Assignment # 7.1 – EXPLORING APACHE PIG

Date: 10-Nov-2018

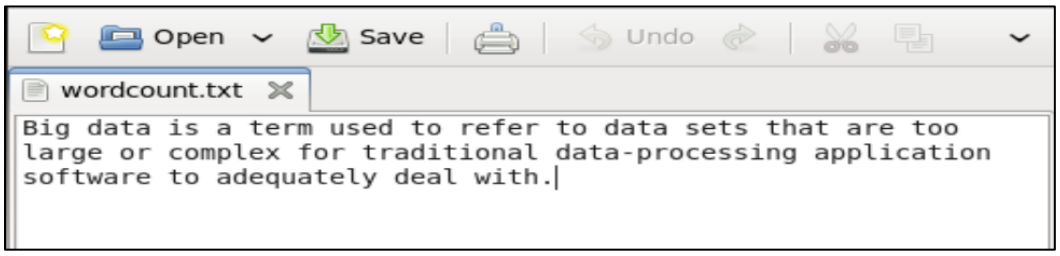
TASK 1:

Task:

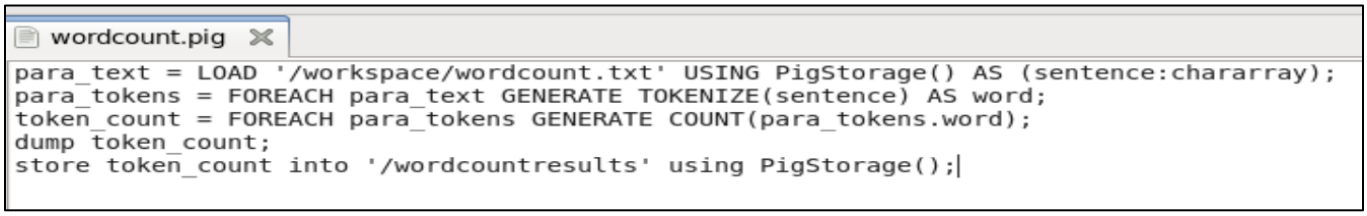
- Write a program to implement word count using Pig

Explanation:

- Create a data file 'wordcount.txt' and place it in local folder path



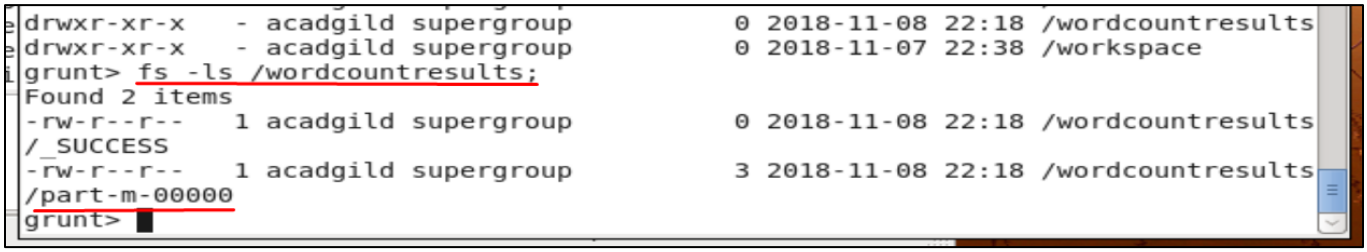
- Create a pig script 'wordcount.pig' with below pig commands (save the file in bin folder of Pig),



- Save it and execute with command $ ./pig 'wordcount.pig'
- Pig script will be executed in mapreduce mode and the result will be stored in /wordcountresults folder in HDFS

Output:

Below is the screenshot of the commands executed in terminal.



