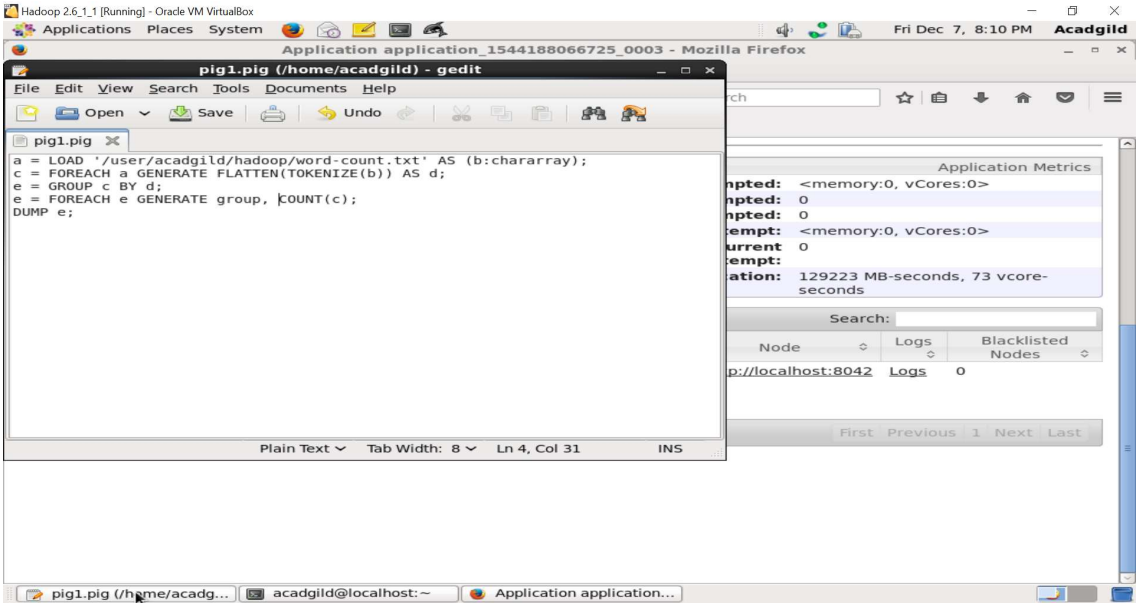
Vijay's Assignment – Pig 1

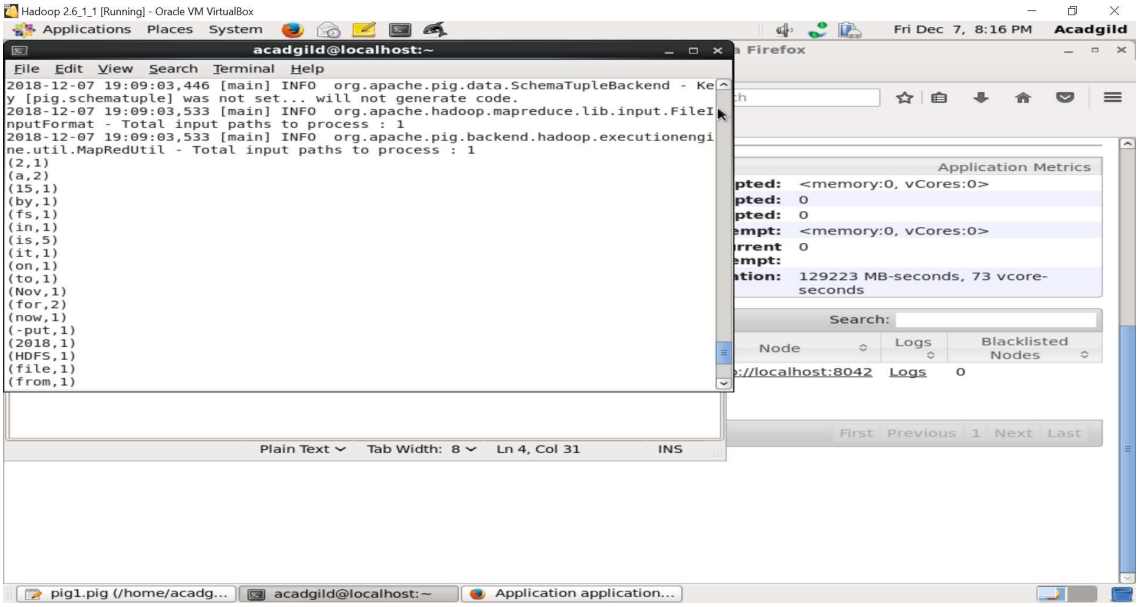
Task 1

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:

