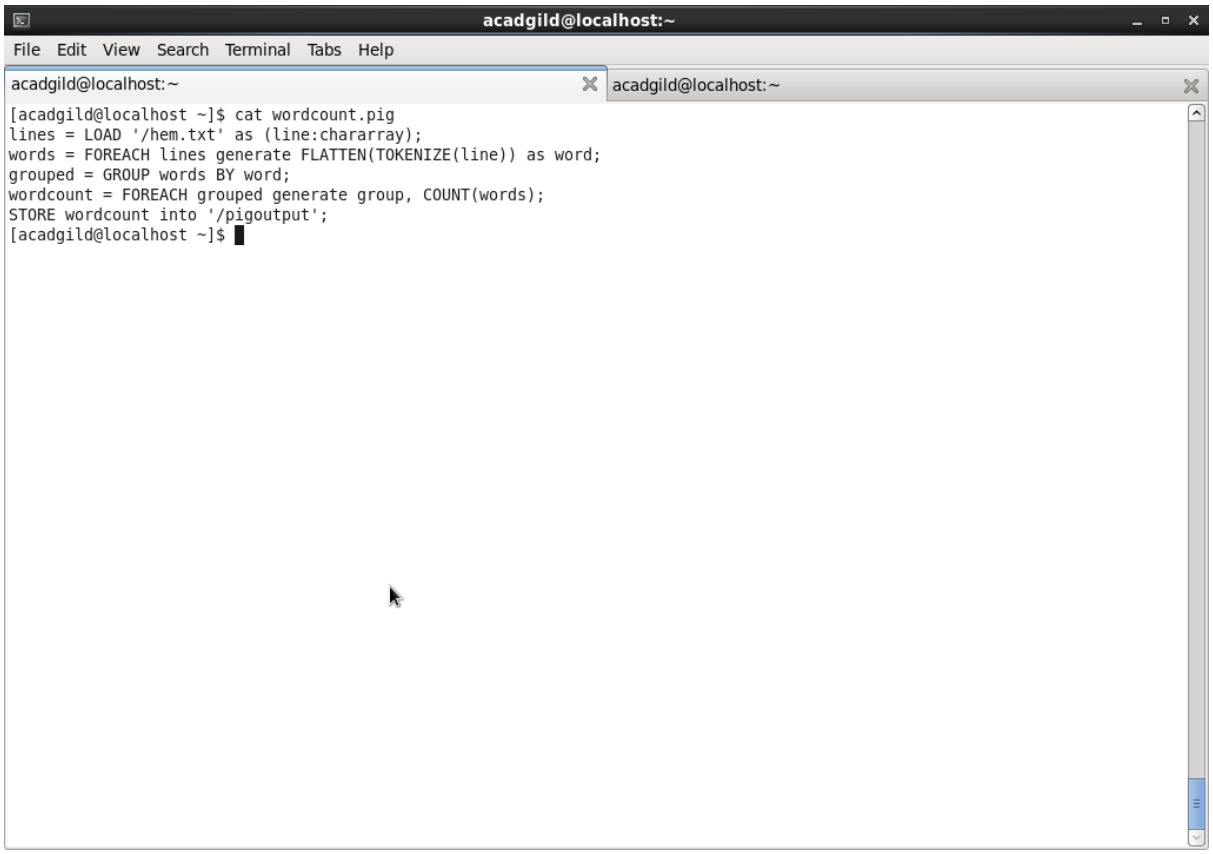
Task 1

Write a program to implement wordcount using Pig.

Pig Script

```
lines = LOAD '/hem.txt' as (line:chararray);

words = FOREACH lines generate FLATTEN(TOKENIZE(line)) as word;

grouped = GROUP words BY word;

wordcount = FOREACH grouped generate group, COUNT(words);

STORE wordcount into '/pigoutput';
```



Output

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~        acadgild@localhost:~
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /pigoutput
18/09/08 00:13:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-08 00:11 /pigoutput/_SUCCESS
-rw-r--r--   1 acadgild supergroup         46 2018-09-08 00:11 /pigoutput/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /pigoutput
18/09/08 00:13:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
cat: `/pigoutput': Is a directory
[acadgild@localhost ~]$ hadoop fs -cat /pigoutput/part-r-00000
18/09/08 00:14:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
a       1
is      1
this    1
Hello   1
session 1
wonderful       1
[acadgild@localhost ~]$
```

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.t
xt
employee_expenses(EmpID,Expence)
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)

Script

```
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (empid:int, name:chararray, salary:int, rating:int);

ordered = order employeedetail by rating desc,name asc;

limitdetail = limit ordered 5;

showoutput = foreach limitdetail generate empid, name;

store showoutput into '/employee_task1';
```

Output

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~    acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fs -cat /employee_task1/part-r-00000
18/09/12 23:17:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
105     Pawan
110     Priyanka
104     Anubhav
109     Katrina
103     Akshay
[acadgild@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)

Script-:

