Project 1.1 - USA Crime Analysis using Apache Pig

Problem Statement 1

Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code.

Steps to calculate the number of cases investigated under each FBI code.

- **In Line 1**: We are registering the *piggybank* jar in order to use the CSVExcelStorage class.
- **In Line 2:** we are defining org.apache.pig.piggybank.storage.CSVExcelStorage as CSVExcelStorage
- In relation **my_rel**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.
- In relation **grouped_by_fbi**, we are grouping relation **my_rel** by "FBI_Code".
- In relation **result_rel**, we are generating the grouped column and the count of each.
- Finally, using dump, we are printing the result.
- Result prints the FBI Code and no. of cases investigated under each FBI Code.

Pig Scripts

```
REGISTER  /usr/local/pig/contrib/piggybank/java/piggybank.jar;

DEFINE CSVExcelStorage  org.apache.pig.piggybank.storage.CSVExcelStorage;

my_rel = Load '/home/acadgild/pig/Crimes_2001_to_present.csv' USING
CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER') AS
(Id:int,Case:chararray,Date:chararray,Block:chararray,IUCR:chararray,Primary_type:chararray,
Description:chararray,Loc_Desc:chararray, Arrest:chararray, Domestic:chararray, Beat:int,
District:int,Ward:int,Community_Area:int,FBI_Code:chararray, X_coord:int, Y_coord:int,
Year:int,Updated_on:chararray,Lati:double,Longi:double,Location:chararray);

grouped_by_fbi= GROUP my_rel BY FBI_Code;

result_rel = FOREACH grouped_by_fbi GENERATE  group, COUNT(my_rel.Id);

dump result_rel;
```

Screenshot of Output in Pig Grunt Shell:

```
2017-11-11 19:17:06,545 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
(02,1502)
(03,10596)
(05,14842)
(06,64329)
(07,11105)
(09,445)
(10,1551)
(11,13757)
(12,27)
(13,57)
(14,31301)
(15,3694)
(16,1787)
(17,1126)
(18,25207)
(19,434)
(20,1267)
(22,371)
(24,4046)
(26,29474)
(01A,533)
(01B,6)
(04A,4994)
(04B,7711)
(08A,14167)
(08B,46938)
(,1)
grunt>
```

Result:

(02,1502)
(03,10596)
(05,14842)
(06,64329)
(07,11105)
(09,445)
(10,1551)
(11,13757)
(12,27)
(13,57)
(14,31301)
(15,3694)
(16,1787)
(17,1126)
(18,25207)
(19,434)
(20,1267)
(22,371)
(24,4046)
(26,29474)
(01A,533)
(01B,6)
(04A,4994)
(04B,7711)
(08A,14167)
(08B,46938)
(,1)

Problem Statement 2

Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.

Steps to find the number of cases investigated under FBI code 32.

- **In Line 1**: We are registering *piggybank* jar in order to use the CSVExcelStorage class.
- **In Line 2:** we are defining org.apache.pig.piggybank.storage.CSVExcelStorage as CSVExcelStorage
- In relation **my_rel**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.
- In relation **filter_fbi_32**, we are filtering the **my_rel** based on FBI_Code='32'. So relation filter_fbi_32 will point to the data which was investigated user FBI Code 32
- In relation **filtered_grp_32**, we are grouping the data.
- In relation **filtered_count_32**, we are finding the number of cases investigated under FBI code 32.
- Finally, using dump, we are printing the result.

Pig Scripts

```
REGISTER /usr/local/pig/lib/piggybank.jar;

DEFINE CSVExcelStorage  org.apache.pig.piggybank.storage.CSVExcelStorage;

my_rel = Load '/home/acadgild/pig/Crimes_2001_to_present.csv' USING
CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER') AS
(Id:int,Case:chararray,Date:chararray,Block:chararray,IUCR:chararray,Primary_type:chararray,
Description:chararray,Loc_Desc:chararray, Arrest:chararray, Domestic:chararray, Beat:int,
District:int,Ward:int,Community_Area:int,FBI_Code:chararray, X_coord:int, Y_coord:int,
Year:int,Updated_on:chararray,Lati:double,Longi:double,Location:chararray);

filter_fbi_32= FILTER my_rel BY FBI_Code=='32';

filtered_grp_32 = GROUP filter_fbi_32 ALL;

filtered_count_32 = FOREACH filtered_grp_32 GENERATE group, COUNT(filter_fbi_32);

dump filtered_count_32;
```

Output:

```
2017-11-11 19:25:27,308 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
grunt>
```

Result:

After running the pig scripts, the result it empty as shown in the output screenshot above, implies there is no case investigated under FBI_Code=32.

Problem Statement 3

Write a MapReduce/Pig program to calculate the number of arrests in theft district wise.

Steps to calculate the number of arrests in theft district wise.

- **In Line 1**: We are registering *piggybank* jar in order to use the CSVExcelStorage class.
- **In Line 2:** we are defining org.apache.pig.piggybank.storage.CSVExcelStorage as CSVExcelStorage
- In relation **my_rel**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.
- In relation **filter_by_theft**, we are filtering the **my_rel** based on Primary_type='Theft'. So relation filter_by_theft will point to the data which was under primary_type theft.
- In relation **grouped_by_theft**, we are grouping the filter_by_theft, based on column "District."
- In relation **result_rel_theft**, we are generating the grouped column and the count of each.
- Finally, using dump, we are printing the result.
- Result prints the district & no of arrest **done in theft** for the district.

