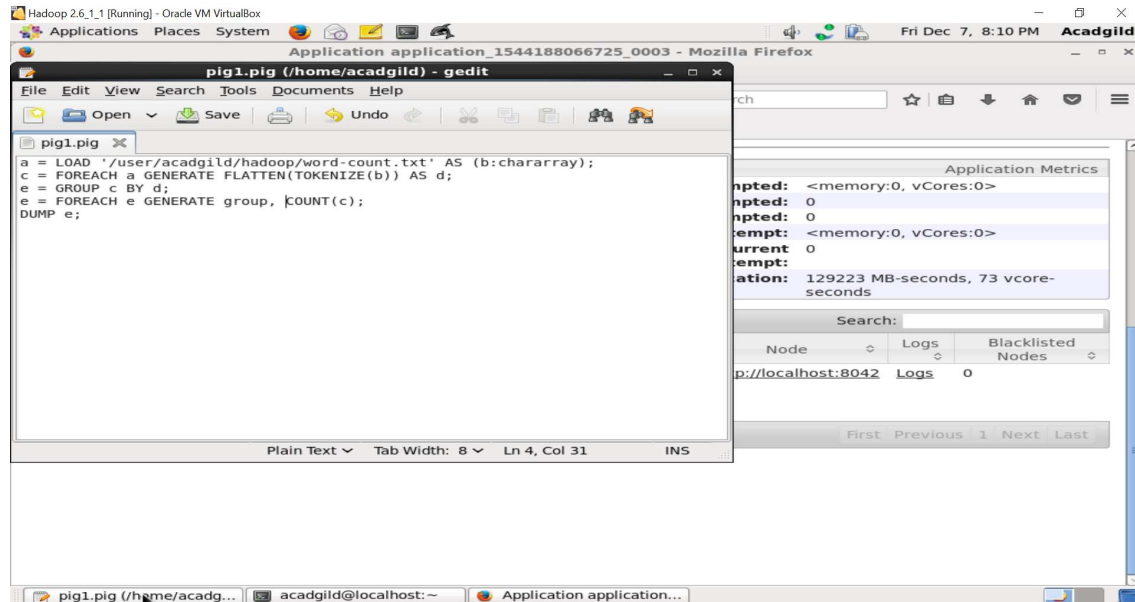


Vijay's Assignment – Pig 1

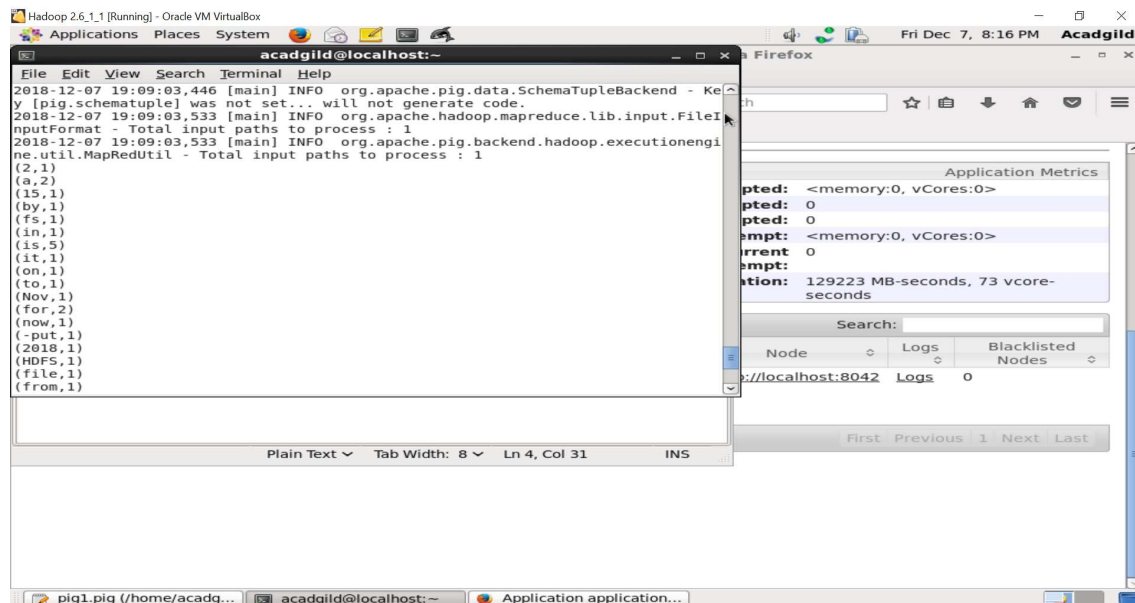
Task 1

Write a program to implement wordcount using Pig.

- 1) A new program pig1.pig was created in /home/acadgild directory in local



- 2) This program Loads the input file word-count.txt present inside the hdfs path given above
- 3) Once the data is loaded, the words in each line are separated using TOKENIZE and they are combined as a tuple using FLATTEN command to remove the level of nesting
- 4) The data is then grouped by each word and each word the count is generated using another foreach statement
- 5) The final data is written to output using DUMP command



