

## Session 7 - EXPLORING APACHE PIG

### Assignment 1

**Task1** : Write a program to implement wordcount using Pig.

```
lines = LOAD '/user/acadgild/hadoop/word-count.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);  
DUMP wordcount;
```

Loading the target text file.

Tokenize the line and generate each word.

Same words are grouped together.

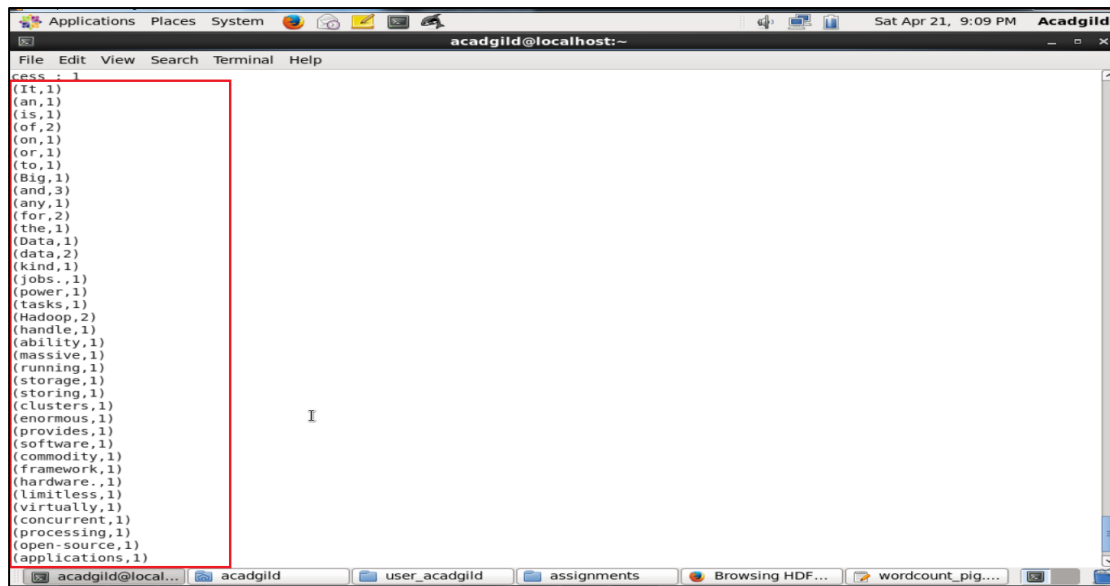
For each group, the number of words are counted.

Display the result on the screen.

Command to run :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/wordcount_pig.pig
```

Output :



```
File Edit View Search Terminal Help
acacgild@localhost:~
cess : 1
(it,1)
(an,1)
(is,1)
(of,2)
(on,1)
(or,1)
(to,1)
(Big,1)
(and,3)
(any,1)
(for,2)
(the,1)
(Data,1)
(data,2)
(kind,1)
(jobs,1)
(power,1)
(tasks,1)
(Hadoop,2)
(handle,1)
(ability,1)
(massive,1)
(running,1)
(storage,1)
(storing,1)
(clusters,1)
(enormous,1)
(provides,1)
(software,1)
(commodity,1)
(framework,1)
(hardware,1)
(limitless,1)
(virtually,1)
(concurrent,1)
(processing,1)
(open-source,1)
(applications,1)
```

The output shows the count of each word in the text file.

**Task2 :** 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)

```
employee = LOAD '/home/acacgild/user_acacgild/assignments/Pig/employee_details.txt'
USING PigStorage(',') as (id:int,name:chararray,salary:int,rating:int);

emp_rating_order = ORDER employee BY rating DESC, name ASC;

emp_rating_limit = LIMIT emp_rating_order 5;

output = FOREACH emp_rating_limit GENERATE $0,$1;

DUMP output;
```

Loading the employee details text file as a relation.

Sorting the relation employee in a descending order based on rating and ascending order based on the name.

Limiting the result to the top 5 tuples.

Extracting the Employee id and the employee name from each tuple.

Displaying the output on the screen.

Command :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/task2a.pig
```