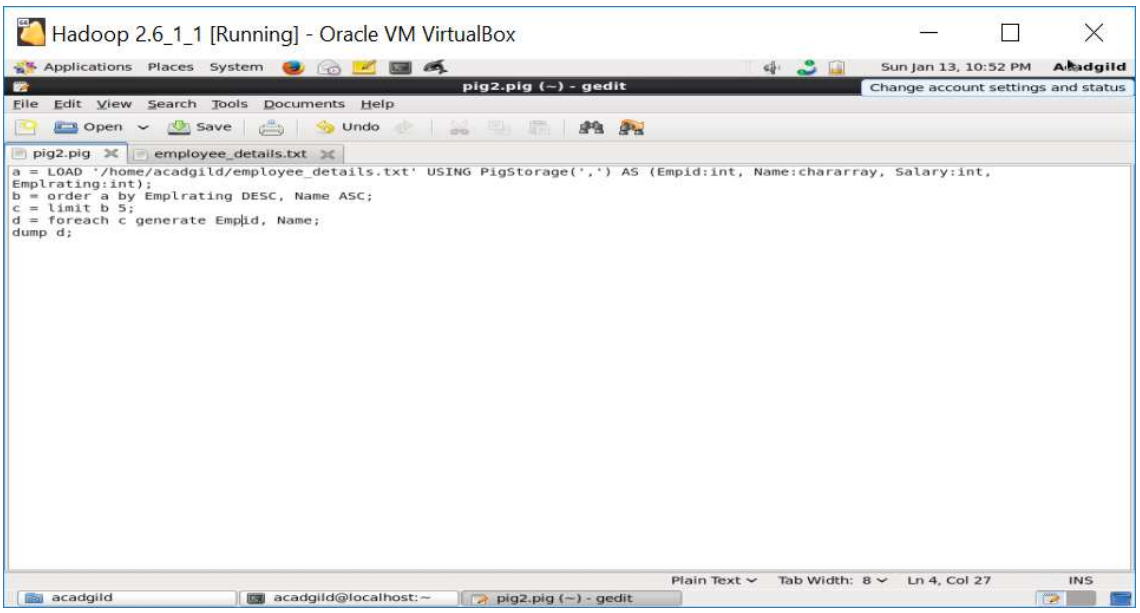
employee_details (EmpID,Name,Salary,EmployeeRating)

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

employee_expenses(EmpID,Expence)

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

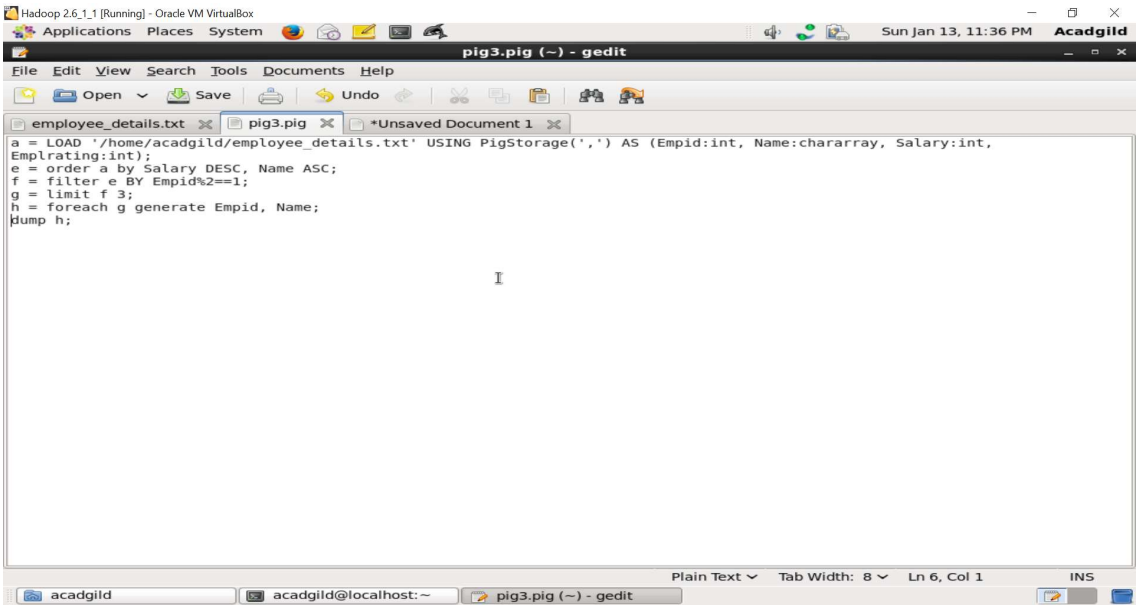
(a) Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)