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:

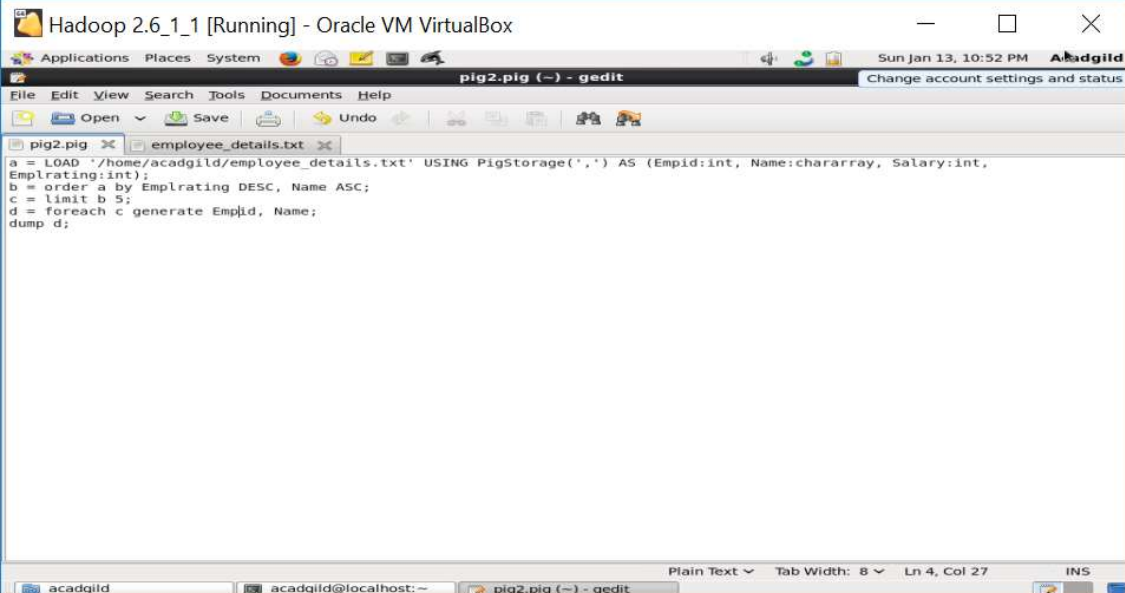
employee_details (EmpID,Name,Salary,EmployeeRating)

https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_details.txt

employee_expenses(EmpID,Expense)

https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_expense_s.txt

(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)



The screenshot shows a virtual machine window titled 'Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox'. Inside the VM, a terminal window is open with a gedit editor. The editor has two tabs: 'pig2.pig' and 'employee_details.txt'. The 'pig2.pig' tab is active and contains the following Pig Latin script:

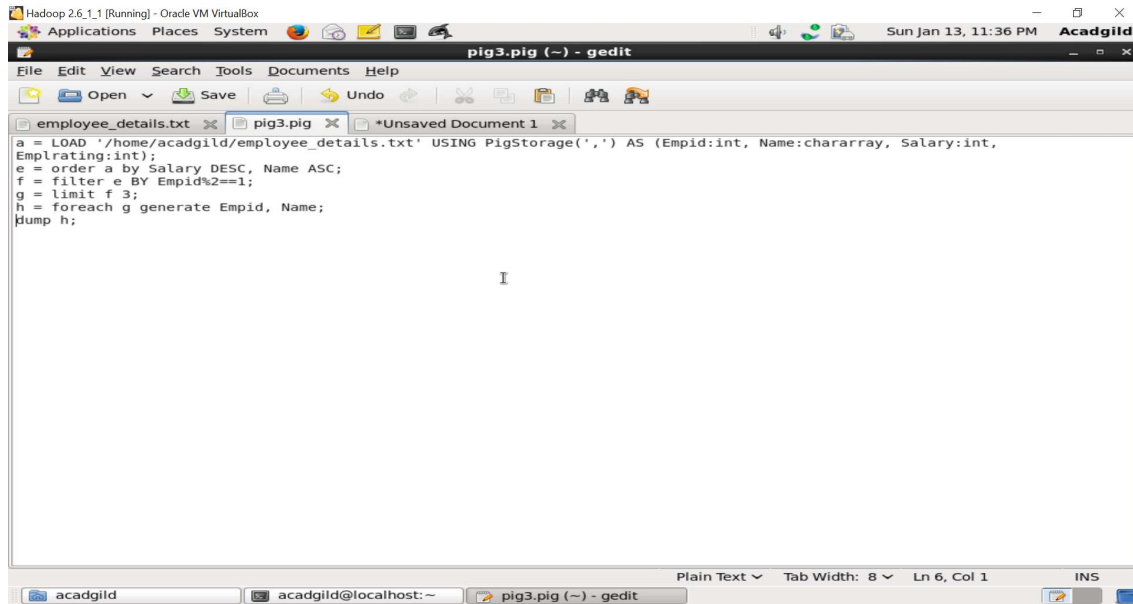
```
a = LOAD '/home/acadgild/employee_details.txt' USING PigStorage(',') AS (Empid:int, Name:chararray, Salary:int, Emprating:int);
b = order a by Emprating DESC, Name ASC;
c = limit b 5;
d = foreach c generate Empid, Name;
dump d;
```

The status bar at the bottom of the gedit window shows 'Plain Text', 'Tab Width: 8', 'Ln 4, Col 27', and 'INS'.

Output

```
2019-01-13 22:50:05,861 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan)
(110,Priyanka)
(104,Anubhav)
(109,Katrina)
(103,Akshay)
```

(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)



```

a = LOAD '/home/acadgild/employee_details.txt' USING PigStorage(',') AS (Empid:int, Name:chararray, Salary:int,
    Emplating:int);
e = order a by Salary DESC, Name ASC;
f = filter e BY Empid%2==1;
g = limit f 3;
h = foreach g generate Empid, Name;
Dump h;