The final result is 26 that is the count of words in the data file 'wordcount.txt'

```
/part-m-00000
grunt> fs -cat /wordcountresults/part-m-00000;
26
grunt>
```

TASK 2:

Task:

- 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

  employee_expenses(EmpID,Expence)
  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)

  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)

  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)

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

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

Explanation:

- Download employee_details.txt from the above mentioned location and the data are as below. The file is stored in local file system

```
employee_details.txt
101,Amitabh,20000,1
102,Shahrukh,10000,2
103,Akshay,11000,3
104,Anubhav,5000,4
105,Pawan,2500,5
106,Aamir,25000,1
107,Salman,17500,2
108,Ranbir,14000,3
109,Katrina,1000,4
110,Priyanka,2000,5
111,Tushar,500,1
112,Ajay,5000,2
113,Jubeen,1000,1
114,Madhuri,2000,2
```

- Create a pig latin script with commands as below and the script is stored in local file system,

```
orderedempdetails.pig | employee_details.txt | employee_expenses.txt | *Unsaved Document 1
emp_details = LOAD '/home/acadgild/Workspace/employee_details.txt' USING PigStorage(',') AS
(empid:int,empname:chararray,salary:int,emprating:int);
ordered_emp_details = ORDER emp_details BY emprating DESC, empname DESC;
limit_ord_emp_details = LIMIT ordered_emp_details 5;
DUMP limit_ord_emp_details;
STORE limit_ord_emp_details INTO '/home/acadgild/Workspace/emporderresults' USING PigStorage('|');
```

- In the above script, ORDER emp_details BY emprating DESC, empname DESC will order the records in descending order of emprating and empname. LIMIT ordered_emp_details 5 will return only top 5 records of the dataset.
- To run pig script from localhost, execute below command,

  $ ./pig –x local '/home/acadgild/Workspace/pig/orderedempdetails.pig

Output:

Below is the screenshot of the command executed in terminal.

Output (a):

```
acadgild@localhost:~/install/pig/pig-0.16.0/bin
File  Edit  View  Search  Terminal  Help
/a      n/a     n/a     n/a     ordered_emp_details             /home/acadgild/W
orkspace/emporderresults,
job_local1484511934_0005        1       0       n/a     n/a     n/a     n/a     0
0       0       0       emp_details     MAP_ONLY
job_local1560897425_0007        1       1       n/a     n/a     n/a     n/a     n
/a      n/a     n/a     n/a     ordered_emp_details     ORDER_BY,COMBINER
job_local1708315880_0006        1       1       n/a     n/a     n/a     n/a     n
/a      n/a     n/a     n/a     ordered_emp_details     SAMPLER

Input(s):
Successfully read 14 records from: "/home/acadgild/Workspace/employee_details.tx
t"

Output(s):
Successfully stored 5 records in: "/home/acadgild/Workspace/emporderresults"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
```

```
part-r-00000   employee
110|Priyanka|2000|5
105|Pawan|2500|5
109|Katrina|1000|4
104|Anubhav|5000|4
108|Ranbir|14000|3
```

Output (b):

```
acadgild@localhost:~/install/pig/pig-0.16.0/bin
File  Edit  View  Search  Terminal  Help
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost bin]$ ./pig -x local /home/acadgild/Workspace/pig/oddemploye
edetails.pig
```

Code script:

```
orderedempdetails.pig    oddemployeedetails.pig
/* Filepath: /home/acadgild/Workspace/pig/oddemployeedetails.pig */
/* Date: 10-Nov-2018 */
/* Author: Saravanan Ponnaiah */
emp_details = LOAD '/home/acadgild/Workspace/employee_details.txt' USING PigStorage(',') AS (empid:int,empname:chararray,salary:int,emprating:int);
odd_employees = FILTER emp_details BY (empid%2) == 1;
ordered_emp_details = ORDER odd_employees BY salary DESC, empname DESC;
limit_ord_emp_details = LIMIT ordered_emp_details 3;
DUMP limit_ord_emp_details;
STORE limit_ord_emp_details INTO '/home/acadgild/Workspace/oddempresults' USING PigStorage('|');
```

