# Assignment 7.1
# Exploring Apache Pig

## Task 1

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

We are using **test.txt** file present in HDFS for wordcount pig script.
In below screenshot, you could see content of test.txt using hadoop fs -cat command.

```
[acadgild@localhost ~]$ hadoop fs -cat /test.txt
18/08/04 12:26:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
My name is Sachin Gorade
I have joined Acadgild Hadoop online course
I am from Hadoop June batch commenced from 2nd June
I have successfully completed six Assignments
It was good learning from all these Assignments
I am from Mumbai, India.[acadgild@localhost ~]$
```

Below you could see content of wordcount.pig script:

```
[acadgild@localhost ~]$ cat wordcount.pig;
A = load '/test.txt';


B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;

C = group B by word;

D = foreach C generate group, COUNT(B);


dump D;
```

We are executing pig script in terminal below using command: **pig wordcount.pig**

```
[acadgild@localhost ~]$ pig wordcount.pig;
18/08/04 12:09:48 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/04 12:09:48 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/04 12:09:48 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-04 12:09:49,035 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-08-04 12:09:49,035 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1533364789032.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-08-04 12:09:50,137 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
2018-08-04 12:09:50,720 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-08-04 12:09:51,102 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use ma
preduce.jobtracker.address
2018-08-04 12:09:51,102 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-04 12:09:51,102 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: hdfs://localhost:8020
2018-08-04 12:09:52,370 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordcount.pig-7c427b19-3f05-413d-84ee-
2227799ef9a1
```

It gives output as in below screenshot :

Here we are getting in output : each word and it's corresponding count.

e.g. for **from** word, we are getting 4 as count as **from** word occurred four times in the test.txt file.

Assignment 7.1
Exploring Apache Pig

```
2018-08-04 12:11:01,795 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-08-04 12:11:01,866 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-04 12:11:01,866 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(I,4)
(It,1)
(My,1)
(am,2)
(is,1)
(2nd,1)
(all,1)
(six,1)
(was,1)
(June,2)
(from,4)
(good,1)
(have,2)
(name,1)
(batch,1)
(these,1)
(Gorade,1)
(Hadoop,2)
(India.,1)
(Mumbai,1)
(Sachin,1)
(course,1)
(joined,1)
(online,1)
(Acadgild,1)
(learning,1)
(commenced,1)
(completed,1)
(Assignments,2)
(successfully,1)
2018-08-04 12:11:02,418 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 14 seconds and 61 milliseconds (74061 ms)
```

## 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)**
**employee_expenses(EmpID,Expence)**

**(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)**

Here, we have stored content of employee_details file in relation A. Then we have sorted Employee Rating in descending order and Name in Ascending order.
Then by using **limit** operator we have restricted these records to top 5.
After this, we are fetching only Employee ID and Name.

**Below is the content of task2_query1.pig file :**

```
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);
B = order A by EmployeeRating desc,Name asc;
C = limit B 5;
D = foreach C generate EmpID,Name;
Dump D;
```

# Assignment 7.1
# Exploring Apache Pig

```
[acadgild@localhost ~]$ hadoop fs -cat /task2_query1.pig
18/08/05 20:35:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

B = order A by EmployeeRating desc,Name asc;

C = limit B 5;

D = foreach C generate EmpID,Name;

Dump D;
```

We have executed pig quiz1.file using command : **pig task2_query1.pig** .

```
[acadgild@localhost ~]$ pig task2_query1.pig;
18/08/04 19:09:43 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/04 19:09:43 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/04 19:09:43 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-04 19:09:43,365 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-08-04 19:09:43,365 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1533389983362.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-08-04 19:09:45,161 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
2018-08-04 19:09:45,939 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-08-04 19:09:46,564 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use ma
preduce.jobtracker.address
2018-08-04 19:09:46,564 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-04 19:09:46,564 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: hdfs://localhost:8020
2018-08-04 19:09:48,341 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-task2_query1.pig-170db084-a61e-40a3-93
4b-2d960446e850
2018-08-04 19:09:48,341 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-08-04 19:09:49,947 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-04 19:09:50,494 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,LIMIT
2018-08-04 19:09:50,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
```