Output

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

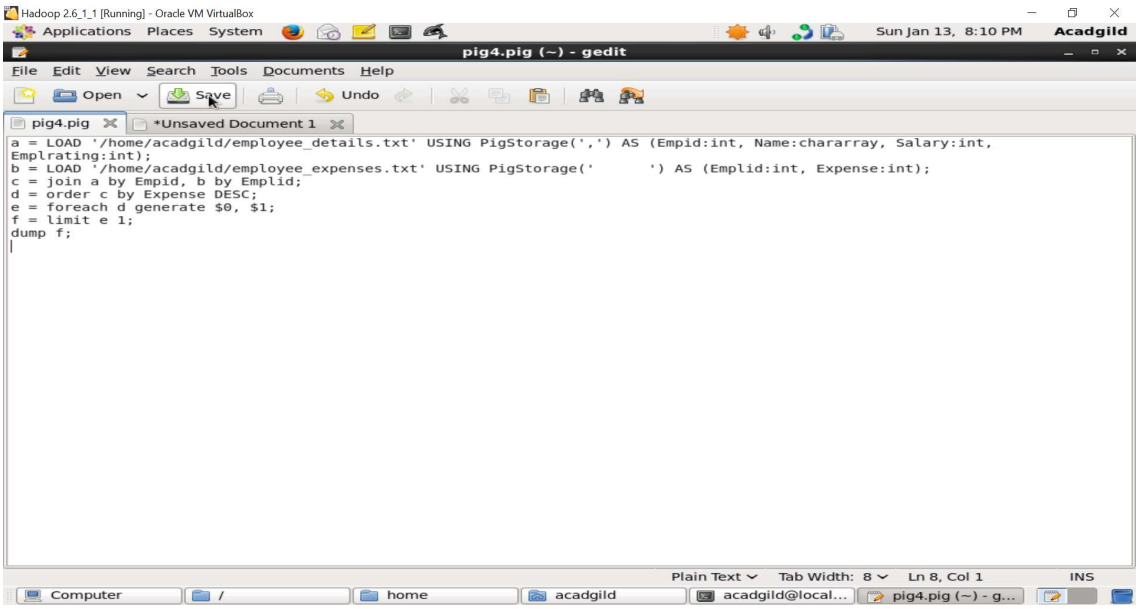
(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference)



