

Problem Statement 1:

Find out the top 5 most visited destinations.

Solution:

```
-- Register piggybank.jar to use CSVExcelStorage
REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar';
--Load flight details
flight_data = load '/home/acadgild/ankita/Assignment5_2/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');
-- Get the destinations
flight_dest_details = foreach flight_data generate (chararray) $18 as dest;

-- Filter the destinations - get only not null destinations
dest_data = filter flight_dest_details by dest is not null;
-- group result by destination
dest_group = group dest_data by dest;

-- count destinations for every group
dest_count = foreach dest_group generate group, COUNT(dest_data.dest);

-- sort results by count of destinations in descending order
top_dest = order dest_count by $1 DESC;

-- fetch top 5 rows from the result
top5_dest = LIMIT top_dest 5;
-- load Airport details
airport_data = load '/home/acadgild/ankita/Assignment5_2/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER');
-- get destination, city and country
airport_country_city = foreach airport_data generate (chararray)$0 as dest, (chararray)$2 as city,
(chararray)$4 as country;
-- join flight data and airport data to get destination's country and city
dest_country_city = join top5_dest by $0, airport_country_city by dest;
-- sort final result by count of destinations
final_result = order dest_country_city by $1 DESC ;

-- display the results
dump final_result;
```

Output:

```
acadgild@localhost:~/ankita/Assignment5_2
File Edit View Search Terminal Help
2017-10-27 11:44:17,741 [main] INFO org.apache.hadoop.metrics.jvm.JvmM
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= -
initialized
2017-10-27 11:44:17,766 [main] INFO org.apache.pig.backend.hadoop.exec
ne.mapReduceLayer.MapReduceLauncher - Success!
2017-10-27 11:44:17,807 [main] INFO org.apache.hadoop.conf.Configurati
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
2017-10-27 11:44:17,815 [main] INFO org.apache.hadoop.conf.Configurati
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-10-27 11:44:17,815 [main] INFO org.apache.hadoop.conf.Configurati
ation - mapreduce.job.counters.limit is deprecated. Instead, use mapred
ounters.max
2017-10-27 11:44:17,815 [main] WARN org.apache.pig.data.SchemaTupleBac
hemaTupleBackend has already been initialized
2017-10-27 11:44:17,903 [main] INFO org.apache.hadoop.mapreduce.lib.in
putFormat - Total input paths to process : 1
2017-10-27 11:44:17,903 [main] INFO org.apache.pig.backend.hadoop.exec
ne.util.MapRedUtil - Total input paths to process : 1
(ORD,108984,ORD,Chicago,USA)
(ATL,106898,ATL,Atlanta,USA)
(DFW,70657,DFW,Dallas-Fort Worth,USA)
(DEN,63003,DEN,Denver,USA)
(LAX,59969,LAX,Los Angeles,USA)
grunt>
```

Problem Statement 2 :

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

Solution:

```
-- Register piggybank.jar to use CSVExcelStorage
REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar';
--Load flight details
flight_data = load '/home/acadgild/ankita/Assignment5_2/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADE
R');
-- Get the month, cancelled and cancellation code
flight_details = foreach flight_data generate (chararray) $2 as month, (int) $22 as
cancelled, (chararray) $23 as reason;

-- Filter the flight details - get only cancelled flights due to bad weather
cancelled_flights_badweather = filter flight_details by cancelled==1 AND reason=='B';
-- group results by month
group_cancelled_flights_by_month = group cancelled_flights_badweather by month;

-- count bad weather cancellations by month
```

```

count_cancelled_flights_by_month = foreach group_cancelled_flights_by_month generate group,
COUNT(cancelled_flights_badweather.reason);

-- sort results by count of cancellations in descending order
sort_cancellation_by_month = order count_cancelled_flights_by_month by $1 DESC;

-- fetch top row from the result
top_cancellation_month = LIMIT sort_cancellation_by_month 1;
dump top_cancellation_month ;

```

Output:

```

acadgild@localhost:~/ankita/Assignment5_2
File Edit View Search Terminal Help
initialized
2017-10-27 12:27:43,296 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-10-27 12:27:43,302 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-10-27 12:27:43,326 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2017-10-27 12:27:43,338 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-10-27 12:27:43,342 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-10-27 12:27:43,343 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.c
ounters.max
2017-10-27 12:27:43,343 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2017-10-27 12:27:43,421 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-10-27 12:27:43,421 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt>

```

Problem Statement 3:

Top ten origins with the highest AVG departure delay

Solution:

```

-- Register piggybank.jar to use CSVExcelStorage
REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar';
--Load flight details
flight_data = load '/home/acadgild/ankita/Assignment5_2/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER
');
-- Get the month, cancelled and cancellation code

