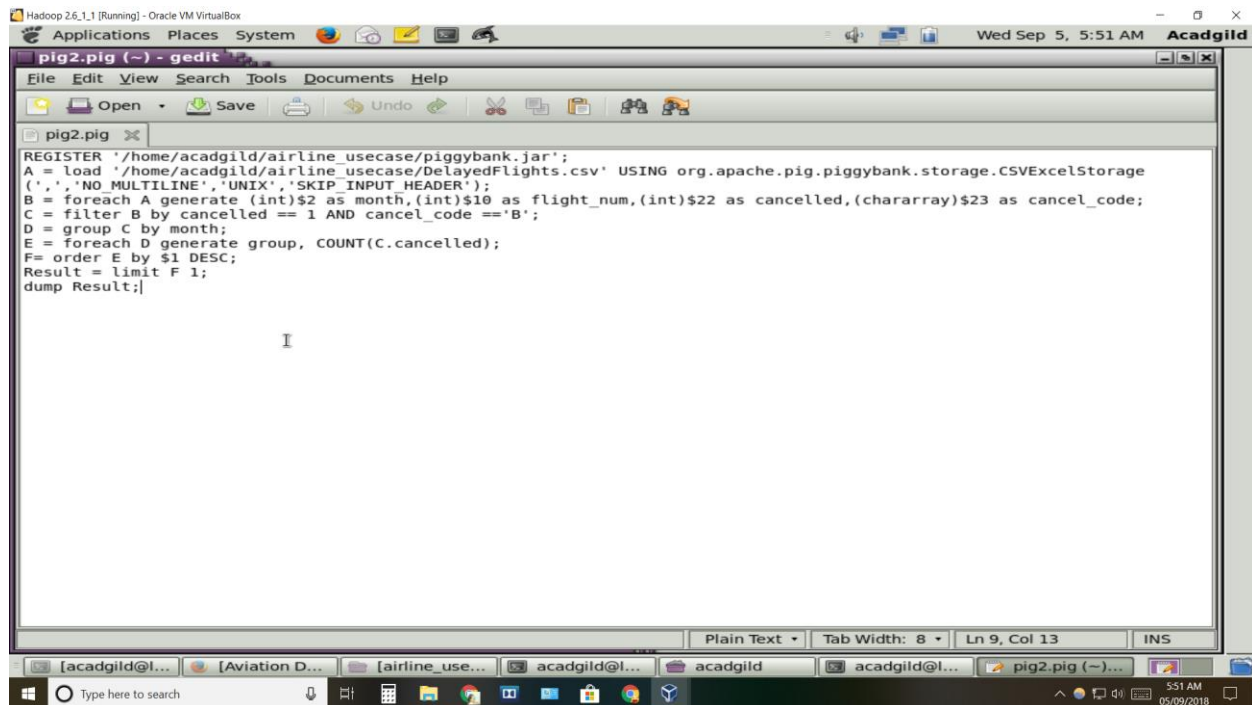


```
acadmild@localhost:~$ pig -x local pig1.pig
2018-09-05 05:39:00,457 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-05 05:39:00,458 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-05 05:39:00,459 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-05 05:39:00,464 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-05 05:39:00,465 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-05 05:39:00,469 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-09-05 05:39:00,479 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-09-05 05:39:00,490 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-05 05:39:00,490 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2018-09-05 05:39:00,518 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2018-09-05 05:39:00,518 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.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)
2018-09-05 05:39:00,665 [main] INFO org.apache.pig.Main - Pig script completed
in 1 minute, 19 seconds and 255 milliseconds (79255 ms)
You have new mail in /var/spool/mail/acadmild
acadmild@localhost ~$
```

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

A screenshot of a virtual machine window titled 'Hadoop 2.6.1.1 [Running] - Oracle VM VirtualBox'. The window shows a desktop environment with a taskbar at the bottom. The main application is a text editor named 'gedit' with the file 'pig2.pig' open. The script contains Pig Latin code for loading a dataset, filtering, grouping, and ordering. The status bar at the bottom of the editor shows 'Plain Text', 'Tab Width: 8', 'Ln 9, Col 13', and 'INS' mode. The taskbar includes icons for various applications and the system clock shows '5:51 AM 05/09/2018'.