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)
```

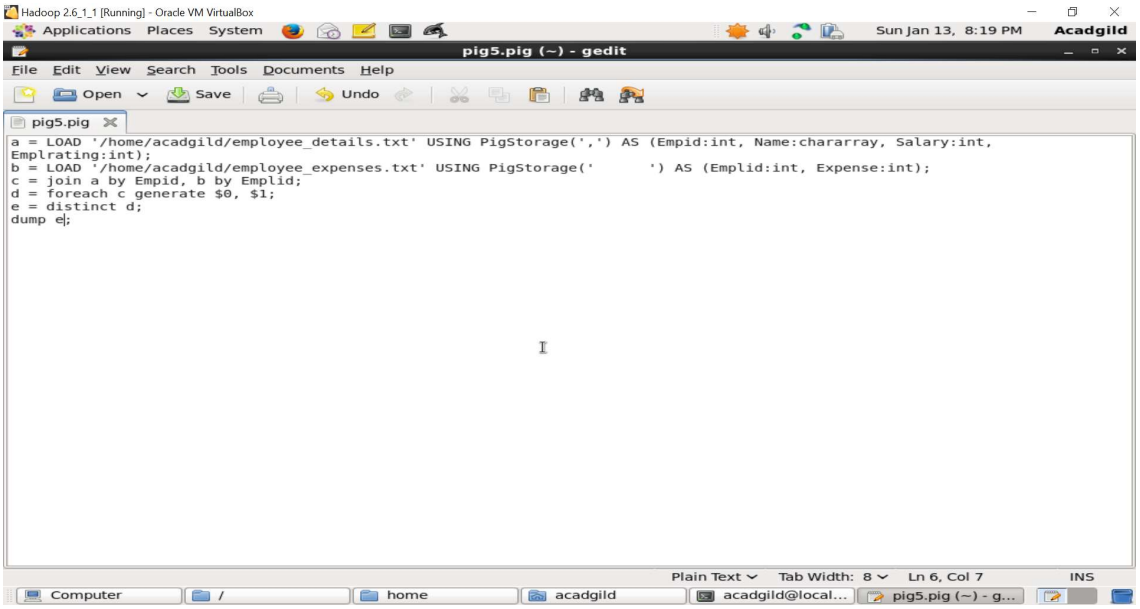
(c) Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)



Output

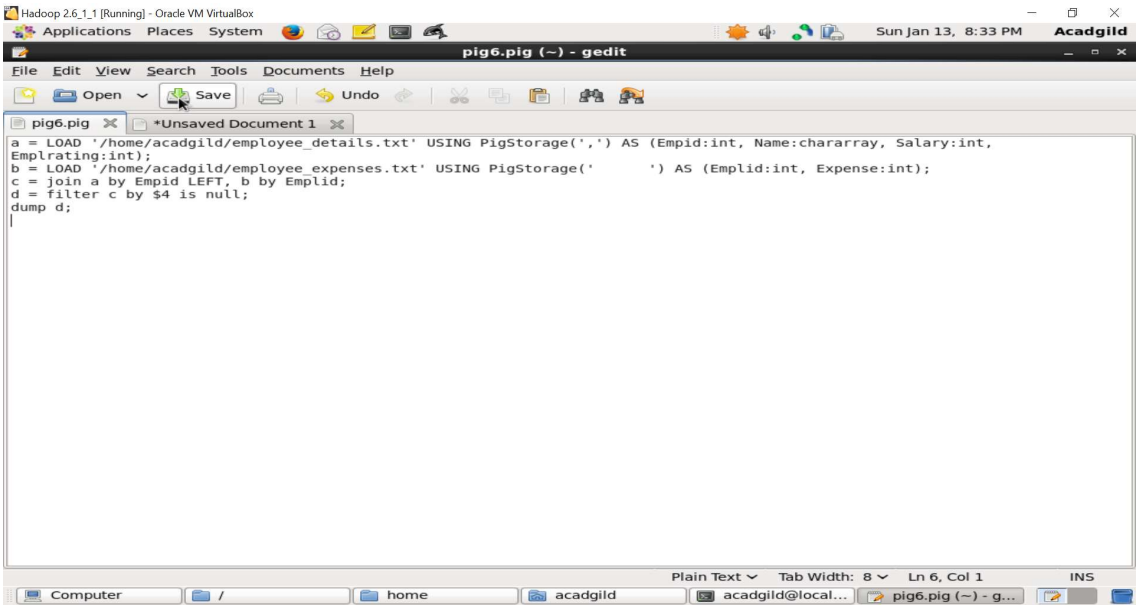
```
2019-01-13 20:10:22,997 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(110,Priyanka)
```

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



```
2019-01-13 20:14:16,064 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 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.