```

```

flight_details = foreach flight_data generate (int) $16 as delay, (chararray) $17 as origin;
-- Filter the flight details - delay time should not be null and origin should not be null
flight_details_filter = FILTER flight_details by origin is not null and delay is not null ;

-- group result by origin
group_by_origin = group flight_details_filter by origin;

-- get the average of delays by origin
sum_delay = foreach group_by_origin generate group, AVG(flight_details_filter.delay);

-- sort results by descending order of delay
sort_flight_delays = order sum_delay by $1 DESC;

-- fetch top 10 rows from the result
top10_flight_delays = LIMIT sort_flight_delays 10;
-- load Airport details
airport_data = load '/home/acadgild/ankita/Assignment5_2/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER
');
-- get destination, city and country
airport_country_city = foreach airport_data generate (chararray)$0 as origin, (chararray)$2
as city, (chararray)$4 as country;
-- join flight data and airport data to get destination's country and city
dest_country_city = join top10_flight_delays by $0, airport_country_city by origin;
-- sort final result by count of destinations
final_result = order dest_country_city by $1 DESC ;

-- display the results
dump final_result;

```

Output:

```

acadgild@localhost:~/ankita/Assignment5_2
2017-10-27 12:37:33,404 [main] INFO org.apache.hadoop.conf.Configuration - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-
2017-10-27 12:37:33,405 [main] INFO org.apache.hadoop.conf.Configuration - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-10-27 12:37:33,411 [main] INFO org.apache.hadoop.conf.Configuration - mapreduce.job.counters.limit is deprecated. Instead, use mapredc
ounters.max
2017-10-27 12:37:33,411 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-27 12:37:33,498 [main] INFO org.apache.hadoop.mapreduce.lib.input.Format - Total input paths to process : 1
2017-10-27 12:37:33,499 [main] INFO org.apache.pig.backend.hadoop.execution.util.MapRedUtil - Total input paths to process : 1
(CMX,116.1470588235294,CMX,Hancock,USA)
(PLN,93.76190476190476,PLN,Pellston,USA)
(SPI,83.84873949579831,SPI,Springfield,USA)
(ALO,82.2258064516129,ALO,Waterloo,USA)
(MQT,79.55665024630542,MQT,NA,USA)
(ACY,79.3103448275862,ACY,Atlantic City,USA)
(MOT,78.66165413533835,MOT,Minot,USA)
(HHH,76.53005464480874,HHH,NA,USA)
(EGE,74.12891986062718,EGE,Eagle,USA)
(BGM,73.15533980582525,BGM,Binghamton,USA)
grunt>

```

Problem Statement 4:

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

Solution:

```

-- Register piggybank.jar to use CSVExcelStorage
REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar';
--Load flight details
flight_data = load '/home/acadgild/ankita/Assignment5_2/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER
');
-- Get the origin, destination and diverted details
flight_details = foreach flight_data generate (chararray) $17 as origin,(chararray)$18 as
dest, (int)$24 as diverted;
-- Filter the flight details - origin and destination should not be null and it should be
diverted.
flight_details_filter = FILTER flight_details by origin is not null and dest is not null and
diverted==1 ;

-- group result by origin and destination
group_by_origindest = group flight_details_filter by (origin,dest);

-- get the number of diversions

```

```

count_diversions = foreach group_by_origindest generate group,
COUNT(flight_details_filter.diverted);

-- sort results by descending order of diversions
sort_flight_diversions = order count_diversions by $1 DESC;

-- fetch top 10 rows from the result
top10_flight_diversions = LIMIT sort_flight_diversions 10;
dump top10_flight_diversions ;

```

Output:

```

acadgild@localhost:~/ankita/Assignment5_2
2017-10-27 12:44:42,047 [main] INFO org.apache.hadoop.conf.Configuration
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-c
2017-10-27 12:44:42,052 [main] INFO org.apache.hadoop.conf.Configuration
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-10-27 12:44:42,052 [main] INFO org.apache.hadoop.conf.Configuration
ation - mapreduce.job.counters.limit is deprecated. Instead, use mapredue
ounters.max
2017-10-27 12:44:42,052 [main] WARN org.apache.pig.data.SchemaTupleBacke
hemaTupleBackend has already been initialized
2017-10-27 12:44:42,111 [main] INFO org.apache.hadoop.mapreduce.lib.inpu
nputFormat - Total input paths to process : 1
2017-10-27 12:44:42,111 [main] INFO org.apache.pig.backend.hadoop.execu
ne.util.MapRedUtil - Total input paths to process : 1
((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> █

```