Result:

```
/oddempresults,
job_local674675352_0006 1       1       n/a     n/a     n/a     n/a     n/a     n
/a      n/a     n/a     ordered_emp_details     SAMPLER

Input(s):
Successfully read 14 records from: "/home/acadgild/Workspace/employee_details.tx
t"

Output(s):
Successfully stored 3 records in: "/home/acadgild/Workspace/oddempresults"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

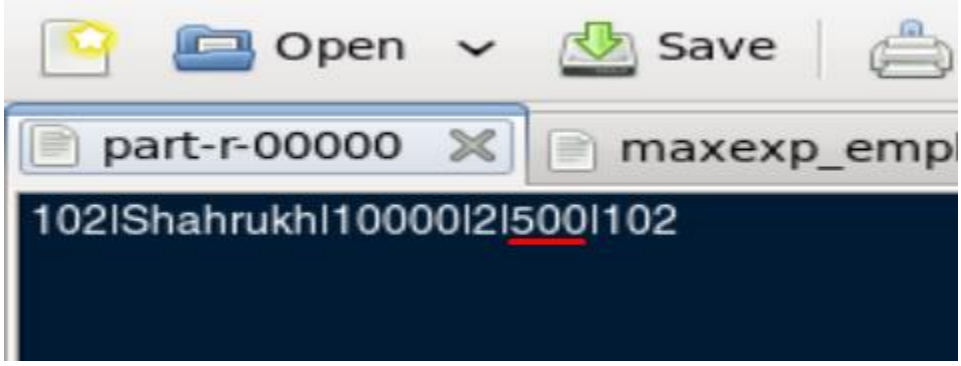
```
part-r-00000    orde
101|Amitabh|20000|1
107|Salman|17500|2
103|Akshay|11000|3
```

Output (c):

Pig Latin script

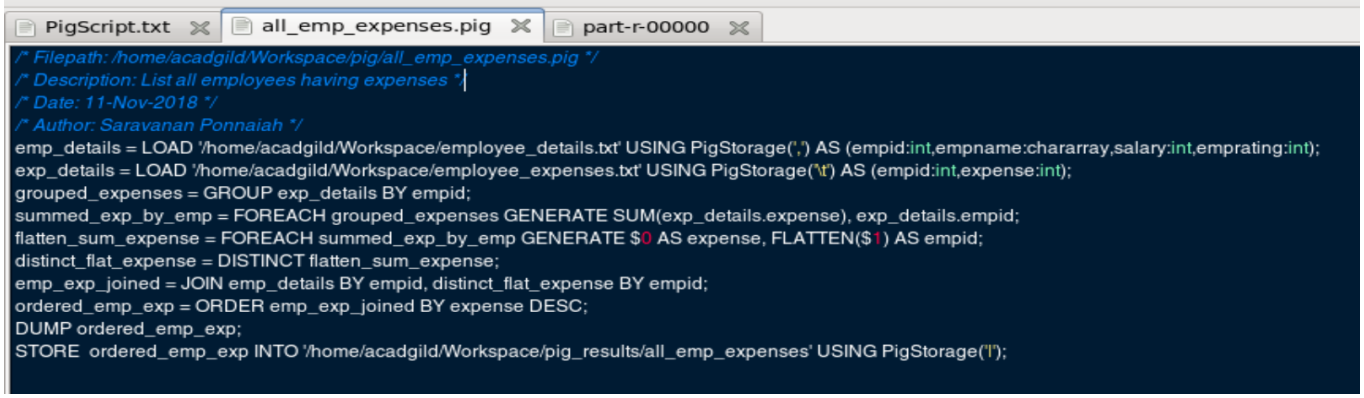
```
maxexp_employee.pig    PigScript.txt    part-r-00000
/* Filepath: /home/acadgild/Workspace/pig/maxexp_employee.pig */
/* Date: 10-Nov-2018 */
/* Author: Saravanan Ponnaiah */
emp_details = LOAD '/home/acadgild/Workspace/employee_details.txt' USING PigStorage(',') AS (empid:int,empname:chararray,salary:int,emprating:int);
exp_details = LOAD '/home/acadgild/Workspace/employee_expenses.txt' USING PigStorage('\t') AS (empid:int,expense:int);
grouped_expenses = GROUP exp_details BY empid;
summed_exp_by_emp = FOREACH grouped_expenses GENERATE SUM(exp_details.expense), exp_details.empid;
flatten_sum_expense = FOREACH summed_exp_by_emp GENERATE $0 AS expense, FLATTEN($1) AS empid;
distinct_flat_expense = DISTINCT flatten_sum_expense;
emp_exp_joined = JOIN emp_details BY empid, distinct_flat_expense BY empid;
ordered_emp_exp = ORDER emp_exp_joined BY expense DESC;
top_expense_emp = LIMIT ordered_emp_exp 1;
DUMP top_expense_emp;
STORE  top_expense_emp INTO '/home/acadgild/Workspace/pig_results/empexpenses' USING PigStorage('|');
```

Result:

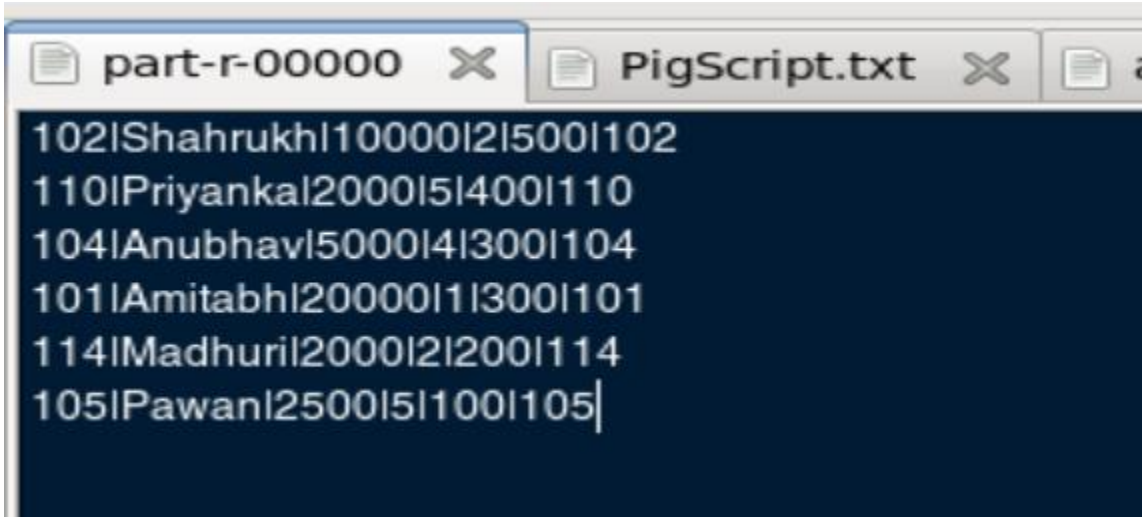


Output (d):

Pig Latin script:



Result:



Output (e):

Pig Latin script:

*PigScript.txt | all_emp_no_expenses.pig | part-r-00000

