Module 4: Pig Assignment – 2

1. Find out the top five most visited destinations

Solution:

```
grunt> flight_data = load '/home/bitnami/pigdata/flights_details.csv' USING PigStorage(',');
2021-11-05 19:19:41,991 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> gen_flight_data = foreach flight_data generate (int)$1 as year, (int)$10 as
>> flight_num, (chararray)$17 as origin,(chararray) $18 as dest;
grunt> fiter_flight_not_null = filter gen_flight_data by dest is not null;
grunt> grp_dest = group fiter_flight_not_null by dest;
grunt> gen_count_dest = foreach grp_dest generate group, COUNT(fiter_flight_not_null.dest);
grunt> ord_count_desc = order gen_count_dest by $1 DESC;
grunt> lmt_dest_cnt = LIMIT ord_count_desc 5;
grunt> airport_data = load '/home/bitnami/pigdata/airports.csv' USING PigStorage(',');
2021-11-05 19:20:12,842 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> gen_airport_data = foreach airport_data generate (chararray)$0 as dest, (chararray)$2 as city, (c
hararray)$4 as country;
grunt> joined_table = join lmt_dest_cnt by $0, gen_airport_data by dest;
grunt> dump joined_table;
```

Job Details:

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime  M
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1085589048_0001        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      fiter_flight_not_null,flight_data,gen_count_dest,gen_flight_data,grp_dest       GROUP_BY,COMBINE
R
job_local1670033492_0005        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      airport_data,gen_airport_data,joined_table      HASH_JOIN       file:/tmp/temp620259091/tmp14209
12278,
job_local1714413284_0003        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ord_count_desc  ORDER_BY,COMBINER
job_local1918540700_0004        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ord_count_desc
job_local1991973631_0002        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ord_count_desc  SAMPLER

Input(s):
Successfully read 1338 records from: "/home/bitnami/pigdata/flights_details.csv"
Successfully read 1477 records from: "/home/bitnami/pigdata/airports.csv"

Output(s):
Successfully stored 39 records in: "file:/tmp/temp620259091/tmp1420912278"

Counters:
Total records written : 39
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1085589048_0001        ->      job_local1991973631_0002,
job_local1991973631_0002        ->      job_local1714413284_0003,
job_local1714413284_0003        ->      job_local1918540700_0004,
job_local1918540700_0004        ->      job_local1670033492_0005,
job_local1670033492_0005
```

Output:

```
(BWI,65,BWI,Baltimore,USA)
(BWI,65,BWI,Fair Haven,USA)
(BWI,65,BWI,Hudson,USA)
(LAS,93,LAS,Dyersburg,USA)
(LAS,93,LAS,Meeker,USA)
(LAS,93,LAS,Keene,USA)
(LAS,93,LAS,Eek,USA)
(LAS,93,LAS,Needles,USA)
(LAS,93,LAS,Enterprise,USA)
(LAS,93,LAS,Jal,USA)
(LAS,93,LAS,Wickenburg,USA)
(LAS,93,LAS,Whiteriver,USA)
(LAS,93,LAS,Gruver,USA)
(LAS,93,LAS,Graham,USA)
(LAS,93,LAS,Andrews,USA)
(LAS,93,LAS,Tatum,USA)
(LAS,93,LAS,Lovington,USA)
(LAS,93,LAS,Hatch,USA)
(LAS,93,LAS,Eunice,USA)
(LAS,93,LAS,Monahans,USA)
(LAS,93,LAS,Duluth,USA)
(LAS,93,LAS,Douglas Bisbee,USA)
(MDW,79,MDW,Tatitlek,USA)
(MDW,79,MDW,McGehee,USA)
(MDW,79,MDW,Mount Ida,USA)
(MDW,79,MDW,Skagway,USA)
(MDW,79,MDW,South Sioux City,USA)
(MDW,79,MDW,Hartford,USA)
(MDW,79,MDW,Osceola,USA)
(OAK,62,OAK,Hilliard,USA)
(OAK,62,OAK,Gatesville,USA)
(OAK,62,OAK,Griffith,USA)
(OAK,62,OAK,Clanton,USA)
(OAK,62,OAK,Belmont,USA)
(PHX,78,PHX,Bonifay,USA)
(PHX,78,PHX,Delphi,USA)
(PHX,78,PHX,Goldsby,USA)
(PHX,78,PHX,Lindsay,USA)
(PHX,78,PHX,Freehold,USA)
```

2. Which month has seen the greatest number of cancellations due to bad weather?

Solution:

```
grunt> flight_data = load '/home/bitnami/pigdata/flights_details.csv' USING PigStorage(',');
2021-11-05 19:23:13,745 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> gen_flight_data = foreach flight_data generate (int)$2 as month,(int)$10 as
>> flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt> fltr_data = filter gen_flight_data by cancelled == 1 AND cancel_code =='B';
grunt> grp_mnth = group fltr_data by month;
grunt> gen_grp = foreach grp_mnth generate group, COUNT(fltr_data.cancelled);
grunt> ord_yr= order gen_grp by $1 DESC;
grunt> Result = limit ord_yr 1;
grunt> dump Result;
```

Job Details:

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.1   0.17.0  root    2021-11-05 19:24:20     2021-11-05 19:24:22     GROUP_BY,ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   M
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1131785180_0009        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ord_yr          file:/tmp/temp620259091/tmp993652038,
job_local1193560756_0008        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ord_yr  ORDER_BY,COMBINER
job_local1334007339_0007        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ord_yr  SAMPLER
job_local961774135_0006 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     f
light_data,fltr_data,gen_flight_data,gen_grp,grp_mnth   GROUP_BY,COMBINER

Input(s):
Successfully read 1338 records from: "/home/bitnami/pigdata/flights_details.csv"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp620259091/tmp993652038"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local961774135_0006 ->      job_local1334007339_0007,
job_local1334007339_0007        ->      job_local1193560756_0008,
job_local1193560756_0008        ->      job_local1131785180_0009,
job_local1131785180_0009
```