```

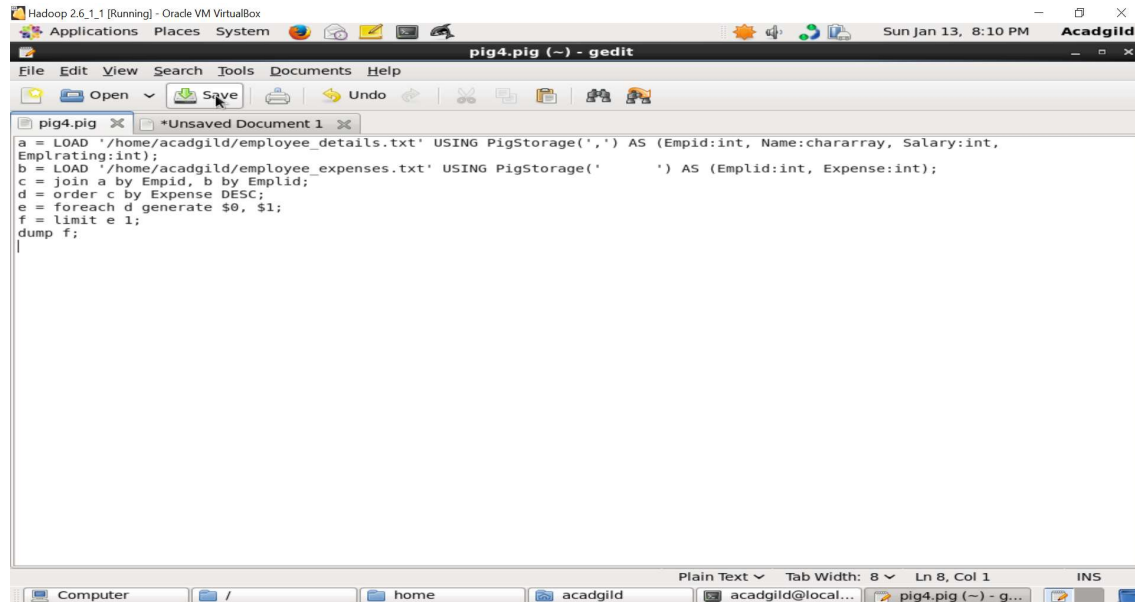
Output

```

2019-01-13 23:32:37,615 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)

```

(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)

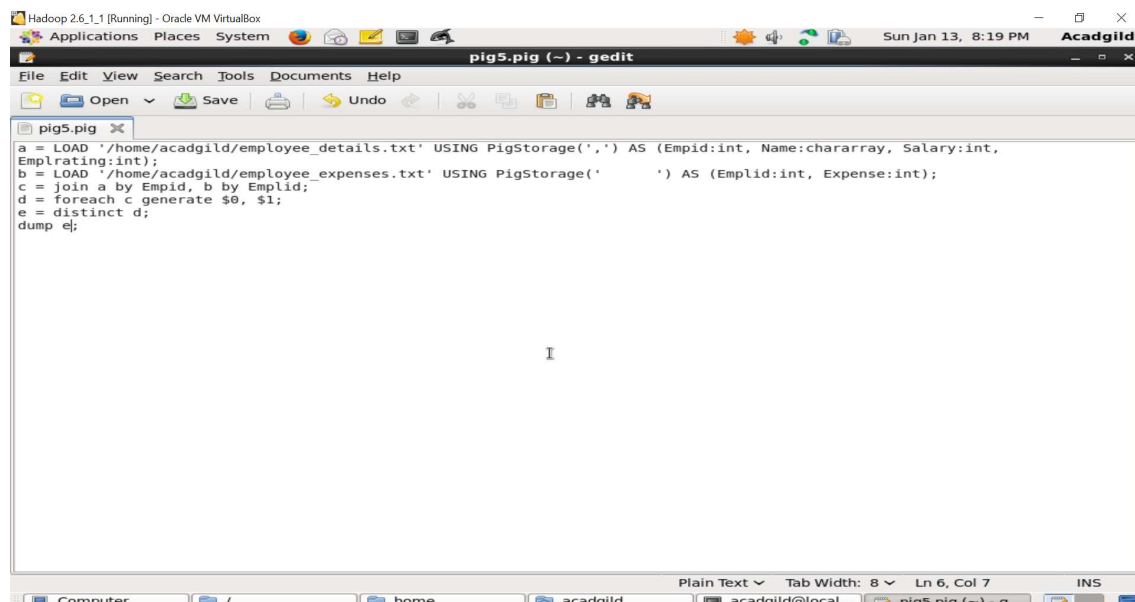


```
a = LOAD '/home/acadgild/employee_details.txt' USING PigStorage(',') AS (Empid:int, Name:chararray, Salary:int, Emplrating:int);
b = LOAD '/home/acadgild/employee_expenses.txt' USING PigStorage(',') AS (Emplid:int, Expense:int);
c = join a by Empid, b by Emplid;
d = order c by Expense DESC;
e = foreach d generate $0, $1;
f = limit e 1;
dump f;
```

