AVIATION DATA ANALYSIS USING APACHE PIG

BIG DATA
HADOOP
&
SPARK
TRAINING

ACADGILD

ASSIGNMENT
5.2

BY :-

SAHIL
KHURANA

Associated Data Files

Delayed_Flights.csv
https://drive.google.com/file/d/0B_Qjau8wv1KoWTVDUVFOdzlJNWM/view

There are 29 columns in this dataset. Some of them have been mentioned below:
Year: 1987 – 2008
Month: 1 – 12
FlightNum: Flight number
Canceled: Was the flight canceled?
CancelleationCode: The reason for cancellation.

Airports.csv
https://drive.google.com/file/d/0B_Qjau8wv1KocDR3djk1Qm96Mmc/view

Data: the international airport abbreviation code
name of the airport
city and country in which airport is located.
lat and long: the latitude and longitude of the airport

Note: - To solve the Assignment, I have created a VM with Ubuntu 16.04 OS and configured Hadoop 2.8.2 and pig-0.17.0 on the same

Put the dataset in HDFS location

```
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hadoop/Pig/airline_usecase
Found 2 items
-rw-r--r--   1 sahil supergroup  247963212 2017-12-05 11:09 /u01/hadoop/Pig/airline_usecase/DelayedFlights.csv
-rw-r--r--   1 sahil supergroup     244438 2017-12-05 10:19 /u01/hadoop/Pig/airline_usecase/airports.csv
sahil@ubuntu:~/Desktop$
sahil@ubuntu:~/Desktop$
```

Problem Statement 1
Find out the top 5 most visited destinations.

Commands Used:-

```
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_IN
PUT_HEADER');
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 '/u01/hadoop/Pig/airline_usecase/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_IN
PUT_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;
```

```
sahil@ubuntu:/usr/local$ pig -x mapreduce
17/12/05 10:48:13 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
17/12/05 10:48:13 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17/12/05 10:48:13 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
17/12/05 10:48:13 WARN pig.Main: Cannot write to log file: /usr/local/pig_1512499693516.log
2017-12-05 10:48:13,519 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2017-12-05 10:48:13,989 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/sahil/.pigbootup not found
2017-12-05 10:48:15,667 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.j
obtracker.address
2017-12-05 10:48:15,667 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://
localhost:9000
2017-12-05 10:48:27,077 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-9421feae-b642-4cdc-98fb-5dc4809fad58
2017-12-05 10:48:27,077 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UN
IX','SKIP_INPUT_HEADER');
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 '/u01/hadoop/Pig/airline_usecase/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;
```

Result:-

```
sahil@ubuntu: ~/Desktop
2017-12-05 11:25:20,171 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2017-12-05 11:25:20,171 [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)
grunt>
```

(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?
Commands Used:-

```
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_IN
PUT_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> C = filter B by cancelled == 1 AND cancel_code =='B';
grunt> 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;

```
sahil@ubuntu:~/Desktop$ pig -x mapreduce
17/12/05 11:36:55 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
17/12/05 11:36:55 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17/12/05 11:36:55 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-12-05 11:36:55,817 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2017-12-05 11:36:55,817 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/sahil/Desktop/pig_1512502615468.log
2017-12-05 11:36:57,752 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/sahil/.pigbootup not found
2017-12-05 11:36:59,257 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.
jobtracker.address
2017-12-05 11:36:59,257 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs:/
/localhost:9000
2017-12-05 11:37:04,558 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-f241902f-1e8f-40b6-b75a-50719ce598ad
2017-12-05 11:37:04,558 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','U
NIX','SKIP_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> C = filter B by cancelled == 1 AND cancel_code =='B';
grunt> 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;
```

Results:-

```
sahil@ubuntu: ~/Desktop
2017-12-05 11:43:04,744 [main]
(12,250)
grunt>
grunt>
grunt>
```

(12,250)

Problem Statement 3

Top ten origins with the highest AVG departure delay

Commands Used:-

grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
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 '/u01/hadoop/Pig/airline_usecase/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
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;

```
sahil@ubuntu: ~/Desktop
sahil@ubuntu:~/Desktop$ pig -x mapreduce
17/12/05 12:11:58 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
17/12/05 12:11:58 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17/12/05 12:11:58 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-12-05 12:11:58,174 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2017-12-05 12:11:58,174 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/sahil/Desktop/pig_1512504718173.log
2017-12-05 12:11:58,196 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/sahil/.pigbootup not found
2017-12-05 12:11:58,878 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.
jobtracker.address
2017-12-05 12:11:58,878 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs:/
/localhost:9000
2017-12-05 12:11:59,447 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-a23b008a-2c00-4229-8c22-6280ea866cb5
2017-12-05 12:11:59,447 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','U
NIX','SKIP_INPUT_HEADER');
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 '/u01/hadoop/Pig/airline_usecase/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UN
IX','SKIP_INPUT_HEADER');
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;
```

Results:-

```
sahil@ubuntu: ~/Desktop
2017-12-05 12:17:14,800 [main] INFO  org.apac
2017-12-05 12:17:14,800 [main] INFO  org.apac
(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>
```

(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)

Problem Statement 4

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

Commands Used:-

```
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_IN
PUT_HEADER');
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;
```

```
sahil@ubuntu:~/Desktop$ pig -x mapreduce
17/12/05 12:10:31 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
17/12/05 12:10:31 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17/12/05 12:10:31 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-12-05 12:10:31,450 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2017-12-05 12:10:31,450 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/sahil/Desktop/pig_1512504631449.log
2017-12-05 12:10:31,489 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/sahil/.pigbootup not found
2017-12-05 12:10:32,194 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.
jobtracker.address
2017-12-05 12:10:32,194 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs:/
/localhost:9000
2017-12-05 12:10:32,761 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-fd05af66-ab63-4e03-9197-64d9c9831aca
2017-12-05 12:10:32,761 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load '/u01/hadoop/Pig/airline_usecase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','U
NIX','SKIP_INPUT_HEADER');
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;
```

Result:-

```
sahil@ubuntu: ~/Desktop
2017-12-05 12:07:08,285 [main]
((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>
grunt>
```

((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)