```
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (empid:int, name:chararray, salary:int, rating:int);

empodd = filter employeedetail by (empid % 2 == 1);

ordered = order empodd by salary desc,name asc;

limitsalary = limit ordered 3;

showoutput = foreach limitsalary generate empid, name;

store showoutput into '/employee_task2';
```

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~    acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fs -cat /employee_task2/part-r-00000
18/09/12 23:17:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
101     Amitabh
107     Salman
103     Akshay
[acadgild@localhost ~]$
```

Output-:

(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)

Script-:

```
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (detailempid:int,
name:chararray, salary:int, rating:int);

employeeexpense = LOAD '/employee_expenses.txt' as (expenseempid:int, expense:int);

joined = join employeedetail by detailempid , employeeexpense by expenseempid;

ordered = order joined by expense desc,name asc;

limitoutput = limit ordered 1;

generateoutput = foreach limitoutput generate detailempid,name;

store generateoutput into '/employee_task3';
```

Output-:

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~    acadgild@localhost:~
[acadgild@localhost ~]$ cat employee_task3.pig
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (detailempid:int, name:chararray, salary:int, rating:i
nt);
employeeexpense = LOAD '/employee_expenses.txt' as (expenseempid:int, expense:int);
joined = join employeedetail by detailempid , employeeexpense by expenseempid;
ordered = order joined by expense desc,name asc;
limitoutput = limit ordered 1;
generateoutput = foreach limitoutput generate detailempid,name;
store generateoutput into '/employee_task3';
[acadgild@localhost ~]$
```

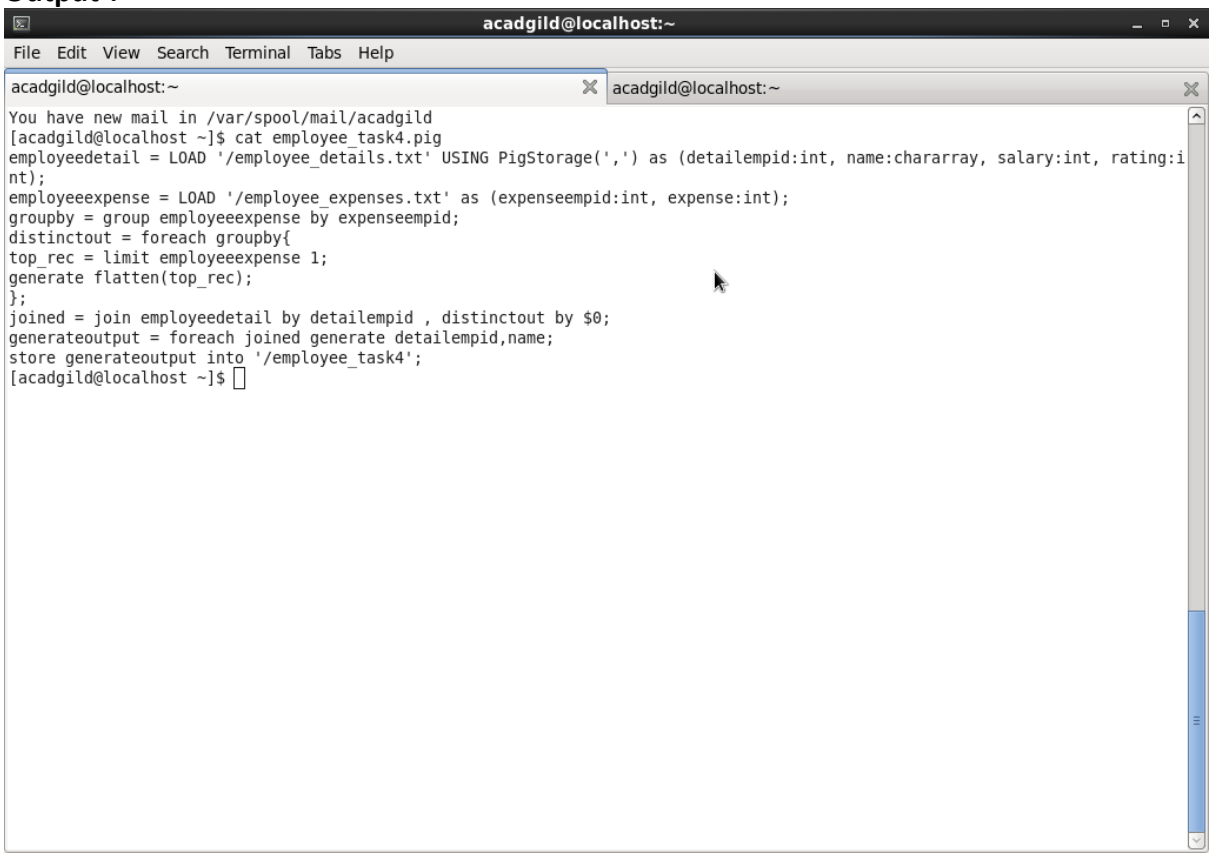
```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~    acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fs -cat /employee_task3/part-r-00000
18/09/12 23:18:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
110     Priyanka
[acadgild@localhost ~]$
```

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

Script-:

