Assignment 2.5

Problem Statement 1

Find out the top 5 most visited destinations.

Solution:

```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar'

A = load '/home/acadgild/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as dest;

C = filter B by dest is not null;

D = group C by dest;

E = foreach D generate group, COUNT(C.dest);

F = order E by $1 DESC;

Result = LIMIT F 5;

A1 = load '/home/acadgild/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;

joined_table = join Result by $0, A2 by dest;

dump joined_table;
```

```
[acadgild@localhost java]$ pig -x LOCAL
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.
7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/imp
l/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/05/19 19:51:55 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/05/19 19:51:55 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-05-19 19:51:55,412 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016,
 23:10:49
2018-05-19 19:51:55,413 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/install/pig/pig-
0.16.0/contrib/piggybank/java/pig_1526739715404.log
2018-05-19 19:51:55,453 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup n
ot found
2018-05-19 19:51:55,827 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depreca
ted. Instead, use mapreduce.jobtracker.address
2018-05-19 19:51:55,828 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
2018-05-19 19:51:55,833 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to h
adoop file system at: file:///
2018-05-19 19:51:55,970 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 19:51:56,002 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-07cb0d82-3
4b1-4fc0-a70f-fae8201499f4
2018-05-19 19:51:56,002 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled s
et to false
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar'
2018-05-19 19:52:38,236 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 19:52:38,236 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MU
LTILINE','UNIX','SKIP_INPUT_HEADER');
2018-05-19 19:56:43,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 19:56:43,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
grunt> B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18 as de
st;
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>
```

```
grunt> A1 = load '/home/acadgild/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILI
NE','UNIX','SKIP_INPUT_HEADER');
2018-05-19 20:02:58,131 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 20:02:58,131 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
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;
```

```
(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>
```

Problem Statement 2

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

Solution:

REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar';

A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;

C = filter B by cancelled == 1 AND cancel_code =='B';

D = group C by month;

E = foreach D generate group, COUNT(C.cancelled);

F= order E by $1 DESC;

Result = limit F 1;

dump Result;

```
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar';
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MU
LTILINE','UNIX','SKIP_INPUT_HEADER');
2018-05-19 20:08:12,118 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 20:08:12,118 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
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;
```

```
(12,250)
```

Problem Statement 3

Top ten origins with the highest AVG departure delay

Solution:

```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar';

A = load '/home/acadgild/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

B1 = foreach A generate (int)$16 as dep_delay, (chararray)$17 as origin;

C1 = filter B1 by (dep_delay is not null) AND (origin is not null);

D1 = group C1 by origin;

E1 = foreach D1 generate group, AVG(C1.dep_delay);

Result = order E1 by $1 DESC;

Top_ten = limit Result 10;

Lookup = load '/home/acadgild/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

Lookup1 = foreach Lookup generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;

Joined = join Lookup1 by origin, Top_ten by $0;

Final = foreach Joined generate $0,$1,$2,$4;

Final_Result = ORDER Final by $3 DESC;

dump Final_Result;
```

```
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar';
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MU
LTILINE','UNIX','SKIP_INPUT_HEADER');
2018-05-19 20:14:34,723 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 20:14:34,723 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
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 '/home/acadgild/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MUL
TILINE','UNIX','SKIP_INPUT_HEADER');
2018-05-19 20:15:21,031 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 20:15:21,031 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
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;
```

```
(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?

Solution:

```
REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar';

A = load '/home/acadgild/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');

B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;

C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);

D = GROUP C by (origin,dest);

E = FOREACH D generate group, COUNT(C.diversion);

F = ORDER E BY $1 DESC;

Result = limit F 10;

dump Result;
```

```
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/contrib/piggybank/java/piggybank.jar';
grunt> A = load '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MU
LTILINE','UNIX','SKIP_INPUT_HEADER');
2018-05-19 20:21:04,862 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecated. Instead, use dfs.bytes-per-checksum
2018-05-19 20:21:04,862 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
. Instead, use fs.defaultFS
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;
```

```
((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)
```