Output :

```
2018-04-22 07:35:04,909 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=
e=JobTracker, sessionId= - already initialized
2018-04-22 07:35:04,909 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=
e=JobTracker, sessionId= - already initialized
2018-04-22 07:35:04,910 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=
e=JobTracker, sessionId= - already initialized
2018-04-22 07:35:04,919 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=
e=JobTracker, sessionId= - already initialized
2018-04-22 07:35:04,926 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-04-22 07:35:04,955 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 07:35:04,955 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 07:35:04,993 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 07:35:04,993 [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)
2018-04-22 07:35:05,150 [main] INFO org.apache.pig.Main - Pig script completed in 13 seconds and 733 milliseconds (13733 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Alphabetical ordered list of the employees with highest rating.

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)

```
employee = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_details.txt'
USING PigStorage(',') as (id:int,name:chararray,salary:int,rating:int);
```

```
odd_id = FILTER employee By id%2==1;
```

```
emp_high_salary = ORDER odd_id by salary DESC, name ASC;
```

```
emp_salary_limit = LIMIT emp_high_salary 3;
```

```
output2 = FOREACH emp_salary_limit GENERATE $0,$1;
```

```
DUMP output2;
```

Loading the employee details text file as a relation.

Filtering the tuples with employee id as odd number.

Sorting in a descending order based on salary and ascending order based on the name.

Limiting the result to the top 3 tuples.

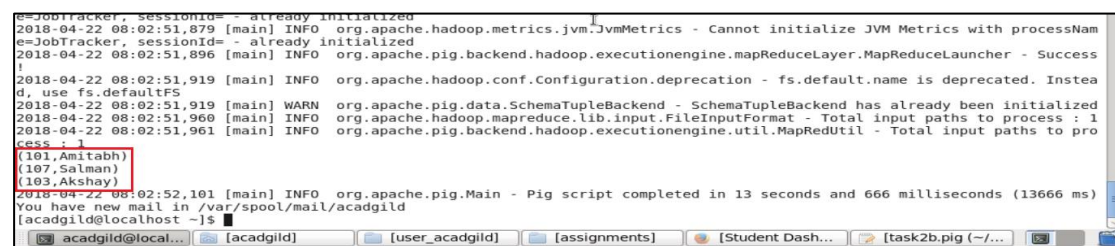
Extracting the Employee id and the employee name from each tuple.

Displaying the output on the screen.

Command :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/task2b.pig
```

Output :



Top 3 employees with highest salary,whose employee id is an odd number.

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)

```
details = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_details.txt'  
USING PigStorage(',') as (id:int,name:chararray,salary:int,rating:int);
```

```
expenses = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_expenses.txt'  
USING PigStorage('\t') as (id:int,expense:int);
```

```
result = JOIN details BY id, expenses BY id;
```

```
top_expense_list = GROUP result BY expense;
```

```
B = ORDER top_expense_list BY $0 DESC;
```

```
top_expense = LIMIT B 1;
```

```
C = FOREACH top_expense GENERATE FLATTEN(top_expense.$1);
```

```
output3 = FOREACH C GENERATE $0,$1;
```

```
DUMP output3;
```

Loading the employee details and employee expenses text files as two relations.

Join the 2 relations based on equality of employee ids.

Group the resultant relation based on the expense amount and then arrange in the descending order of expense.

Limiting the result to the top 1 tuples. This will give all the tuples with the max expense amount.

Flattening the resultant bag into a tuple.

Extracting the Employee id and the employee name from each tuple.

Displaying the output on the screen.

Command :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/task2c.pig
```

Output:

```
e=JobTracker, sessionId= - already initialized
2018-04-22 11:18:46,943 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:18:46,944 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:18:46,946 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:18:46,959 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2018-04-22 11:18:46,980 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 11:18:46,980 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 11:18:47,011 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 11:18:47,011 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
(102,Shahrukh)
2018-04-22 11:18:47,156 [main] INFO org.apache.pig.Main - Pig script completed in 17 seconds and 312 milliseconds (17312 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Alphabetical ordered list of employees with highest expense.

D) List of employees (employee id and employee name) having entries in employee\_expenses file.

```
details = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_details.txt'
USING PigStorage(',') as (id:int,name:chararray,salary:int,rating:int);
```

```
expenses = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_expenses.txt'
USING PigStorage('\t') as (id:int,expense:int);
```

```
result = JOIN details BY id, expenses BY id;
```

```
output4 = DISTINCT (FOREACH result GENERATE $0,$1);
```

```
DUMP output4;
```

Loading the employee details and employee expenses text files as two relations.

Join the 2 relations based on equality of employee ids.

Since this is an equijoin, and an equijoin returns only the tuples that have equivalent values, we get only the employees whose entries are present in the employee expenses file.