```
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (detailempid:int, name:chararray, salary:int, rating:int);
employeeexpense = LOAD '/employee_expenses.txt' as (expenseempid:int, expense:int);
groupby = group employeeexpense by expenseempid;
distinctout = foreach groupby{
top_rec = limit employeeexpense 1;
generate flatten(top_rec);
};
joined = join employeedetail by detailempid , distinctout by $0;
generateoutput = foreach joined generate detailempid,name;
store generateoutput into '/employee_task4';
```

Output-:



```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~        acadgild@localhost:~
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /employee_task4/part-r-00000
18/09/12 23:19:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
101     Amitabh
102     Shahrukh
104     Anubhav
105     Pawan
110     Priyanka
114     Madhuri
[acadgild@localhost ~]$
```

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

Script-:

```
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (detailempid:int,
name:chararray, salary:int, rating:int);

employeeexpense = LOAD '/employee_expenses.txt' as (expenseempid:int, expense:int);

joined = join employeedetail by detailempid left outer , employeeexpense by $0;

e = filter joined by expenseempid is null;

generateoutput = foreach e generate detailempid,name;

store generateoutput into '/employee_task5';
```

Output-:

acadgild@localhost:~

File Edit View Search Terminal Tabs Help

acadgild@localhost:~ | acadgild@localhost:~

```
[acadgild@localhost ~]$ hadoop fs -cat /employee_task5/part-r-00000
18/09/12 23:20:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
103     Akshay
106     Aamir
107     Salman
108     Ranbir
109     Katrina
111     Tushar
112     Ajay
113     Jubeen
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

acadgild@localhost:~

File Edit View Search Terminal Tabs Help

acadgild@localhost:~ | acadgild@localhost:~

```
[acadgild@localhost ~]$ cat employee_task5.pig
employeedetail = LOAD '/employee_details.txt' USING PigStorage(',') as (detailempid:int, name:chararray, salary:int, rating:i
nt);
employeeexpense = LOAD '/employee_expenses.txt' as (expenseempid:int, expense:int);
joined = join employeedetail by detailempid left outer , employeeexpense by $0;
e = filter joined by expenseempid is null;
generateoutput = foreach e generate detailempid,name;
store generateoutput into '/employee_task5';
[acadgild@localhost ~]$
```

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/

Problem Statement 1

Find out the top 5 most visited destinations.

Script-:

```
REGISTER '/home/acadgild/piggybank.jar';

A = LOAD '/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;

store Result into '/aviation_task1';
```

Output-:

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~        acadgild@localhost:~
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ cat aviation_task1.pig
REGISTER '/home/acadgild/piggybank.jar';
A = LOAD '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPU
T_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 $1 DESC;
Result = LIMIT F 5;
store Result into '/aviation_task1';
[acadgild@localhost ~]$
```

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~        acadgild@localhost:~
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /aviation_task1/part-r-00000
18/09/12 23:21:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
ORD     108984
ATL     106898
DFW     70657
DEN     63003
LAX     59969
[acadgild@localhost ~]$
```

Problem Statement 2

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

Script-:

```
REGISTER '/home/acadgild/piggybank.jar';

A = LOAD '/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 cancelledcode;

C = filter B by cancelled == 1 AND cancelledcode == 'B';

D = group C by month;

E = foreach D generate group, COUNT(C.cancelled);

F = order E by $1 DESC;

Result = LIMIT F 1;

store Result into '/aviation_task2.pig';
```

Output-:

acadgild@localhost:~

File Edit View Search Terminal Tabs Help

acadgild@localhost:~ | acadgild@localhost:~

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ cat aviation_task2.pig
REGISTER '/home/acadgild/piggybank.jar';
A = LOAD '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPU
T_HEADER');
B = foreach A generate (int)$2 as month, (int)$10 as flight_num, (int)$22 as cancelled,(chararray)$23 as cancelledcode;
C = filter B by cancelled == 1 AND cancelledcode == 'B';
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F = order E by $1 DESC;
Result = LIMIT F 1;
store Result into '/aviation_task2.pig';
[acadgild@localhost ~]$
```

Acadgild

acadgild@localhost:~

File Edit View Search Terminal Tabs Help

acadgild@localhost:~ | acadgild@localhost:~

