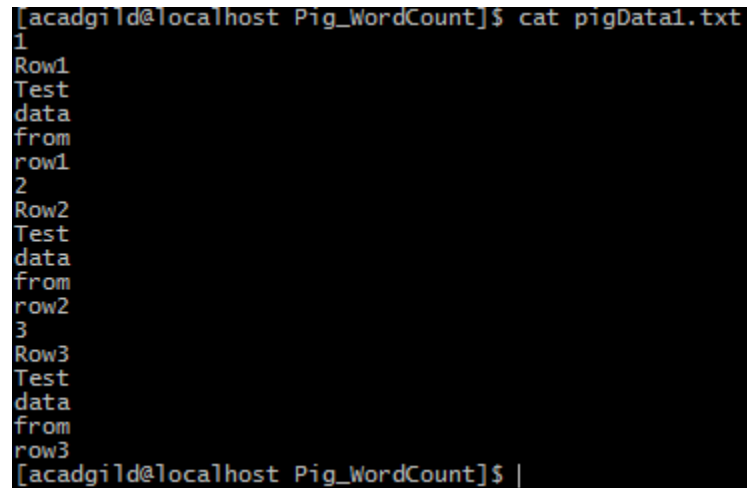


Task 1:

Write a program to implement wordcount using Pig.

File:

A terminal window with a black background and green text. The prompt is [acadgild@localhost Pig\_WordCount]\$ and the command is cat pigData1.txt. The output shows three rows of data, each starting with a row number (1, 2, 3) and a label (Row1, Row2, Row3), followed by the words 'Test', 'data', and 'from' on separate lines.

```
[acadgild@localhost Pig_WordCount]$ cat pigData1.txt
1
Row1
Test
data
from
row1
2
Row2
Test
data
from
row2
3
Row3
Test
data
from
row3
[acadgild@localhost Pig_WordCount]$ |
```

Script:

```
myLine = LOAD '/home/acadgild/myCode/Pig_wordCount/pigData1.txt' AS
(line:chararray);

GroupWords = Group myLine by line;

myCount = FOREACH GroupWords GENERATE group, COUNT(myLine.line);
dump myCount;

countTest = FILTER myLine BY line == 'Test';
GroupTest = Group countTest by line;
DUMP GroupTest;
```

```

2018-10-10 21:58:36,359 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 21:58:36,360 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 21:58:36,384 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-10 21:58:36,391 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-10 21:58:36,392 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-10 21:58:36,417 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-10 21:58:36,417 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,1)
(2,1)
(3,1)
(Row1,1)
(Row2,1)
(Row3,1)
(Test,3)
(data,1)
(from,1)
(row1,1)
(row2,1)
(data,2)
(from,2)
(row3,1)

```

## 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](https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee_details.txt)

employee\_expenses(EmpID,Expenditure)

[https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/employee\\_expenses.txt](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)

Script:

```
1 myLine = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_details.txt' U
  SING PigStorage(',') AS (EmpID:int,Name:chararray,Salary:int,EmployeeRating:
    int);
2
3 OrderRatings = order myLine by EmployeeRating DESC;
4
5 --! dump OrderRatings;
6
7 Top5Emp = Limit OrderRatings 5;
8
9 dump Top5Emp;
10
11
~
~
~
```

Results:

```
2018-10-10 22:49:42,541 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,555 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,559 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,560 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,572 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,575 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,575 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-10 22:49:42,584 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRe
duceLauncher - Success!
2018-10-10 22:49:42,594 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name
is deprecated. Instead, use fs.defaultFS
2018-10-10 22:49:42,595 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has al
ready been initialized
2018-10-10 22:49:42,614 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total inpu
t paths to process : 1
2018-10-10 22:49:42,614 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - To
tal input paths to process : 1
(110,Priyanka,2000,5)
(105,Pawan,2500,5)
(109,Katrina,1000,4)
(104,Anubhav,5000,4)
(108,Ranbir,14000,3)
2018-10-10 22:49:42,693 [main] INFO org.apache.pig.Main - Pig script completed in 9 seconds and 961 mi
lliseconds (9961 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Pig_Assignment1]$ |
```

Big Data Hadoop and Spark Development

(b) Top 3 employees (employee id and employee name) with highest salary, whose employee id

Script:

Results:

(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)

```
1 myEmpDetail = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_details.txt' USING PigStorage(',')
2 ) AS (EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);
3 myEmpExp = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_expenses.txt' USING PigStorage('\t')
4 AS (EEmpID:int,Expense:int);
5 dump myEmpDetail;
6 dump myEmpExp;
7
8
9 RelateTables = JOIN myEmpDetail by EmpID, myEmpExp by EEmpID;
10 dump RelateTables;
11
12 GroupTableOrder = order RelateTables by Expense DESC;
13 dump GroupTableOrder;
14 describe GroupTableOrder;
15
16 myExp = FILTER GroupTableOrder by Expense==400;
17 dump myExp;
18
19 FocusColumn = FOREACH myExp GENERATE $0, $1;
20 dump FocusColumn;|
```

