

PIG Assignment 1**Task 1:**

Write a program to implement word count using PIG.

Code:

```
lines = LOAD '/home/acadgild/Desktop/wordcount.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;
```

Output:

```
acadgild@localhost:~$ pig
Counters:
Total records written : 11
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local207202730_0001

2018-10-09 18:51:53,067 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-09 18:51:53,083 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-09 18:51:53,096 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-09 18:51:53,244 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-09 18:51:53,290 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-09 18:51:53,296 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-09 18:51:53,446 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-09 18:51:53,448 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(a,1)
(be,1)
(bad,2)
(big,1)
(day,2)
(data,1)
(good,2)
(learn,1)
(bigdata,1)
(morning,2)
(engineer,1)
2018-10-09 18:51:53,958 [main] INFO org.apache.pig.Main - Pig script completed in 20 seconds and 594 milliseconds (20594 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Explanation:

- Text file was read into variable lines using default delimiter as comma (,).
- Tokenized lines into words based on delimiter.
- Grouped the whole set of words based on each word.
- Taken count of each word in set.

-
the result set.

Dumped

Task 2:

2(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)

```
emp_details = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_details.txt' using PigStorage(',')
as(emp_id:int,ename:chararray,sal:int,ranking:int);

sort_by_rank_ename = ORDER emp_details by ranking DESC,ename ASC;

top5 = LIMIT sort_by_rank_ename 5;

DUMP top5;
```

Output:

```

e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,631 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,641 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,742 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,755 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,778 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,846 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,848 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,861 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,917 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,930 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,935 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:20:46,969 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-10-09 19:20:47,000 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-10-09 19:20:47,006 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-09 19:20:47,135 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-09 19:20:47,135 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(105,Pawan,2500,5)
(110,Priyanka,2000,5)
(104,Anubhav,5000,4)
(109,Katrina,1000,4)
(103,Akshay,11000,3)
2018-10-09 19:20:47,783 [main] INFO org.apache.pig.Main - Pig script completed in 30 seconds and 127 milliseconds (30127 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

Explanation:

- Read emp details into PIG variable.
- Done double sorting – First on rankings DESC and then by ename ASC.
- Fetched top 5 by using limit operator.
- Dumped the result.

2(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)

```

emp_details = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_details.txt' using PigStorage(',')
as(emp_id:int,ename:chararray,sal:int,ranking:int);

odd_emp_id = FILTER emp_details by emp_id%2!=0;

sort_by_sal_ename = ORDER odd_emp_id by sal DESC,ename ASC;

top3 = LIMIT sort_by_sal_ename 3;

res = FOREACH top3 GENERATE $0,$1;

DUMP res;

```

Output:

```

2018-10-09 19:35:25,974 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:25,979 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:25,988 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,040 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,045 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,102 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,116 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,128 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,215 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,227 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,238 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-09 19:35:26,275 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-10-09 19:35:26,347 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-10-09 19:35:26,349 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-09 19:35:26,433 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-09 19:35:26,435 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)
2018-10-09 19:35:27,077 [main] INFO org.apache.pig.Main - Pig script completed in 33 seconds and 582 milliseconds (33582 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

Explanation:

- Load emp details into a PIG variable.
- Filter data set with only odd emp id's using FILTER clause emp_id%2!=0.
- Done double sorting – First on sal DESC and then by ename ASC.
- Fetched top 3 by using limit operator.
- Fetched empid and ename out of the result set using FOREACH.
- Dumped the result.

2(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_expenses = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_expenses.txt' using
PigStorage('\t') as(emp_id:int,expenses:int);
```

```
emp_details = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_details.txt' using
PigStorage(',') as(emp_id:int,ename:chararray,sal:int,ranking:int);
```

```
grp_all = GROUP emp_expenses all;
```

```
max_exp = FOREACH grp_all GENERATE MAX(emp_expenses.expenses);
```

```
join_res = JOIN emp_expenses by expenses,max_exp by $0;
```

Output:

```

2018-10-10 08:44:13,159 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,271 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,278 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,282 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini
2018-10-10 08:44:13,365 [main] INFO c
vm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini
2018-10-10 08:44:13,368 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,384 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,450 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,460 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,463 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,530 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,536 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,541 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 08:44:13,596 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-10-10 08:44:13,635 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-10-10 08:44:13,636 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-10 08:44:13,794 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-10 08:44:13,794 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(110,Priyanka)
(102,Shahrakh)
2018-10-10 08:44:14,343 [main] INFO org.apache.pig.Main - Pig script completed in 49 seconds and 891 milliseconds (49891 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

Explantion:

- Read emp details and expenses into two different PIG variables.
- Group emp expenses all.
- Find MAX of expenses out of above result set.
- Join MAX(expense) with emp expenses data set to get emp id.
- Join emp id yielded in above step with emp details.
- Sort data using ename.
- Dumped the result.

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