Output

```
2019-01-13 20:10:22,997 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
```

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



```
a = LOAD '/home/acadgild/employee_details.txt' USING PigStorage(',') AS (Empid:int, Name:chararray, Salary:int, Emplrating:int);
b = LOAD '/home/acadgild/employee_expenses.txt' USING PigStorage(',') AS (Emplid:int, Expense:int);
c = join a by Empid, b by Emplid;
d = foreach c generate $0, $1;
e = distinct d;
dump e;
```

```

2019-01-13 20:14:16,064 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(102,Sharukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)

```

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

```

a = LOAD '/home/acadgild/employee_details.txt' USING PigStorage(',') AS (Empid:int, Name:chararray, Salary:int, Emplrating:int);
b = LOAD '/home/acadgild/employee_expenses.txt' USING PigStorage(',') AS (Emplid:int, Expense:int);
c = join a by Emplid LEFT, b by Emplid;
d = filter c by $4 is null;
dump d;

```

Output

```

2019-01-13 20:33:07,137 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Akshay,11000,3,,)
(106,Aamir,25000,1,,)
(107,Salman,17500,2,,)
(108,Ranbir,14000,3,,)
(109,Katrina,1000,4,,)
(111,Tushar,500,1,,)
(112,Akay,5000,2,,)
(113,Jubeen,1000,1,,)

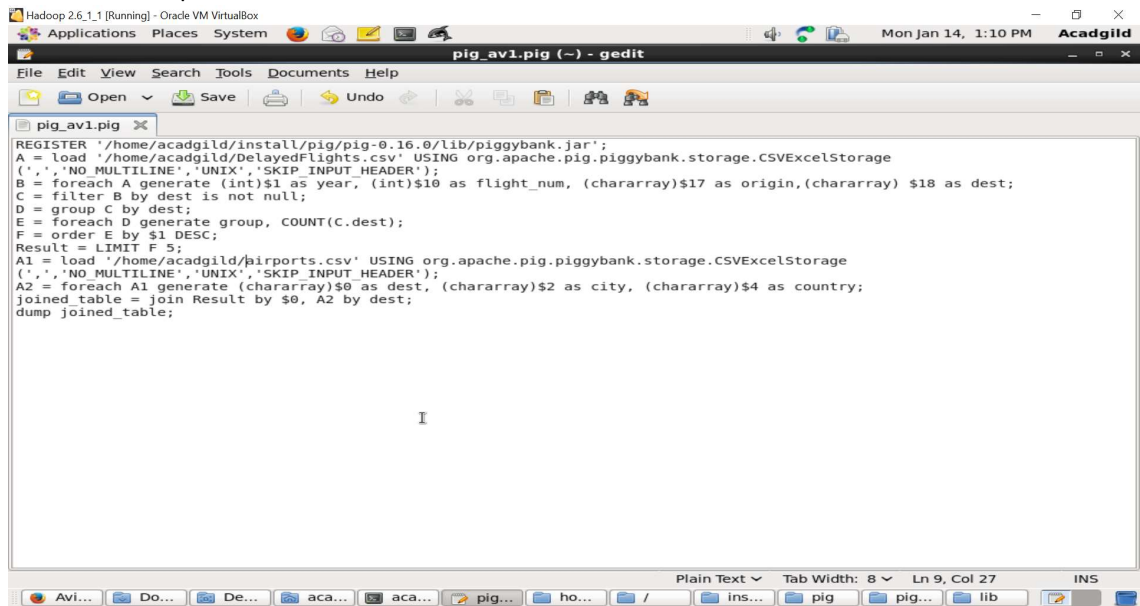
```

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) Find out the top 5 most visited destinations.