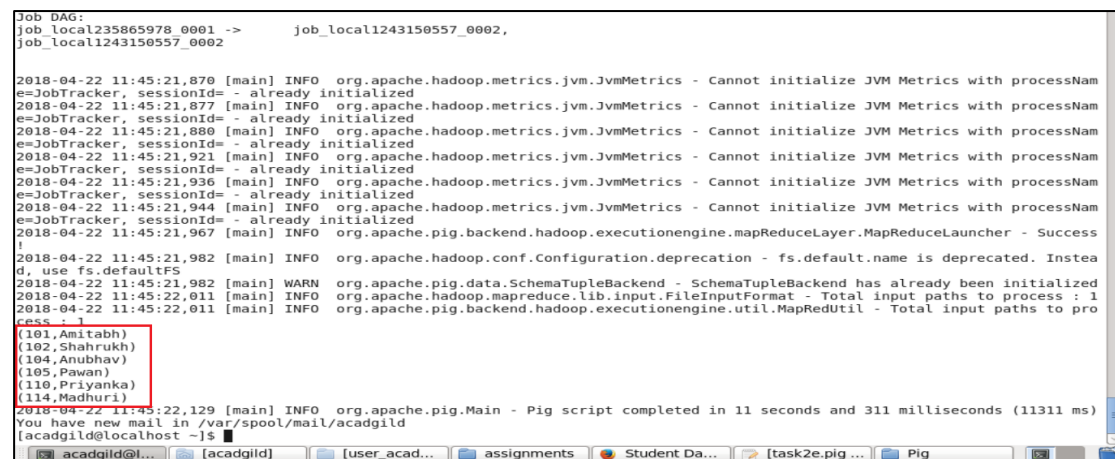
Extracting the Employee id and the employee name from each tuple.

Displaying the output on the screen.

Command :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/task2d.pig
```

Output :



```
Job DAG:
job_local1235865978_0001 ->      job_local1243150557_0002,
job_local1243150557_0002

2018-04-22 11:45:21,870 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:45:21,877 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:45:21,880 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:45:21,921 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:45:21,936 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:45:21,944 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:45:21,967 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-22 11:45:21,982 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 11:45:21,982 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 11:45:22,011 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 11:45:22,011 [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)
2018-04-22 11:45:22,129 [main] INFO  org.apache.pig.Main - Pig script completed in 11 seconds and 311 milliseconds (11311 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

List of employees with entries in the employee\_expenses file.

E) List of employees (employee id and employee name) having no entry in employee\_expenses file.

```
details = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_details.txt'
USING PigStorage(',') as (id:int,name:chararray,salary:int,rating:int);
```

```
expenses = LOAD '/home/acadgild/user_acadgild/assignments/Pig/employee_expenses.txt'
USING PigStorage('\t') as (id:int,expense:int);
```

```
result = JOIN details BY id LEFT OUTER, expenses BY id;
```

```
emp_without_exp = FILTER result BY expenses::id is null;
output5 = FOREACH emp_without_exp GENERATE $0,$1;
DUMP output5;
```

Loading the employee details and employee expenses text files as two relations.

Perform an Left Outer join based on employee ids.

A left outer join preserves the unmatched rows from the first (left) relation, joining them with a NULL tuple in the second (right) relation.

Then we filter out the tuples where the id from the expenses (right) relation is null. Thus those tuples who have entries in the expenses relation remain.

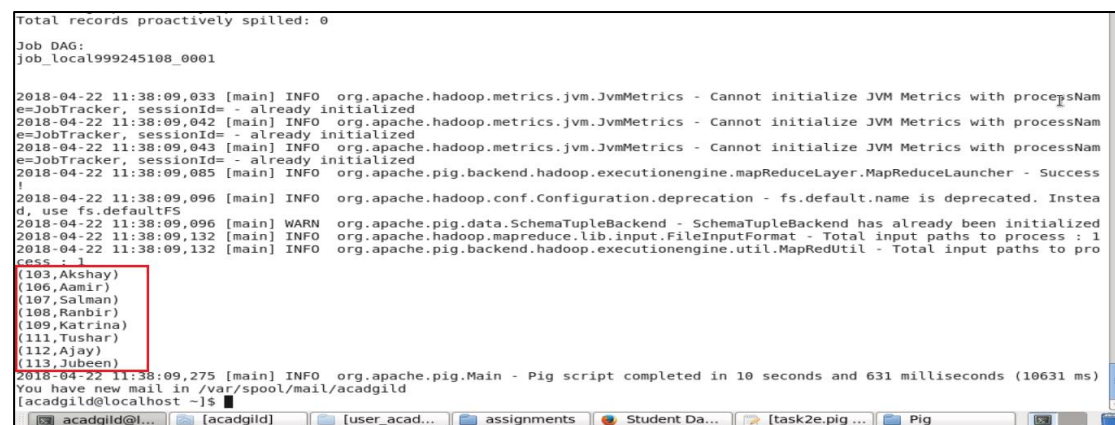
Extracting the Employee id and the employee name from each tuple.