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

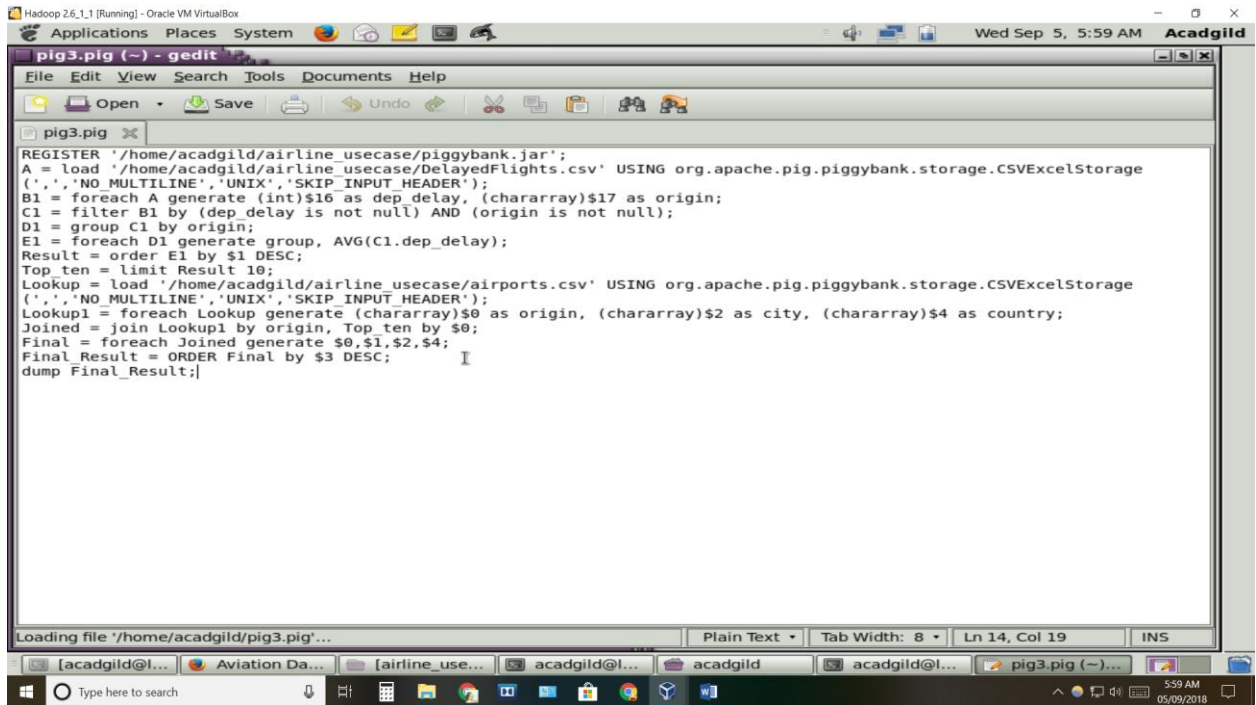
Pig -x local pig2.pig


```
2018-09-05 05:47:58,699 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,714 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,714 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,718 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,743 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,744 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,745 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-09-05 05:47:58,766 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-09-05 05:47:58,779 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-09-05 05:47:58,779 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-09-05 05:47:58,814 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-05 05:47:58,814 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
2018-09-05 05:47:58,886 [main] INFO org.apache.pig.Main - Pig script completed in 40 seconds and 746 milliseconds (40746 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ dump Result;
bash: dump: command not found
[acadgild@localhost ~]$
```