```
[acadgild@localhost ~]$ hadoop fs -cat /aviation_task2.pig/part-r-00000
18/09/12 23:23:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
12      250
[acadgild@localhost ~]$
```

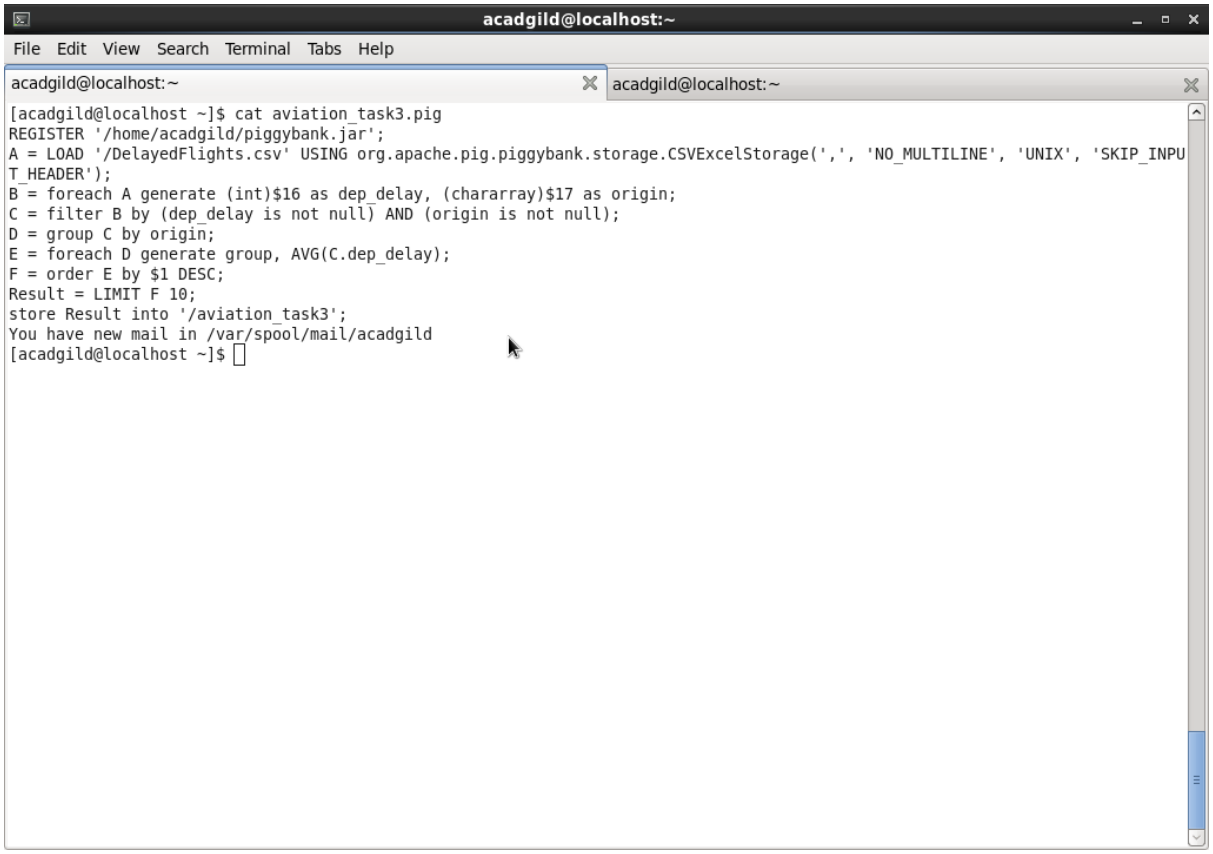
Problem Statement 3

Top ten origins with the highest AVG departure delay

Script-:

```
REGISTER '/home/acadgild/piggybank.jar';

A = LOAD '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',',
'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');

B = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;

C = filter B by (dep_delay is not null) AND (origin is not null);

D = group C by origin;

E = foreach D generate group, AVG(C.dep_delay);

F = order E by $1 DESC;

Result = LIMIT F 10;

store Result into '/aviation_task3';
```

Output-:



```
(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)
```

Problem Statement 4

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

Script-:

```
REGISTER '/home/acadgild/piggybank.jar';

A = LOAD '/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 1;

store Result into '/aviation_task4';
```

Output-:

acadgild@localhost:~

File Edit View Search Terminal Tabs Help

acadgild@localhost:~ | acadgild@localhost:~

```
[acadgild@localhost ~]$ cat aviation_task4.pig
REGISTER '/home/acadgild/piggybank.jar';
A = LOAD '/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPU
T_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 $1 DESC;
Result = LIMIT F 1;
store Result into '/aviation_task4';
[acadgild@localhost ~]$
```

```
(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)
```