Output:

```
2021-11-05 19:24:23,003 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - To
al input paths to process : 1
(1,60)
```

3. Find out the top ten origins with the highest AVG departure delay

Solution:

```
grunt> A = load '/home/bitnami/pigdata/flights_details.csv' USING PigStorage(',');
2021-11-05 19:25:48,709 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
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/bitnami/pigdata/airports.csv' USING PigStorage(',');
2021-11-05 19:26:13,027 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
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;
```

Job Details:

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime  M
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1778115246_0011        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      Result  SAMPLER
job_local216947728_0010 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     A
,B1,C1,D1,E1    GROUP_BY,COMBINER
job_local41679641_0012  1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     R
esult   ORDER_BY,COMBINER
job_local440058059_0014 2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F
inal,Joined,Lookup,Lookup1      HASH_JOIN
job_local504372126_0013 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     R
esult
job_local597010120_0015 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F
inal_Result     SAMPLER
job_local98024150_0016  1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F
inal_Result     ORDER_BY        file:/tmp/temp620259091/tmp-1877256182,

Input(s):
Successfully read 1338 records from: "/home/bitnami/pigdata/flights_details.csv"
Successfully read 1477 records from: "/home/bitnami/pigdata/airports.csv"

Output(s):
Successfully stored 31 records in: "file:/tmp/temp620259091/tmp-1877256182"

Counters:
Total records written : 31
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local216947728_0010 ->      job_local1778115246_0011,
job_local1778115246_0011        ->      job_local41679641_0012,
job_local41679641_0012  ->      job_local504372126_0013,
job_local504372126_0013 ->      job_local440058059_0014,
job_local440058059_0014 ->      job_local597010120_0015,
job_local597010120_0015 ->      job_local98024150_0016,
job_local98024150_0016
```

Output:

```
(MDW,South Sioux City,USA,48.950980392156865)
(MDW,Tatitlek,USA,48.950980392156865)
(MDW,McGehee,USA,48.950980392156865)
(MDW,Mount Ida,USA,48.950980392156865)
(MDW,Skagway,USA,48.950980392156865)
(MDW,Osceola,USA,48.950980392156865)
(MDW,Hartford,USA,48.950980392156865)
(LAS,Dyersburg,USA,41.830601092896174)
(LAS,Douglas Bisbee,USA,41.830601092896174)
(LAS,Duluth,USA,41.830601092896174)
(LAS,Monahans,USA,41.830601092896174)
(LAS,Eunice,USA,41.830601092896174)
(LAS,Hatch,USA,41.830601092896174)
(LAS,Lovington,USA,41.830601092896174)
(LAS,Tatum,USA,41.830601092896174)
(LAS,Andrews,USA,41.830601092896174)
(LAS,Graham,USA,41.830601092896174)
(LAS,Gruver,USA,41.830601092896174)
(LAS,Whiteriver,USA,41.830601092896174)
(LAS,Wickenburg,USA,41.830601092896174)
(LAS,Jal,USA,41.830601092896174)
(LAS,Enterprise,USA,41.830601092896174)
(LAS,Needles,USA,41.830601092896174)
(LAS,Eek,USA,41.830601092896174)
(LAS,Keene,USA,41.830601092896174)
(LAS,Meeker,USA,41.830601092896174)
(OAK,Clanton,USA,41.294117647058826)
(OAK,Belmont,USA,41.294117647058826)
(OAK,Hilliard,USA,41.294117647058826)
(OAK,Gatesville,USA,41.294117647058826)
(OAK,Griffith,USA,41.294117647058826)
```

4. Which route (origin and destination) has seen the maximum diversion?

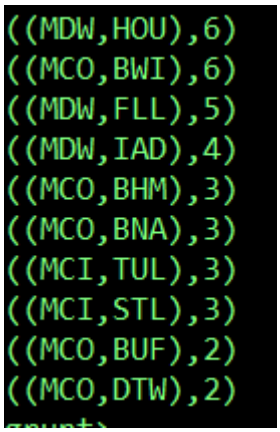
Solution:

```
grunt> A = load '/home/bitnami/pigdata/flights_details.csv' USING PigStorage(',');
2021-11-05 19:28:11,986 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
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;_
```

Job Details:

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.1   0.17.0  root    2021-11-05 19:28:58     2021-11-05 19:29:00     GROUP_BY,ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   M
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1764430980_0019        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n
/a      F       ORDER_BY,COMBINER
job_local507804641_0018 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F
SAMPLER
job_local684917989_0020 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     F
file:/tmp/temp620259091/tmp260684207,
job_local909901284_0017 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     A
,B,C,D,E        GROUP_BY,COMBINER

Input(s):
Successfully read 1338 records from: "/home/bitnami/pigdata/flights_details.csv"

Output(s):
Successfully stored 10 records in: "file:/tmp/temp620259091/tmp260684207"

Counters:
Total records written : 10
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local909901284_0017 ->      job_local507804641_0018,
job_local507804641_0018 ->      job_local1764430980_0019,
job_local1764430980_0019        ->      job_local684917989_0020,
job_local684917989_0020
```

Output:


((MDW,HOU),6)
((MCO,BWI),6)
((MDW,FLL),5)
((MDW,IAD),4)
((MCO,BHM),3)
((MCO,BNA),3)
((MCI,TUL),3)
((MCI,STL),3)
((MCO,BUF),2)
((MCO,DTW),2)