```
emp_expenses = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_expenses.txt' using
PigStorage('\t') as(emp_id:int,expenses:int);

emp_details = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_details.txt' using PigStorage(',')
as(emp_id:int,ename:chararray,sal:int,ranking:int);

emp_in_exp = FOREACH emp_expenses GENERATE emp_id;

dist_emp = DISTINCT emp_in_exp;

join_left_outer = JOIN emp_details by emp_id LEFT OUTER,dist_emp by $0;

res = FILTER join_left_outer by $4 IS NOT NULL;

res1 = FOREACH res GENERATE $0,$1;

DUMP res1;
```

Output:

```
acadgild@localhost:~
File Edit View Search Terminal Help
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1403633480_0001  -> job_local64684135_0002,
job_local64684135_0002

2018-10-10 08:45:38,946 [main] INFO d
e=JobTracker, sessionId= - already ini
2018-10-10 08:45:38,957 [main] INFO d
e=JobTracker, sessionId= - already ini
2018-10-10 08:45:38,971 [main] INFO d
e=JobTracker, sessionId= - already ini
2018-10-10 08:45:39,085 [main] INFO d
e=JobTracker, sessionId= - already ini
2018-10-10 08:45:39,093 [main] INFO d
e=JobTracker, sessionId= - already ini
2018-10-10 08:45:39,101 [main] INFO d
e=JobTracker, sessionId= - already ini
2018-10-10 08:45:39,166 [main] INFO d
!
2018-10-10 08:45:39,195 [main] INFO d
d, use fs.defaultFS
2018-10-10 08:45:39,198 [main] WARN d
2018-10-10 08:45:39,262 [main] INFO d
2018-10-10 08:45:39,264 [main] INFO d
cess : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
2018-10-10 08:45:39,716 [main] INFO d
org.apache.pig.Main - Pig script completed in 27 seconds and 920 milliseconds (27920 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Explanation:

- Load expenses and details into two different PIG variables.
- As there are duplicates in expenses find out distinct entries in expenses using DISTINCT.
- Do a left outer join between emp details and distinct employees on emp id.
- Filter data set whose emp id from dist emp(dataset) is NOT NULL.
- Fetch empid and ename.
- Dumped the result.

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

```
emp_expenses = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_expenses.txt' using
PigStorage('\t') as(emp_id:int,expenses:int);

emp_details = LOAD '/home/acadgild/Desktop/Pig_asgn/emp_details.txt' using PigStorage(',')
as(emp_id:int,ename:chararray,sal:int,ranking:int);

emp_in_exp = FOREACH emp_expenses GENERATE emp_id;

dist_emp = DISTINCT emp_in_exp;

join_left_outer = JOIN emp_details by emp_id LEFT OUTER,dist_emp by $0;

res = FILTER join_left_outer by $4 IS NULL;

res1 = FOREACH res GENERATE $0,$1;

DUMP res1;
```


Output:

```

File Edit View Search Terminal Help
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local532055924_0001 ->      job_local79656566_0002,
job_local79656566_0002

2018-10-10 08:46:36,756 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 08:46:36,799 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 08:46:36,804 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 08:46:36,889 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 08:46:36,895 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 08:46:36,907 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 08:46:36,958 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-10 08:46:37,021 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-10 08:46:37,022 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-10 08:46:37,099 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-10 08:46:37,099 [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)
2018-10-10 08:46:37,542 [main] INFO  org.apache.pig.Main - Pig script completed in 26 seconds and 884 milliseconds (26884 ms)
[acadgild@localhost ~]$

```

Explanation:

- Load expenses and details into two different PIG variables.
- As there are duplicates in expenses find out distinct entries in expenses using DISTINCT.
- Do a left outer join between emp details and distinct employees on emp id.
- Filter data set whose emp id from dist emp(dataset) is NULL.
- Fetch empid and ename.
- Dumped the result.

Task 3:

Implement the use case present in below blog link and share the complete steps along with screenshot(s) from your end.

a)Find out the top 5 most visited destinations.