Below is the final output which shows Top 5 records with Highest Rating in descending order and their Name in Ascending order.

```
2018-08-04 19:13:35,261 [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)
2018-08-04 19:13:35,504 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 53 seconds and 478 milliseconds (233478 ms)
```

**(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)**

Here, we have stored content of employee_details file in relation A.
After this, we have taken only those records having odd Employee ID by using **filter A by EmpID%2 == 1.**
Then we have sorted Employee Rating in descending order and Name in Ascending order.
Then by using **limit** operator we have restricted these records to top 5.
After this, we are fetching only Employee ID and Name.

# Assignment 7.1
# Exploring Apache Pig

**Below is the content of task2_query2.pig file :**

A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

B = filter A by EmpID%2 == 1;

C = order B by EmployeeRating desc,Name asc;

D = limit C 3;

E = foreach D generate EmpID,Name;

Dump E;

```
[acadgild@localhost ~]$ hadoop fs -cat /task2_query2.pig;
18/08/04 19:32:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

B = filter A by EmpID%2 == 1;

C = order B by EmployeeRating desc,Name asc;

D = limit C 3;

E = foreach D generate EmpID,Name;

Dump E;
```

We have executed pig script using command : **pig task2_query2.pig** .

```
[acadgild@localhost ~]$ pig task2_query2.pig;
18/08/04 19:32:24 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/04 19:32:24 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/04 19:32:24 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-04 19:32:24,727 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-08-04 19:32:24,727 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1533391344720.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-08-04 19:32:25,907 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
2018-08-04 19:32:26,490 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-08-04 19:32:26,918 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use ma
preduce.jobtracker.address
2018-08-04 19:32:26,918 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-04 19:32:26,918 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: hdfs://localhost:8020
2018-08-04 19:32:28,245 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-task2_query2.pig-29349d54-f9ac-4b68-bd
20-443ffa96d881
2018-08-04 19:32:28,248 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-08-04 19:32:30,029 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-04 19:32:30,634 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIMIT
```

Below is the final output which shows Top 3 records having odd Employee ID with Highest Rating in descending order and their Name in Ascending order.

```
2018-08-04 19:36:02,651 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-04 19:36:02,651 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(105,Pawan)
(109,Katrina)
(103,Akshay)
2018-08-04 19:36:03,019 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 39 seconds and 29 milliseconds (219029 ms)
```

# Assignment 7.1
# Exploring Apache Pig

**(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)**

**Below is the content of task2_query3.pig file :**

```
[acadgild@localhost ~]$ hadoop fs -cat /task2_query3.pig
18/08/05 19:43:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

X = load '/employee_expenses.txt' using PigStorage('\t') as (Emp_ID:int, Expense:int);

combine = join A by EmpID , X by Emp_ID;

Descexpense = order combine by Expense desc,Name asc;

final = limit Descexpense 1;

finally = foreach final generate EmpID,Name;

dump finally;
```

We have executed pig script using command : **pig task2_query3.pig** .

```
[acadgild@localhost ~]$ pig task2_query3.pig;
18/08/05 19:41:52 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/05 19:41:52 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/05 19:41:52 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-05 19:41:52,390 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-08-05 19:41:52,391 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1533478312377.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-08-05 19:41:53,634 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
2018-08-05 19:41:54,252 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-08-05 19:41:54,675 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use ma
preduce.jobtracker.address
2018-08-05 19:41:54,675 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 19:41:54,675 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: hdfs://localhost:8020
2018-08-05 19:41:55,906 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-task2_query3.pig-6f77206c-f481-4fb6-88
86-c5279142519b
2018-08-05 19:41:55,909 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-08-05 19:41:57,355 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 19:41:57,691 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
```

Below is the final output which shows employee id and employee name having Maximum expense.

```
tus=SUCCEEDED. Redirecting to job history server
2018-08-05 19:45:39,082 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-08-05 19:45:39,099 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-08-05 19:45:39,200 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-08-05 19:45:39,212 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2018-08-05 19:45:39,318 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-08-05 19:45:39,325 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 19:45:39,326 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-08-05 19:45:39,333 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-05 19:45:39,333 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priyanka)
2018-08-05 19:45:39,560 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 48 seconds and 42 milliseconds (228042 ms)
```

# Assignment 7.1
# Exploring Apache Pig

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

Here, we have stored content of employee_details file in relation A and content of employee_expenses file into relation X.
After this, we have joined A and X by EmpID field.
Then we are fetching only Employee ID and Name and taking only distinct values.

