Session 5: EXPLORIN PIG

Assignment 5.2

Student Name: Subham Vishal

Course: Big Data Hadoop & Spark Training

Assignment 5.2–
Implement the use case present in below blog link and share the complete steps along with Screenshot from your end.

NOTE: You must submit a word file containing steps and screenshots.

Contents

Input Data Sets:

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

Problem Statement 1

Find out the top 5 most visited destinations

Codes:

1. **REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');**
3. **B = FOREACH A GENERATE (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;**
4. **C = FILTER B by dest is NOT Null;**
5. **D = GROUP C BY dest;**
6. **E = FOREACH D GENERATE group, COUNT(C.dest);**
7. **F = ORDER E by $1 DESC;**
8. **Result = LIMIT F 5;**
9. **A1 = LOAD '/home/acadgild/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');**
10. **A2 = FOREACH A1 GENERATE (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;**
11. **joined_table = JOIN Result by $0, A2 by dest;**
12. **DUMP joined_table;**

1.
```
grunt>
grunt>
grunt> REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';
```
2.
3.
```
grunt> A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
```
4.
5.
```
grunt> B = FOREACH A GENERATE (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
```
6.
7.
```
grunt> C = FILTER B by dest is NOT Null;
```
8.
9.
```
grunt> D = GROUP C BY dest;
```
10.
11.
```
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/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

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;
```

Output: the top 5 most visited destinations

Original Output:

```
(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)
```

Executed Output:

```
(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)
```

Problem Statement 2

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

Codes:

1. **REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');**
3. **B = FOREACH A GENERATE(int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;**
4. **C = FILTER B by cancelled == 1 AND cancel_code =='B';**
5. **D = group C BY month;**
6. **E = FOREACH D GENERATE group, COUNT(C.cancelled);**
7. **F= ORDER E BY $1 DESC;**
8. **Result = limit F 1;**
9. **DUMP Result;**

```
grunt>
grunt> REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';
```

```
grunt>
grunt>
grunt> A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
```

```
grunt> B = FOREACH A GENERATE(int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt>
grunt>
grunt> C = FILTER B by cancelled == 1 AND cancel_code =='B';
grunt> D = group C BY month;
grunt>
grunt> E = FOREACH D GENERATE group, COUNT(C.cancelled);
grunt>
grunt> F= ORDER E BY $1 DESC;
grunt>
grunt> Result = limit F 1;
grunt>
grunt> DUMP Result;
```

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

Original Output:

```
2016-11-13 11:58:45,906 [main] INFO
max
2016-11-13 11:58:45,907 [main] WARN
2016-11-13 11:58:45,922 [main] INFO
2016-11-13 11:58:45,922 [main] INFO
(12,250)
```

Executed Output:

```
(12,250)
grunt>
```

Problem Statement 3

Top ten origins with the highest AVG departure delay

Codes:

1. **REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');**
3. **B1 = FOREACH A GENERATE (int)$16 as dep_delay, (chararray)$17 as origin;**
4. **C1 = FILTER B1 BY (dep_delay is not null) AND (origin is not null);**
5. **D1 = group C1 by origin;**
6. **E1 = FOREACH D1 generate group, AVG(C1.dep_delay);**
7. **Result = ORDER E1 by $1 DESC;**
8. **Top_ten = LIMIT Result 10;**
9. **Lookup = load '/home/acadgild/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');**
10. **Lookup1 = FOREACH Lookup GENERATE (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;**
11. **Joined = JOIN Lookup1 by origin, Top_ten by $0;**
12. **Final = FOREACH Joined GENERATE $0,$1,$2,$4;**
13. **Final_Result = ORDER Final by $3 DESC;**
14. **DUMP Final_Result;**

```
grunt>
grunt> REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';
grunt> A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
```

```
grunt> B1 = FOREACH A GENERATE (int)$16 as dep_delay, (chararray)$17 as origin;
grunt>
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>
grunt> Lookup = load '/home/acadgild/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP
_INPUT_HEADER');
```

```
grunt>
grunt>
grunt> Lookup1 = FOREACH Lookup GENERATE (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
grunt>
grunt> Joined = JOIN Lookup1 by origin, Top_ten by $0;
grunt> Final = FOREACH Joined GENERATE $0,$1,$2,$4;
grunt>
grunt> Final_Result = ORDER Final by $3 DESC;
grunt> DUMP Final_Result;
```

Output Top ten origins with the highest AVG departure delay

Original Output

```
(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)
```

Executed Output:

```
(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)
grunt>
```

Problem Statement 4

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

Codes:

1. **REGISTER '/home/acadgild/airline_usecase/piggybank.jar';**
2. **A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');**
3. **B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;**
4. **C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);**
5. **D = GROUP C by (origin,dest);**
6. **E = FOREACH D generate group, COUNT(C.diversion);**
7. **F = ORDER E BY $1 DESC;**
8. **Result = LIMIT F 10;**

9. DUMP Result;

```
grunt>
grunt> REGISTER '/home/acadgild/hadoop/piggybank-0.15.0.jar';
grunt> A = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKI
P_INPUT_HEADER');
```

```
grunt> B1 = FOREACH A GENERATE (int)$16 as dep_delay, (chararray)$17 as origin;
grunt>
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;
```

Output which route (origin & destination) has seen the maximum diversion?

Original output:

```
(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)
```

Executed output:

```
((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)
```