

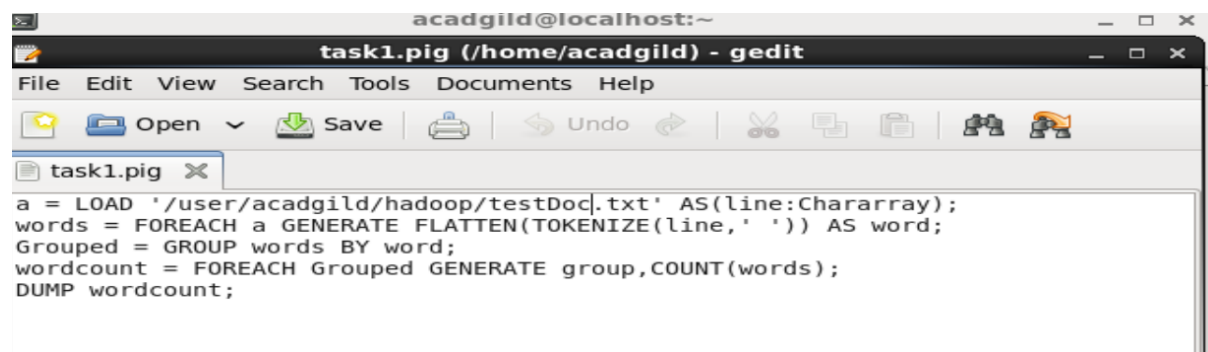
## Session 7: EXPLORING APACHE PIG

### Assignment 7

#### Task 1:

Write a program to implement wordcount using Pig.

#### Solution:

A screenshot of a gedit window titled 'task1.pig (/home/acadgild) - gedit'. The window shows a Pig Latin script for wordcount. The script is as follows:

```
a = LOAD '/user/acadgild/hadoop/testDoc.txt' AS (line:Chararray);
words = FOREACH a GENERATE FLATTEN(TOKENIZE(line, ' ')) AS word;
Grouped = GROUP words BY word;
wordcount = FOREACH Grouped GENERATE group, COUNT(words);
DUMP wordcount;
```

#### Output:

```
2018-12-15 03:58:22,020 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-12-15 03:58:22,827 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(is,1)
(for,1)
(file,1)
(test,2)
(this,1)
(word,1)
(count,1)
2018-12-15 03:58:23,078 [main] INFO org.apache.pig.Main - Pig script completed
in 3 minutes, 11 seconds and 630 milliseconds (191630 ms)
You have new mail in /var/spool/mail/acadgild
```

#### 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)

**Solution:**

```
Employee = LOAD '/user/acadgild/hadoop/employee_details.txt' USING PigStorage(',') AS  
(EmpID:int, EmpName:CharArray, Salary:int, Rating:int);
```

```
Order_data = ORDER Employee BY Rating DESC,EmpName ASC;
```

```
Top_data = LIMIT Order_data 5;
```

```
Result = FOREACH Top_data GENERATE EmpID,EmpName;
```

```
DUMP Result;
```

**Output:**

```
2018-12-15 06:48:38,186 [m  
2018-12-15 06:48:38,186 [m  
cess : 1  
(105,Pawan)  
(110,Priyanka)  
(104,Anubhav)  
(109,Katrina)  
(103,Akshay)  
grunt>
```

(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)

**Solution:**

```
*task22.pig X  
employee = LOAD '/user/acadgild/hadoop/employee_details.txt' USING PigStorage(',') AS( EmpID:int, EmpName:CharArray,  
Salary:int,Rating:int);  
filter_data = FILTER employee BY EmpID%2==1;  
order_data = ORDER filter_data BY Salary DESC,EmpName ASC;  
result = LIMIT order_data 3;  
final_result = FOREACH result GENERATE EmpID,EmpName;  
DUMP final_result;
```

**Output:**

```
2018-12-15 07:06:26,301 [main]  
2018-12-15 07:06:26,302 [main]  
cess : 1  
(101,Amitabh)  
(107,Salman)  
(103,Akshay)  
2018-12-15 07:06:26,500 [main]
```

(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)

**Solution:**

```
task23.pig
employee = LOAD '/user/acadgild/hadoop/employee_details.txt' USING PigStorage(',') AS( EmpID:int, EmpName:CharArray, Salary:int,Rating:int);
expense = LOAD '/user/acadgild/hadoop/employee_expenses.txt' USING PigStorage('\t') AS( ID:int,Exp:double);
group_data = GROUP expense BY ID;
sum_data = FOREACH group_data GENERATE group,SUM(expense.Exp) as final_expense;
emp_expense = JOIN employee BY EmpID,sum_data BY group;
top_expense = ORDER emp_expense BY final_expense DESC,EmpName ASC;
result = LIMIT top_expense 1;
final_result = FOREACH result GENERATE EmpID,EmpName;
DUMP final_result;
```

**Output:**

```
-----
2018-12-15 07:55:15,568 [main] INFO org.ap
2018-12-15 07:55:15,568 [main] INFO org.ap
cess : 1
(102,Shahrukh)
2018-12-15 07:55:15,806 [main] INFO org.ap
-----
```

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

**Solution:**

```
task23.pig
employee = LOAD '/user/acadgild/hadoop/employee_details.txt' USING PigStorage(',') AS( EmpID:int, EmpName:CharArray, Salary:int,Rating:int);
expense = LOAD '/user/acadgild/hadoop/employee_expenses.txt' USING PigStorage('\t') AS( ID:int,Exp:double);
group_data = GROUP expense BY ID;
sum_data = FOREACH group_data GENERATE group,SUM(expense.Exp) as final_expense;
emp_expense = JOIN employee BY EmpID,sum_data BY group;
final_result = FOREACH emp_expense GENERATE EmpID,EmpName;
DUMP final_result;
```

**Output:**

```
-----
cess : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
2018-12-15 08:01:47,654 [main] IN
-----
```

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

**Solution:**

```
task23.pig
employee = LOAD '/user/acadgild/hadoop/employee_details.txt' USING PigStorage(',') AS( EmpID:int, EmpName:CharArray, Salary:int,Rating:int);
expense = LOAD '/user/acadgild/hadoop/employee_expenses.txt' USING PigStorage('\t') AS( ID:int,Exp:double);
group_data = GROUP expense BY ID;
sum_data = FOREACH group_data GENERATE group,SUM(expense.Exp) as final_expense;
emp_expense = JOIN employee BY EmpID LEFT OUTER,sum_data BY group;
result = FILTER emp_expense BY group is null;
final_result = FOREACH result GENERATE EmpID,EmpName;
DUMP final_result;
```

### Output:

```
2018-12-15 08:13:24,416 [main] I  
cess : 1  
(103,Akshay)  
(106,Aamir)  
(107,Salman)  
(108,Ranbir)  
(109,Katrina)  
(111,Tushar)  
(112,Ajay)  
(113,Jubeen)  
2018-12-15 08:13:24,613 [main] I
```

### 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/>

### Solution:

Couldn't download the csv files from the given link in the blog and got the below error:



**Sorry, the file you have requested does not exist.**

Make sure that you have the correct URL and the file exists.

#### **Get stuff done with Google Drive**

Apps in Google Drive make it easy to create, store and share online documents, spreadsheets, presentations and more.

Learn more at [drive.google.com/start/apps](https://drive.google.com/start/apps).