Pig Scripts

```
REGISTER /usr/local/pig/contrib/piggybank/java/piggybank.jar;

DEFINE CSVExcelStorage  org.apache.pig.piggybank.storage.CSVExcelStorage;

my_rel = Load '/home/acadgild/pig/Crimes_2001_to_present.csv' USING
CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER') AS
(Id:int,Case:chararray,Date:chararray,Block:chararray,IUCR:chararray,Primary_type:chararray,
Description:chararray,Loc_Desc:chararray, Arrest:chararray, Domestic:chararray, Beat:int,
District:int,Ward:int,Community_Area:int,FBI_Code:chararray, X_coord:int, Y_coord:int,
Year:int,Updated_on:chararray,Lati:double,Longi:double,Location:chararray);

filter_by_theft= FILTER my_rel BY Primary_type=='THEFT';

grouped_by_theft= GROUP filter_by_theft BY District;

result_rel_theft = FOREACH grouped_by_theft GENERATE  group,
COUNT(filter_by_theft.Id);

dump result_rel_theft;
```

Output:

```
2017-11-11 19:40:12,409 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
(1,5969)
(2,2696)
(3,2281)
(4,2955)
(5,2044)
(6,3275)
(7,2208)
(8,3977)
(9,2790)
(10,1902)
(11,2172)
(12,4023)
(14,3290)
(15,1481)
(16,2338)
(17,2269)
(18,5643)
(19,4724)
(20,1315)
(22,2080)
(24,1838)
(25,3058)
(31,1)
grunt>
```

Result:

(1,5969)
(2,2696)
(3,2281)
(4,2955)
(5,2044)
(6,3275)
(7,2208)
(8,3977)
(9,2790)
(10,1902)
(11,2172)
(12,4023)
(14,3290)
(15,1481)
(16,2338)
(17,2269)
(18,5643)
(19,4724)
(20,1315)
(22,2080)
(24,1838)
(25,3058)
(31,1)

Problem Statement 4

Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015

Steps to to calculate the number of arrests done between October 2014 and October 2015.

- **In Line 1**: We are registering *piggybank* jar in order to use the CSVExcelStorage class.
- **In Line 2:** we are defining org.apache.pig.piggybank.storage.CSVExcelStorage as CSVExcelStorage
- In relation **my_rel**, we are loading the dataset using CSVExcelStorage because of its effective technique to handle double quotes and headers.
- In relation **S_Date**, we are generating the columns Arrest, SUBSTRING(Date) for generating the month and year which are required for processing.
- In relation **S_Date_int** , we are explicitly type-casting month and year to integer
- In relation **Y_2014**, we are filtering **S_Date_int,** based on Arrest= 'true' and month >9 and Year=2014. This will remove all the records, which are not arrest and which are not between October and December 2014.
- In relation **Y_2015**, we are filtering **S_Date_int,** based on Arrest= 'true' and month <11 and Year=2015. This will remove all the records, which are not arrest and which are not between January and October 2015.
- In relation **union_res**, we are merging the content of two relations **Y_2014** and **Y_2015**.
- In relation **union_group**, we are grouping the data.
- In relation **result_rel**, we are generating the count of grouped column.
- Finally, using dump, we are printing the result.
- Result prints the number of arrests done between October 2014 and October 2015.

Pig Scripts

```
REGISTER /usr/local/pig/contrib/piggybank/java/piggybank.jar;

DEFINE CSVExcelStorage  org.apache.pig.piggybank.storage.CSVExcelStorage;

my_rel = Load '/home/acadgild/pig/Crimes_2001_to_present.csv' USING
CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER') AS
(Id:int,Case:chararray,Date:chararray,Block:chararray,IUCR:chararray,Primary_type:chararray,
Description:chararray,Loc_Desc:chararray, Arrest:chararray, Domestic:chararray, Beat:int,
District:int,Ward:int,Community_Area:int,FBI_Code:chararray, X_coord:int, Y_coord:int,
Year:int,Updated_on:chararray,Lati:double,Longi:double,Location:chararray);

S_Date = FOREACH my_rel Generate Arrest, SUBSTRING(Date,0,2) AS Month,
SUBSTRING(Date,6,10) AS Year;

S_Date_int = FOREACH S_Date Generate Arrest,(int)Month As Month,(int)Year as Year;
```

```
Y_2014 = Filter S_Date_int BY (Arrest=='true') AND (Month>9) AND (Year==2014);

Y_2015 = Filter S_Date_int BY (Arrest=='true') AND (Month<11) AND (Year==2015);

union_res= UNION Y_2014,Y_2015;

union_group = GROUP union_res all;

result_rel = FOREACH union_group GENERATE COUNT(union_res);

dump result_rel;
```

Output:

```
2017-11-11 19:59:42,289 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
(65028)
grunt>
```

Result:

(65028)