Problem Statement 3

Top ten origins with the highest AVG departure delay

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



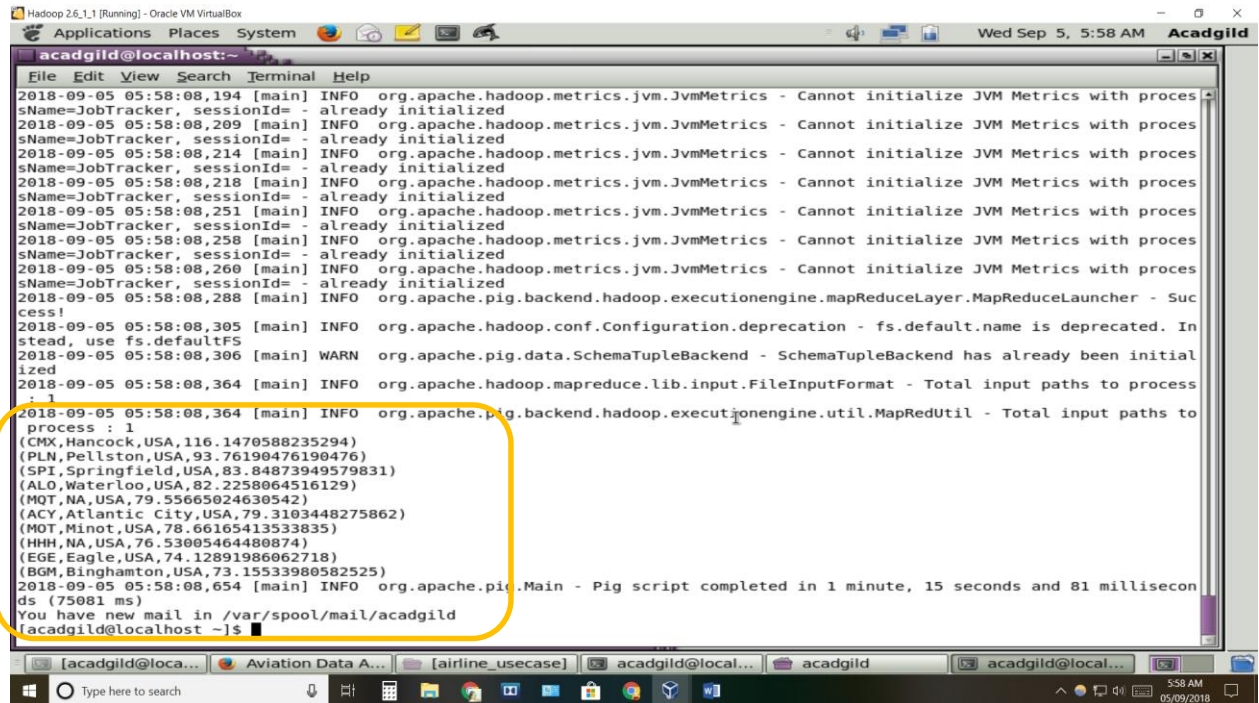
```
REGISTER '/home/acadgild/airline_usecase/piggybank.jar';
A = load '/home/acadgild/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
D1 = group C1 by origin;
E1 = foreach D1 generate group, AVG(C1.dep_delay);
Result = order E1 by $1 DESC;
Top_ten = limit Result 10;
Lookup = load '/home/acadgild/airline_usecase/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
Joined = join Top_ten by origin, Lookup1 by $0;
Final = foreach Joined generate $0,$1,$2,$4;
Final_Result = ORDER Final by $3 DESC;
dump Final_Result;
```

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 **Lookup1**, 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