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

delayed_flights = load '/home/acadgild/Desktop/airline_usecase/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

src_dest = foreach delayed_flights generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as
origin,(chararray) $18 as dest;

dest_not_null = filter src_dest by dest is not null;

group_dest = group dest_not_null by dest;

count_dest = foreach group_dest generate group, COUNT(dest_not_null.dest);

sort_dest = order count_dest by $1 DESC;

top5_dest = LIMIT sort_dest 5;

airports = load '/home/acadgild/Desktop/airline_usecase/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

airports_dest = foreach airports generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as
country;

joined_table = join top5_dest by $0, airports_dest by dest;

dump joined_table;
```

Output:

```

2018-10-10 09:18:31,152 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,157 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,168 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,175 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,176 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,209 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,215 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,224 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,246 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,248 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,251 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:18:31,273 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-10 09:18:31,309 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-10 09:18:31,310 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-10 09:18:31,375 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-10 09:18:31,375 [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)
2018-10-10 09:18:31,640 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 26 seconds and 143 milliseconds (86143 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

Explanation:

- Register piggyback.jar to load CSV data.
- Load delay flights and airports data into PIG variables.
- Fetch destination details which are not null.
- Group by destination.
- Do count on each destination.
- Sort them in descending order.
- Pull top 5 results.
- Perform a join between destination and airports and display the results.

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

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

A = load '/home/acadgild/Desktop/airline_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;
```

Output:

```
2018-10-10 09:23:57,215 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,223 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,227 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,316 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,336 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,348 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,379 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,387 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,398 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,455 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,461 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,470 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 09:23:57,510 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-10 09:23:57,548 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-10 09:23:57,548 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-10 09:23:57,647 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-10 09:23:57,647 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
2018-10-10 09:23:58,116 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 9 seconds and 140 milliseconds (69140 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Explanation:

- Load delayed flights data.
- Fetch required attributes.
- Filter data on cancelled == 1 and cancel code = 'B'
- Group data by month
- Calculate count and do a sort on values
- Limit the top row.
- Dump the result.

c) Top ten origins with the highest AVG departure delay

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

A = load '/home/acadgild/Desktop/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/Desktop/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 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;
```

Output:

```

File Edit View Search Terminal Help
acadgild@localhost:~
e=JobTracker, sessionId= - already initialized
2018-10-10 09:32:16,014 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:32:16,033 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:32:16,047 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:32:16,065 [main] INFO c tialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini tialize JVM Metrics with processNam
2018-10-10 09:32:16,070 [main] INFO c tialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini tialize JVM Metrics with processNam
2018-10-10 09:32:16,080 [main] INFO c tialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini tialize JVM Metrics with processNam
2018-10-10 09:32:16,082 [main] INFO c tialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini tialize JVM Metrics with processNam
2018-10-10 09:32:16,087 [main] INFO c tialize JVM Metrics with processNam
e=JobTracker, sessionId= - already ini tialize JVM Metrics with processNam
2018-10-10 09:32:16,093 [main] INFO c celayer.MapReduceLauncher - Success
!
2018-10-10 09:32:16,106 [main] INFO c .default.name is deprecated. Instea
d, use fs.defaultFS
2018-10-10 09:32:16,107 [main] WARN c ackend has already been initialized
2018-10-10 09:32:16,148 [main] INFO c - Total input paths to process : 1
2018-10-10 09:32:16,149 [main] INFO c pRedUtil - Total input paths to pro
cess : 1
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831
(ALQ,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)
2018-10-10 09:32:16,422 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 28 seconds and 316 milliseconds
(88316 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

Explanation:

- Load Delayed flights data.
- Fetch departure delay and origin.
- Filter data set which is not null.
- Group by origin
- Calculated AVG of dep delay against each origin.
- Sort by avg dep delay DESC
- Fetch top 10
- Join with airports based on origin to fetch airport details.
- Dump the results.

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

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

A = load '/home/acadgild/Desktop/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;
```

Output:

```

Applications Places System acadgild@localhost:~
File Edit View Search Terminal Help
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,483 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,485 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,536 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,543 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,549 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,584 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,595 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,598 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-10-10 09:34:01,694 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-10-10 09:34:01,818 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-10-10 09:34:01,818 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-10 09:34:01,938 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-10 09:34:01,938 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 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)
2018-10-10 09:34:02,248 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 1 second and 888 milliseconds (6
1888 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

Explanation:

- Load Filghts delay data.
- Filter by diversion ==1
- Group by data
- Calculate count
- Order data
- Limit top 10
- Dump the result.
-