```
/* Filepath: /home/acadgild/Workspace/pig/all_emp_no_expenses.pig */
/* Description: List all employees having no expenses */
/* Date: 11-Nov-2018 */
/* Author: Saravanan Ponnaiah */
emp_details = LOAD '/home/acadgild/Workspace/employee_details.txt' USING PigStorage(',') AS (empid:int,empname:chararray,salary:int,emprating:int);
exp_details = LOAD '/home/acadgild/Workspace/employee_expenses.txt' USING PigStorage('\t') AS (empid:int,expense:int);
grouped_expenses = GROUP exp_details BY empid;
summed_exp_by_emp = FOREACH grouped_expenses GENERATE SUM(exp_details.expense), exp_details.empid;
flatten_sum_expense = FOREACH summed_exp_by_emp GENERATE $0 AS expense, FLATTEN($1) AS empid;
distinct_flat_expense = DISTINCT flatten_sum_expense;
emp_exp_joined = JOIN emp_details BY empid LEFT, distinct_flat_expense BY empid;
ordered_emp_exp = ORDER emp_exp_joined BY expense DESC;
emp_no_expense = FILTER ordered_emp_exp BY ($4) IS NULL;
DUMP emp_no_expense;
STORE  emp_no_expense INTO '/home/acadgild/Workspace/pig_results/all_emp_no_expenses' USING PigStorage('|');
```

Result:

part-r-00000 | *PigScript.txt