20,18 A11

Results:

```
2018-10-11 01:26:52,343 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,348 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,348 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,353 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,353 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,357 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,363 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,363 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,367 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM M
etrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:26:52,371 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRe
duceLauncher - Success!
2018-10-11 01:26:52,373 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name
is deprecated. Instead, use fs.defaultFS
2018-10-11 01:26:52,373 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has al
ready been initialized
2018-10-11 01:26:52,390 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total inpu
t paths to process : 1
2018-10-11 01:26:52,390 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - To
tal input paths to process : 1
(110,Priyanka)
(102,Shahrukh)
2018-10-11 01:26:52,454 [main] INFO org.apache.pig.Main - Pig script completed in 17 seconds and 525 m
illiseconds (17525 ms)
[acadgild@localhost Pig_Assignment1]$ |
```

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

Script:

```

1 myEmpDetail = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_details.txt' USING PigStorage(',')
  AS (EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);
2
3 myEmpExp = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_expenses.txt' USING PigStorage('\t')
  AS (EEmpID:int,Expense:int);
4
5 dump myEmpDetail;
6 dump myEmpExp;
7
8
9 RelateTables = JOIN myEmpDetail by EmpID, myEmpExp by EEmpID;
0 dump RelateTables;
1

```

Result:

```

Job DAG:
job_local579272951_0003

2018-10-11 01:29:40,543 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:29:40,546 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:29:40,548 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:29:40,564 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-11 01:29:40,567 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-11 01:29:40,567 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-11 01:29:40,587 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-11 01:29:40,587 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh,20000,1,101,100)
(101,Amitabh,20000,1,101,200)
(102,Shahrukh,10000,2,102,400)
(102,Shahrukh,10000,2,102,100)
(104,Anubhav,5000,4,104,300)
(105,Pawan,2500,5,105,100)
(110,Priyanka,2000,5,110,400)
(114,Madhuri,2000,2,114,200)
2018-10-11 01:29:40,663 [main] INFO org.apache.pig.Main - Pig script completed in 9 seconds and 877 milliseconds (9877 ms)
You have new mail in /var/spool/mail/acadgild

```

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

Script:

```
myEmpDetail = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_details.txt' USING PigStorage(',') AS
(EmpID:int,Name:chararray,Salary:int,EmployeeRating:int);

myEmpExp = LOAD '/home/acadgild/myCode/Pig_Assignment1/employee_expenses.txt' USING PigStorage('\t') AS
(EEmpID:int,Expense:int);

dump myEmpDetail;
dump myEmpExp;

RelateTables = JOIN myEmpDetail by EmpID LEFT OUTER, myEmpExp by EEmpID;
dump RelateTables;
describe RelateTables;

mOut = FILTER RelateTables by $4 is null;
dump mOut;

~
~
~
~
~
~
"EntryNOTFile.pig" 15L, 507C written                                13,32      A11
```

Results:

```
Job DAG:
job_local390348750_0004

2018-10-11 01:41:27,569 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:41:27,576 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:41:27,579 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 01:41:27,586 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-11 01:41:27,590 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-11 01:41:27,591 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-11 01:41:27,610 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-11 01:41:27,610 [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,Ajay,5000,2,,)
(113,Jubeen,1000,1,,)
2018-10-11 01:41:27,681 [main] INFO org.apache.pig.Main - Pig script completed in 11 seconds and 233 milliseconds (11233 ms)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Pig_Assignment1]$ |
```

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/>



## Use Case 1:

### Steps:

1. Created Working Directory
2. Copy dataset from blog to Working directory
3. Copy jar file into working directory
4. Typed Script
5. Execute pig locally with script

### Script:

```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar'
A = load '/home/acadgild/myCode/Pig_Assignment1/airline_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/myCode/Pig_Assignment1/airline_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;
```

### Results:

```
2018-10-11 02:16:37,013 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 02:16:37,016 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 02:16:37,018 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2018-10-11 02:16:37,025 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-11 02:16:37,033 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-11 02:16:37,033 [main] WARN org.apache.pig.data.SchemaTupleBackend - xSchemaTupleBackend has already been initialized
2018-10-11 02:16:37,050 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-11 02:16:37,051 [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)
2018-10-11 02:16:37,166 [main] INFO org.apache.pig.Main - Pig script completed in 37 seconds and 907 milliseconds (37907 ms)
You have new mail in /var/spool/mail/acadgild
acadgild@localhost: airline_usecase$
```

## Use Case 2:

### Script:







```

1 REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar'
2 A = load '/home/acadgild/myCode/Pig_Assignment1/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
3
4 B = foreach A Generate (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
5
6 C = filter B by (origin is not null) AND (dest is not null) AND (diversion == 1);
7
8 D = group C by (origin,dest) ;
9
10 E = foreach D generate group, COUNT(C.diversion);
11
12 F = order E by $1 DESC;
13
14 Result = limit F 10;
15
16 dump Result ;
17

```

Result:

```

eayd initialized
2018-10-11 02:49:54,563 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - al
eayd initialized
2018-10-11 02:49:54,566 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - al
eayd initialized
2018-10-11 02:49:54,567 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - al
eayd initialized
2018-10-11 02:49:54,578 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - al
eayd initialized
2018-10-11 02:49:54,581 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - al
eayd initialized
2018-10-11 02:49:54,582 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - al
eayd initialized
2018-10-11 02:49:54,589 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-10-11 02:49:54,593 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-10-11 02:49:54,593 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-10-11 02:49:54,614 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-10-11 02:49:54,614 [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)
2018-10-11 02:49:54,706 [main] INFO org.apache.pig.Main - Pig script completed in 26 seconds and 513 milliseconds (26513 ms)
[acadgild@localhost airline_uscases]$

```