**Below is the content of task2_query4.pig file :**

```
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

X = load '/employee_expenses.txt' using PigStorage('\t') as (Emp_ID:int, TotalExpenses:int);

combine = join A by EmpID , X by Emp_ID;

final = foreach combine generate EmpID,Name;

finally = distinct final;

dump finally;
```

```
applicable
[acadgild@localhost ~]$ hadoop fs -cat /task2_query4.pig
18/08/05 14:32:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

X = load '/employee_expenses.txt' using PigStorage('\t') as (Emp_ID:int, TotalExpenses:int);

combine = join A by EmpID , X by Emp_ID;

final = foreach combine generate EmpID,Name;

finally = distinct final;

dump finally;
```

We have executed pig script using command : **pig task2_query4.pig** .

```
[acadgild@localhost ~]$ pig task2_query4.pig;
18/08/05 14:32:39 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/05 14:32:39 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/05 14:32:39 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-05 14:32:40,136 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-08-05 14:32:40,136 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1533459760131.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-08-05 14:32:41,382 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
2018-08-05 14:32:42,122 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-08-05 14:32:42,573 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use ma
preduce.jobtracker.address
2018-08-05 14:32:42,573 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 14:32:42,574 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at
: hdfs://localhost:8020
2018-08-05 14:32:43,957 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-task2_query4.pig-b4fd7b6a-67c3-4ce3-94
93-796c13fa8350
2018-08-05 14:32:43,957 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2018-08-05 14:32:45,468 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 14:32:45,777 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
```

# Assignment 7.1
# Exploring Apache Pig

Below is the final output which shows employee id and employee name in Employee_details file which are also available in Employee_expenses file.

```
2018-08-05 14:34:38,423 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-05 14:34:38,425 [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)
2018-08-05 14:34:38,733 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 59 seconds and 470 milliseconds (119470 ms)
```

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

**Below is the content of task2_query5.pig file :**

```
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

X = load '/employee_expenses.txt' using PigStorage('\t') as (Emp_ID:int, TotalExpenses:int);

nojoin = join A by EmpID LEFT OUTER, X by Emp_ID;

final = filter nojoin by Emp_ID is NULL;

finally = foreach final generate EmpID,Name;

dump finally;
```

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /task2_query5.pig
18/08/05 14:20:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
A = load '/employee_details.txt' using PigStorage(',') as (EmpID:int, Name:chararray, Salary:int, EmployeeRating:int);

X = load '/employee_expenses.txt' using PigStorage('\t') as (Emp_ID:int, TotalExpenses:int);

nojoin = join A by EmpID LEFT OUTER, X by Emp_ID;

final = filter nojoin by Emp_ID is NULL;

finally = foreach final generate EmpID,Name;

dump finally;You have new mail in /var/spool/mail/acadgild
```

We have executed pig script using command : **pig task2_query5.pig** .

```
applicable
[acadgild@localhost ~]$ pig task2_query5.pig
18/08/05 14:10:07 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/05 14:10:07 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/05 14:10:07 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-05 14:10:07,577 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-08-05 14:10:07,578 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1533458407565.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-08-05 14:10:08,889 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...
```

# Assignment 7.1
# Exploring Apache Pig

Below is the final output which shows employee id and employee name in Employee_details file which are not available in Employee_expenses file.