Output

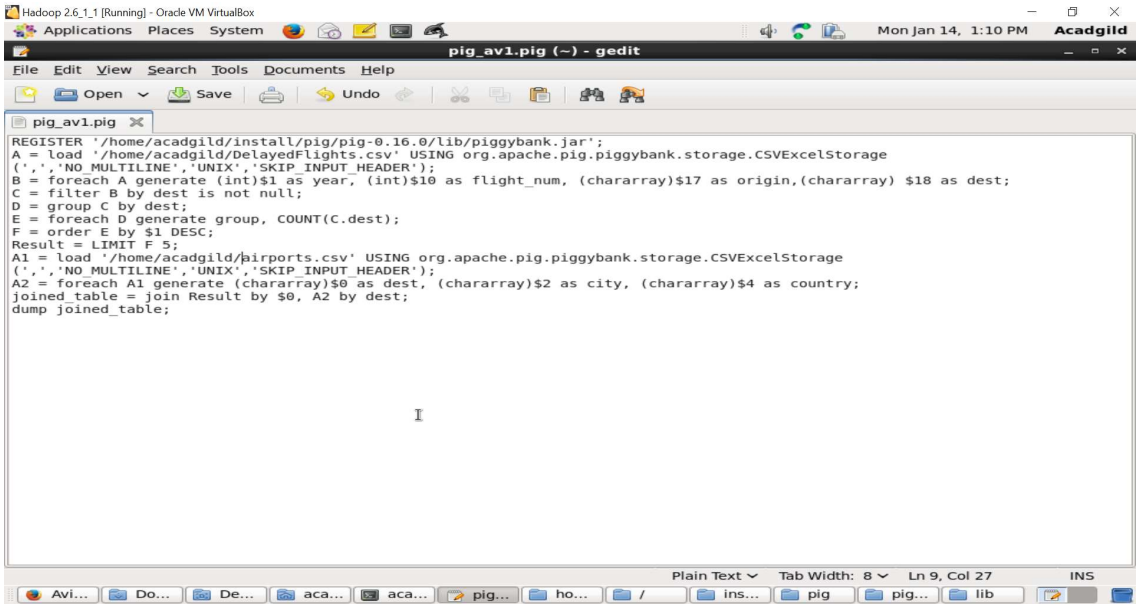
```
2019-01-13 20:33:07,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(103,Akshay,11000,3,,)
(106,Aamir,25000,1,,)
(107,Salman,17500,2,,)
(108,Ranbir,14000,3,,)
(109,Katrina,1000,4,,)
(111,Tushar,500,1,,)
(112,Akay,5000,2,,)
(113,Jubeen,1000,1,,)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

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/

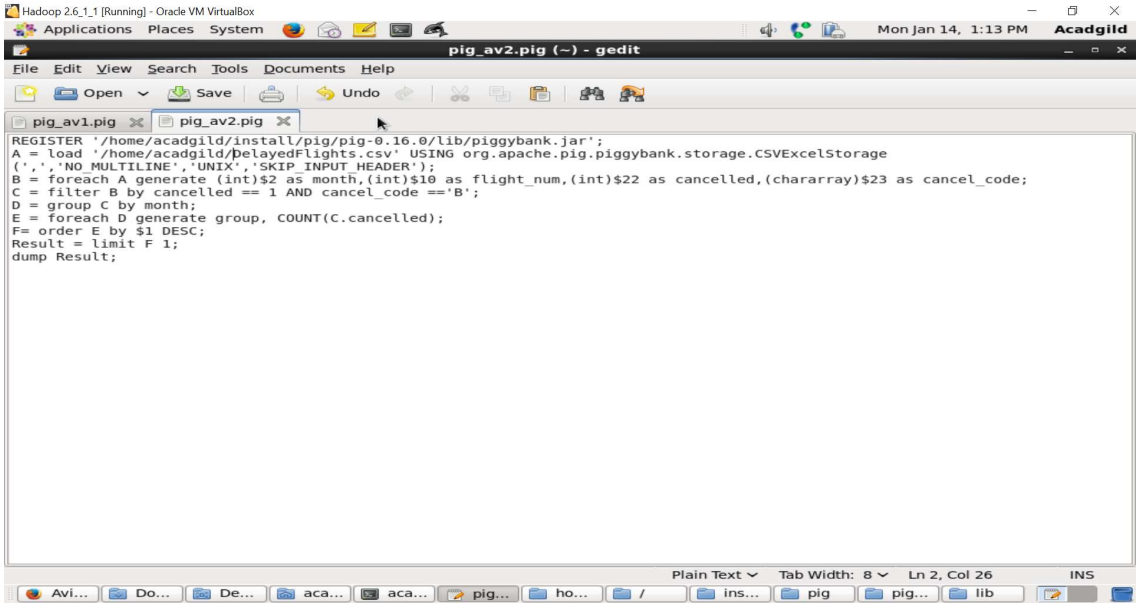
1) Find out the top 5 most visited destinations.



Output:

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

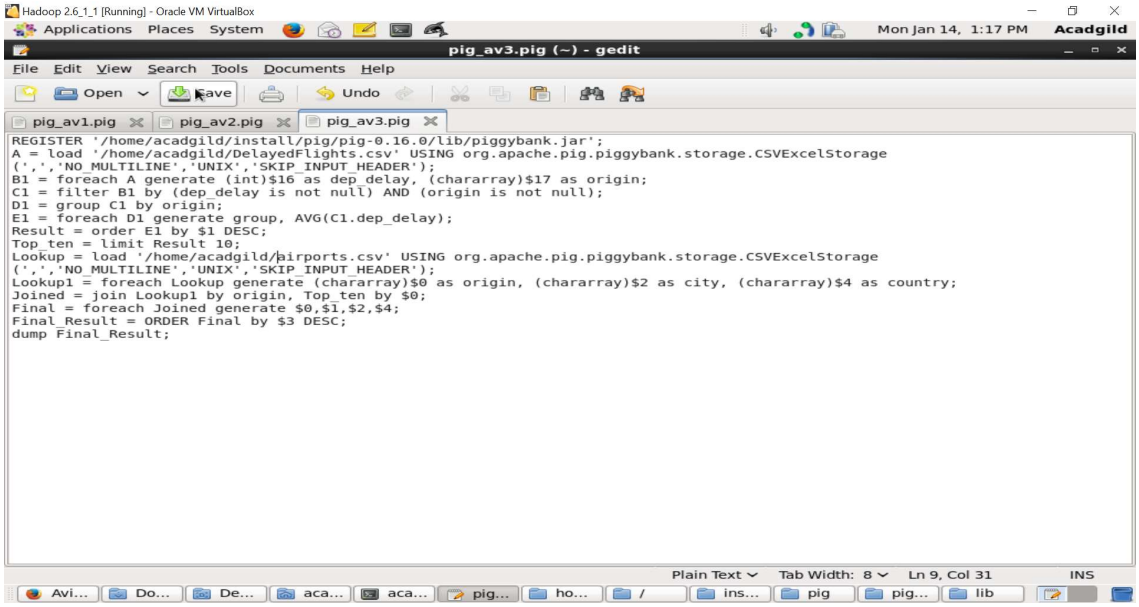
2) Which month has seen the most number of cancellations due to bad weather?



Output