Displaying the output on the screen.

Command :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/task2e.pig
```

Output:



```
Total records proactively spilled: 0
Job DAG:
Job_local999245108_0001

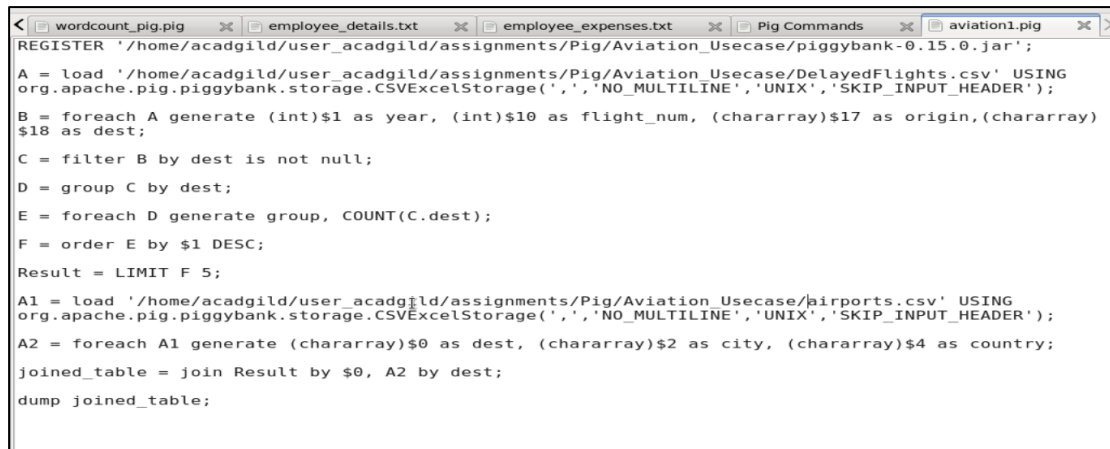
2018-04-22 11:38:09,033 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:38:09,042 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:38:09,043 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 11:38:09,085 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2018-04-22 11:38:09,096 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 11:38:09,096 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 11:38:09,132 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-04-22 11:38:09,132 [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)
2018-04-22 11:38:09,275 [main] INFO org.apache.pig.Main - Pig script completed in 10 seconds and 631 milliseconds (10631 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

List of employees with no entries in employee\_expenses file.

**Task3:** 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



```
REGISTER '/home/acadgild/user_acadgild/assignments/Pig/Aviation_Usecase/piggybank-0.15.0.jar';

A = load '/home/acadgild/user_acadgild/assignments/Pig/Aviation_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/user_acadgild/assignments/Pig/Aviation_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;
```

The piggybank.jar is registered i.e the methods in the jar file can be used in the program.

The DelayedFlights.CSV file is loaded.

The required columns are fetched and iterated.

To filter out the columns whose destination is null, we use a filter function.

Tuples with the same destination are grouped together and the number of tuples for each destination is counted.

Creating a descending ordered list of tuples and limiting them to 5.

Loading the airports.csv to find out the names of the corresponding 5 destinations by joining them.

Display the final result on the output.

Command :

```
pig -x local /home/acadgild/user_acadgild/assignments/Pig/aviation1.pig
```

Output :



```
eJobTracker, sessionId= already initialized
2018-04-22 13:20:14,496 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
eJobTracker, sessionId= already initialized
2018-04-22 13:20:14,514 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
eJobTracker, sessionId= already initialized
2018-04-22 13:20:14,525 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
eJobTracker, sessionId= already initialized
2018-04-22 13:20:14,529 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
eJobTracker, sessionId= already initialized
2018-04-22 13:20:14,555 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-04-22 13:20:14,572 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-04-22 13:20:14,572 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 13:20:14,621 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 13:20:14,621 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 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)
2018-04-22 13:20:14,777 [main] INFO org.apache.pig.Main - Pig script completed in 50 seconds and 332 milliseconds (50332 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

**Problem Statement 2:** Which month has seen the most number of cancellations due to bad weather?