```
2018-08-05 14:11:29,368 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-08-05 14:11:29,378 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 14:11:29,380 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2018-08-05 14:11:29,495 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-05 14:11:29,495 [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)
2018-08-05 14:11:29,984 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 23 seconds and 352 milliseconds (83352 ms)
```

**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/

We have used local mode by using command : **pig -x local**

```
[acadgild@localhost ~]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder
.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
```

1) **Find out the top 5 most visited destinations.**

Please find below all steps performed and final output :

```
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
2018-08-05 20:27:00,629 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2018-08-05 20:27:00,629 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
grunt> C = filter B by dest is not null;
grunt> D = group C by dest;
grunt> E = foreach D generate group, COUNT(C.dest);
grunt> F = order E by $1 DESC;
grunt> Result = LIMIT F 5;
grunt> A1 = load '/home/acadgild/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INP
UT_HEADER');
2018-08-05 20:27:56,193 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2018-08-05 20:27:56,193 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
grunt> joined_table = join Result by $0, A2 by dest;
grunt> dump joined_table;
2018-08-05 20:28:08,856 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDE
```

```
2018-08-05 19:17:06,643 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-05 19:17:06,646 [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)
2018-08-05 19:17:07,035 [main] INFO  org.apache.pig.Main - Pig script completed in 6 minutes, 3 seconds and 743 milliseconds (363743 ms)
```

# Assignment 7.1
# Exploring Apache Pig

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

Please find below all steps performed and final output :

```
grunt> REGISTER '/home/acadgild/Desktop/piggybank.jar';
2018-08-05 19:56:49,599 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2018-08-05 19:56:49,599 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
2018-08-05 19:57:05,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2018-08-05 19:57:05,530 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt>
grunt> B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt> C = filter B by cancelled == 1 AND cancel_code =='B';
grunt> D = group C by month;D = group C by month;
grunt> E = foreach D generate group, COUNT(C.cancelled);
grunt> F= order E by $1 DESC;
grunt> Result = limit F 1;
grunt> dump Result;
```

**Please find below final Output :**

```
2018-08-05 20:00:09,574 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2018-08-05 20:00:09,583 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2018-08-05 20:00:09,599 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2018-08-05 20:00:09,600 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2018-08-05 20:00:09,609 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker
, sessionId= - already initialized
2018-08-05 20:00:09,629 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-08-05 20:00:09,638 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2018-08-05 20:00:09,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2018-08-05 20:00:09,640 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-08-05 20:00:09,661 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-05 20:00:09,662 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt>
```

# Assignment 7.1
# Exploring Apache Pig

### 3. Top 10 origins with the highest AVG departure delay.

Please find below all steps performed and final output :

```
grunt> REGISTER '/home/acadgild/Desktop/piggybank.jar';
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-08-05 20:08:01,238 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-08-05 20:08:01,238 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;
grunt> C1 = filter B1 by (dep_delay is not null) AND (origin is not null);
grunt> D1 = group C1 by origin;
grunt> E1 = foreach D1 generate group, AVG(C1.dep_delay);
grunt> Result = order E1 by $1 DESC;
grunt> Top_ten = limit Result 10;
grunt> Lookup = load '/home/acadgild/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-08-05 20:09:22,902 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-08-05 20:09:22,905 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
grunt> Joined = join Lookup1 by origin, Top_ten by $0;
grunt> Final = foreach Joined generate $0,$1,$2,$4;
grunt> Final_Result = ORDER Final by $3 DESC;
grunt> dump Final_Result;
```

### Please find below final Output :

```
2018-08-05 20:11:22,744 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-08-05 20:11:22,744 [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)
(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)
```

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

### Please find below all steps performed and final output :

```
grunt> REGISTER '/home/acadgild/Desktop/piggybank.jar';
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-08-05 20:15:38,523 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-08-05 20:15:38,523 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
grunt> D = GROUP C by (origin,dest);
grunt> E = FOREACH D generate group, COUNT(C.diversion);
grunt> F = ORDER E BY $1 DESC;
grunt> Result = limit F 10;
grunt> dump Result;
```

### Please find below final Output :

```
2018-08-05 20:18:42,272 [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)
grunt>
```