```
2019-01-14 13:13:30,633 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.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 ~]$
```

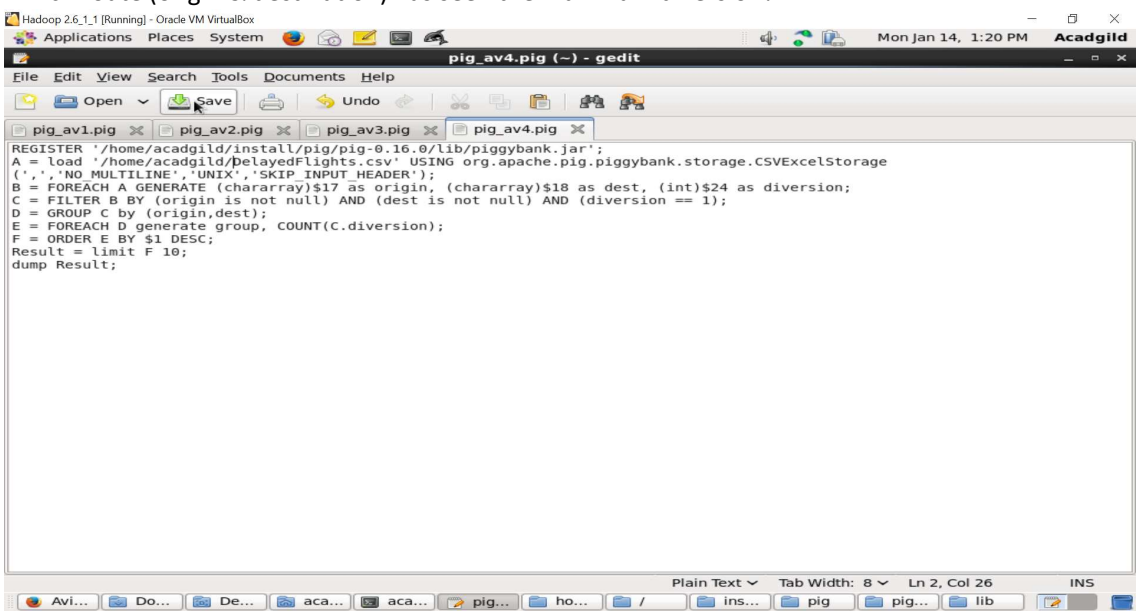
3) Top ten origins with the highest AVG departure delay



Output

```
2019-01-14 13:17:44,375 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.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.executionengi
ne.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)
```