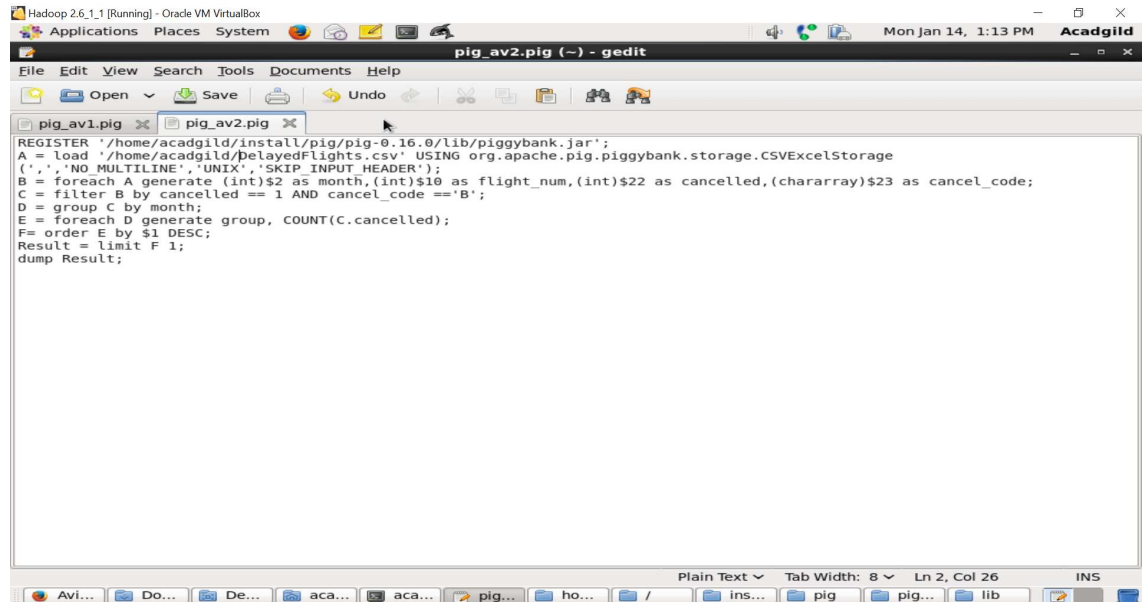


```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as 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/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;
```

Output:

```
2019-01-14 13:09:15,284 [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)
```

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

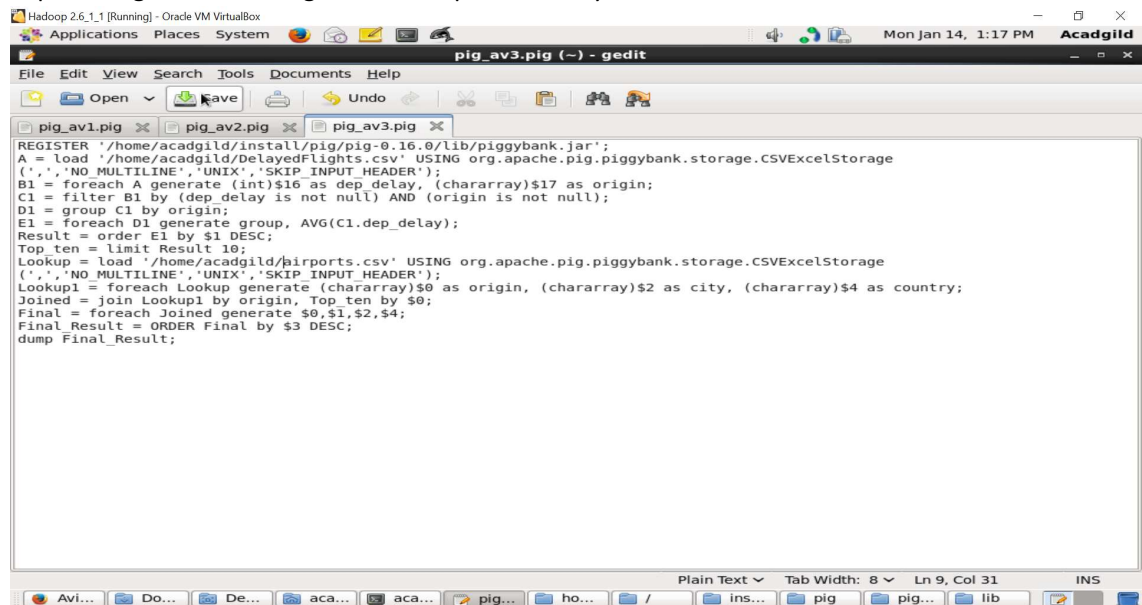


```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO MULTILINE','UNIX','SKIP INPUT HEADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
C = filter B by cancelled == 1 AND cancel_code == 'B';
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F = order E by $1 DESC;
Result = limit F 1;
dump Result;
```