```
113|Jubeen|1000|1||
112|Ajay|5000|2||
111|Tushar|500|1||
109|Katrina|1000|4||
108|Ranbir|14000|3||
107|Salman|17500|2||
106|Aamir|25000|1||
103|Akshay|11000|3||
```

TASK 3:

Task Link: https://acadgild.com/blog/aviation-data-analysis-using-apache-pig

Problem Statement 1:

Find out the top 5 most visited destinations.

Output:

Pig Latin script:

airports_top_dest.pig | PigScript.txt | part-r-00000

```
/* Filepath: /home/acadgild/Workspace/pig/airports_top_dest.pig */
/* Date: 26-Nov-2018 */
/* Author: Saravanan Ponnaiah */
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
delayed_flights = LOAD '/home/acadgild/Workspace/aviation-data/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
airports = LOAD '/home/acadgild/Workspace/aviation-data/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
filtered_flights = FOREACH delayed_flights GENERATE (int)$9 AS flight_num, (chararray)$16 AS origin, (chararray)$17 AS destination;
refined_flights = FILTER filtered_flights BY destination is not null;
grp_dest = GROUP refined_flights BY destination;
calc_grp_dest = FOREACH grp_dest GENERATE group, COUNT(refined_flights.destination);
ordered_dest = ORDER calc_grp_dest BY $1 DESC;
result = LIMIT ordered_dest 5;
refined_airports = FOREACH airports GENERATE (chararray)$0 AS dest, (chararray)$2 AS city, (chararray)$4 AS country;
detail_result = JOIN result BY $0, refined_airports BY dest;
STORE detail_result INTO '/home/acadgild/Workspace/pig_results/airports_top_dest' USING PigStorage('|');
```

Result:

part-r-00000 | airports_top_dest

```
ATL|131613|ATL|Atlanta|USA
DEN|74323|DEN|Denver|USA
DFW|95414|DFW|Dallas-Fort Worth|USA
LAX|58772|LAX|Los Angeles|USA
ORD|125979|ORD|Chicago|USA
```

Problem Statement 2:

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

Output:

Pig Latin script:

airports_cancel_month.pig | *PigScript.txt | *Unsaved Document 1

```
/* Filepath: /home/acadgild/Workspace/pig/airports_cancel_month.pig */
/* Date: 01-Dec-2018 */
/* Author: Saravanan Ponnaiah */
REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
delayed_flights = LOAD '/home/acadgild/Workspace/aviation-data/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
delay_flights = FOREACH delayed_flights GENERATE (int)$2 AS month, (int)$10 AS flight_num, (int)$22 AS cancelled, (chararray)$23 AS cancel_code;
filtered_flights = FILTER delay_flights BY cancelled == 1 AND cancel_code =='B';
grp_flights = GROUP filtered_flights BY month;
grped_flights = FOREACH grp_flights GENERATE group, COUNT(filtered_flights.cancelled);
ord_flights = ORDER grped_flights BY $1 DESC;
result = limit ord_flights 1;
STORE result INTO '/home/acadgild/Workspace/pig_results/airports_cancel_month' USING PigStorage('|');
```

Result:

part-r-00000 | airports

```
12|250
```