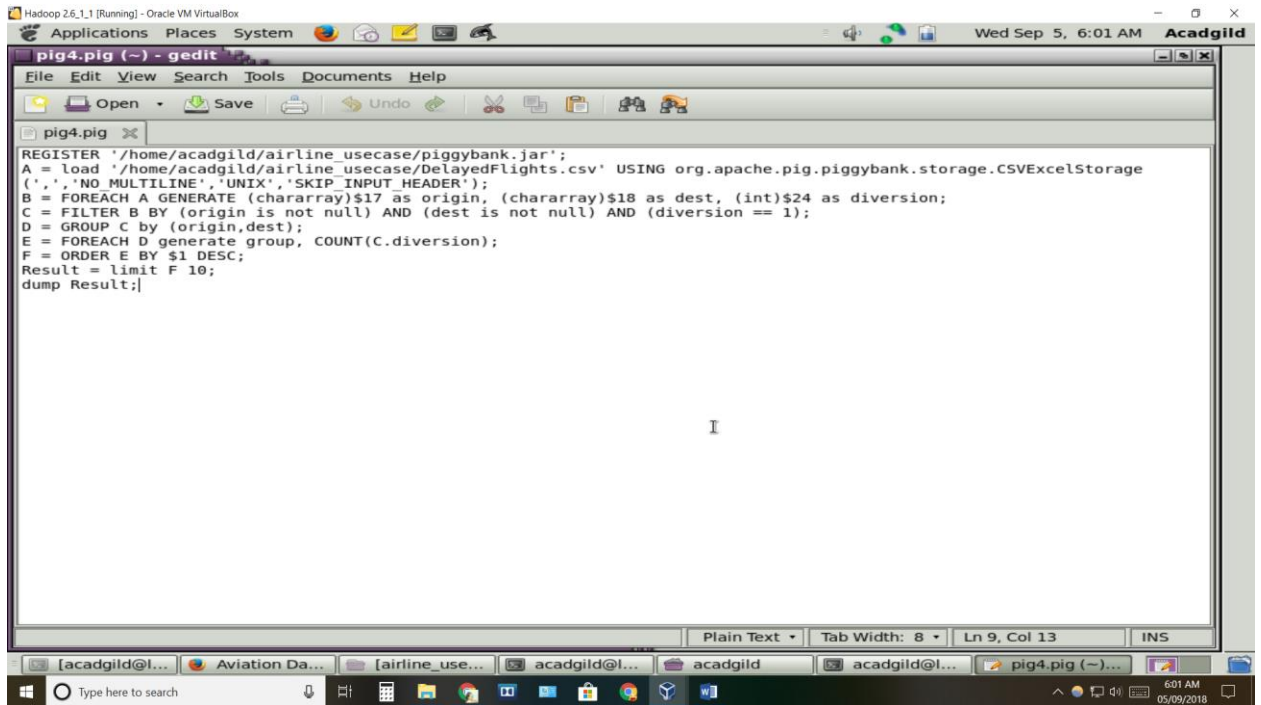


```
acadmild@localhost:~$ pig -x local pig3.pig
2018-09-05 05:58:08,194 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,209 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,214 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,218 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,251 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,258 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,260 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with proces
sName=JobTracker, sessionId= - already initialized
2018-09-05 05:58:08,288 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Suc
cess!
2018-09-05 05:58:08,305 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. In
stead, use fs.defaultFS
2018-09-05 05:58:08,306 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initial
ized
2018-09-05 05:58:08,364 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process
: 1
2018-09-05 05:58:08,364 [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)
2018-09-05 05:58:08,654 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 15 seconds and 81 millisecon
ds (75081 ms)
You have new mail in /var/spool/mail/acadmild
acadmild@localhost ~$
```

Problem Statement 4

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

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



```
REGISTER '/home/acadgild/airline_usecase/piggybank.jar';
A = load '/home/acadgild/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
D = GROUP C by (origin,dest);
E = FOREACH D generate group, COUNT(C.diversion);
F = ORDER E BY $1 DESC;
Result = limit F 10;
dump Result;
```

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. 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

acadmild@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/acadmild
[acadmild@localhost ~]$
```

acadmild@localhost:~

[acadmild@loca... Aviation Data A... [airline_usecase] acadmild@local... acadmild acadmild@local...

Type here to search

6:02 AM 05/09/2018