2019-01-14 13:13:30,633 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
2019-01-14 13:13:30,753 [main] INFO org.apache.pig.Main - Pig script completed in 30 seconds and 800 milliseconds (30800 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]\$

Output

3) Top ten origins with the highest AVG departure delay



```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
A = load '/home/acadgild/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/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 Lookup1 by origin, Top ten by $0;
Final = foreach Joined generate $0,$1,$2,$4;
Final Result = ORDER Final by $3 DESC;
dump Final Result;
```

2019-01-14 13:13:30,633 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
2019-01-14 13:13:30,753 [main] INFO org.apache.pig.Main - Pig script completed in 30 seconds and 800 milliseconds (30800 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]\$

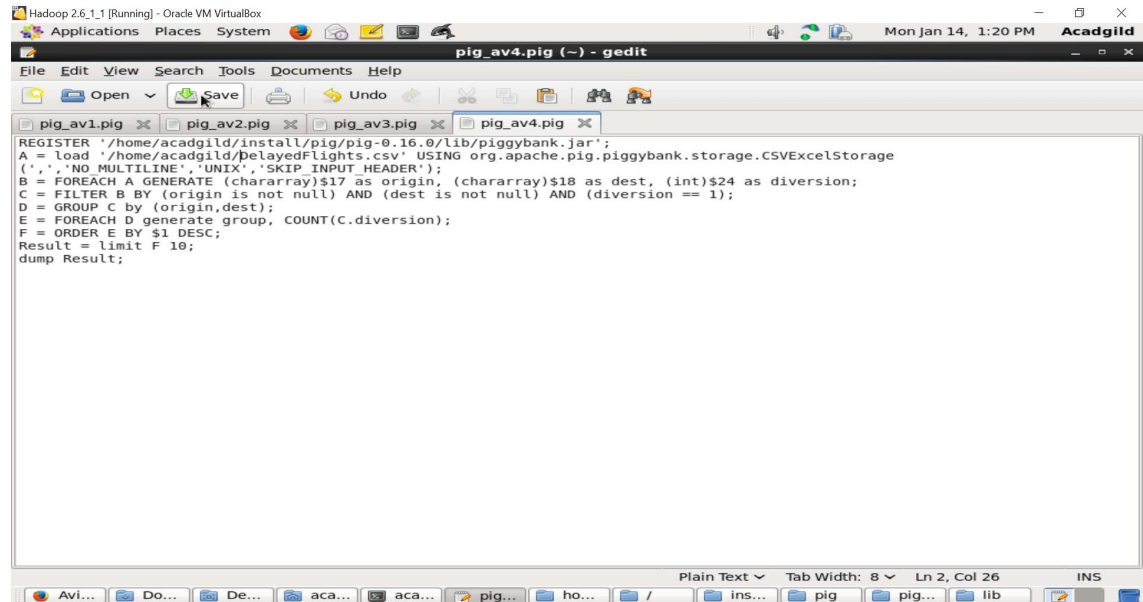
Output

```

2019-01-14 13:17:44,375 [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)
2019-01-14 13:17:44,510 [main] INFO org.apache.pig.Main - Pig script completed
in 46 seconds and 446 milliseconds (46446 ms)

```

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



Output

```

2019-01-14 13:20:53,757 [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)

```