```
word-count.txt wordcount_pig.pig Pig Commands aviation1.pig aviation2.pig
REGISTER '/home/acadgild/user_acadgild/assignments/Pig/Aviation_Usecase/piggybank-0.15.0.jar';
A = load '/home/acadgild/user_acadgild/assignments/Pig/Aviation_Usecase/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;
```

The piggybank.jar is registered I.e the methods in the jar file can be used in the program.

The DelayedFlights.CSV file is loaded.

The required columns are fetched and iterated.

To filter out the columns where canceled = 1 meaning that it has been canceled and cancel\_code = 0 indicating the reason for canceling is due to bad weather.

Tuples are grouped together based on month and the number of tuples for each month is counted.

Creating a descending ordered list of tuples and limiting them to top 1.

Display the final result on the output.

Command :

*pig -x local /home/acadgild/user\_acadgild/assignments/Pig/aviation2.pig*

Output :

```
2018-04-22 13:41:35,062 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 13:41:35,063 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 13:41:35,063 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 13:41:35,086 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2018-04-22 13:41:35,097 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 13:41:35,097 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 13:41:35,128 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 13:41:35,128 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,258)
2018-04-22 13:41:35,249 [main] INFO org.apache.pig.Main - Pig script completed in 32 seconds and 958 milliseconds (32958 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

### Problem Statement 3: Top ten origins with the highest AVG departure delay

```
word-count.txt wordcount_pig.pig Pig Commands aviation1.pig aviation2.pig aviation3.pig
REGISTER '/home/acadgild/user_acadgild/assignments/Pig/Aviation_Usecase/piggybank-0.15.0.jar';
A = load '/home/acadgild/user_acadgild/assignments/Pig/Aviation_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/user_acadgild/assignments/Pig/Aviation_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 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;
```

The piggybank.jar is registered I.e the methods in the jar file can be used in the program.

The DelayedFlights.CSV file is loaded.

The required columns are iterated.

To filter out the columns whose departure delay and origin not is null, we use a filter function.

Tuples with the same origin are grouped together and the average (AVG) departure delay of tuples for each origin is counted.

Creating a descending ordered list of tuples and limiting them to 10.

Loading the airports.csv to find out the names of the corresponding 10 origins' airports by joining them.

Display the final result on the output.

Command :

*pig -x local /home/acadgild/user\_acadgild/assignments/Pig/aviation3.pig*

Output :

```
2018-04-22 14:02:13,596 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 14:02:13,609 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 14:02:13,618 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-22 14:02:13,639 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 14:02:13,641 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 14:02:13,661 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 14:02:13,661 [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)
(MOT,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)
2018-04-22 14:02:13,810 [main] INFO org.apache.pig.Main - Pig script completed in 48 seconds and 671 milliseconds (48671 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

**Problem Statement 4:** Which route (origin & destination) has seen the maximum diversion?

```
wordcount_pig.pig Pig Commands aviation1.pig aviation2.pig aviation3.pig aviation4.pig
REGISTER '/home/acadgild/user_acadgild/assignments/Pig/Aviation_Usecase/piggybank-0.15.0.jar';
A = load '/home/acadgild/user_acadgild/assignments/Pig/Aviation_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;
```

The piggybank.jar is registered I.e the methods in the jar file can be used in the program.

The DelayedFlights.CSV file is loaded.

The required columns are iterated.

To filter out the columns whose destination and origin is not null also the diversion is 1 indicating that the flight was diverted, we use a filter function.

Tuples are grouped together based on destination and origin and the diversions of are counted.

Creating a descending ordered list of tuples and limiting them to 10.

Display the final result on the output.

Command :

*pig -x local /home/acadgild/user\_acadgild/assignments/Pig/aviation4.pig*

Output :

```
2018-04-22 14:10:01,192 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-04-22 14:10:01,211 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
2018-04-22 14:10:01,220 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-22 14:10:01,221 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-22 14:10:01,251 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-22 14:10:01,251 [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)
2018-04-22 14:10:01,374 [main] INFO org.apache.pig.Main - Pig script completed in 33 seconds and 523 milliseconds (33523 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

