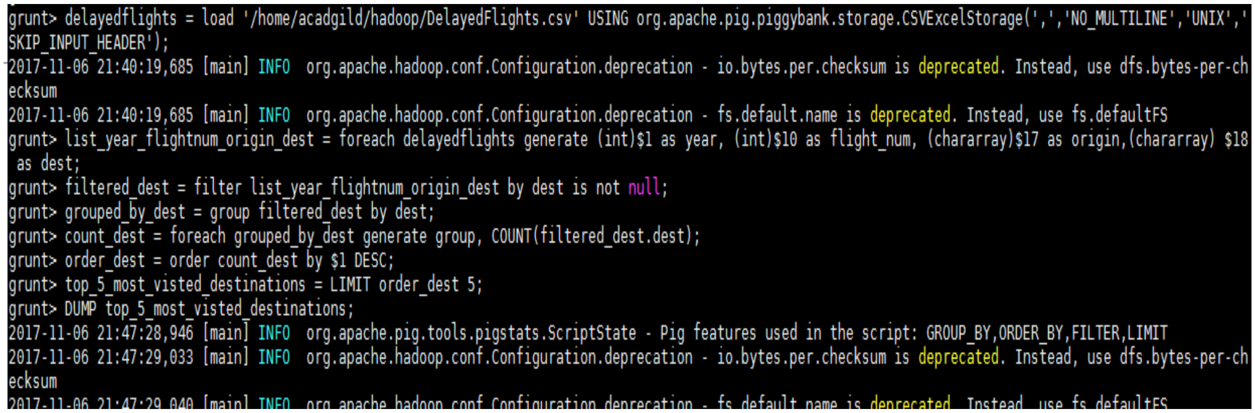
Assignment 5.2

Problem Statement 1:-

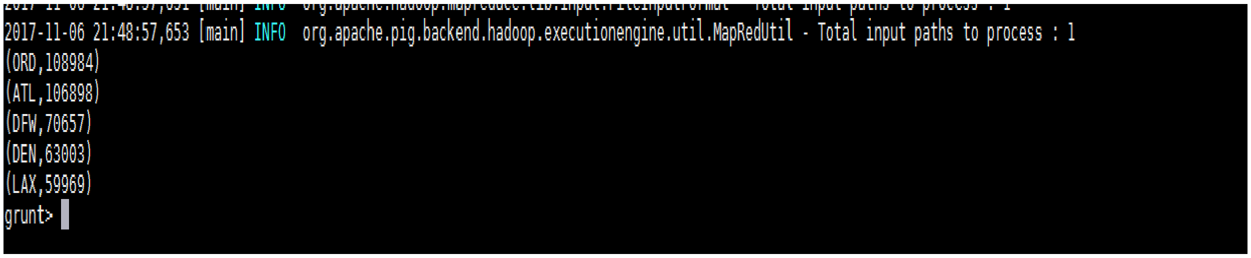
Find out the top 5 most visited destinations

Solution:-

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

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

filtered_dest = filter list_year_flightnum_origin_dest by dest is not null;

grouped_by_dest = group filtered_dest by dest;

count_dest = foreach grouped_by_dest generate group, COUNT(filtered_dest.dest);

order_dest = order count_dest by $1 DESC;

top_5_most_visted_destinations = LIMIT order_dest 5;

DUMP top_5_most_visted_destinations;
```



```
grunt> delayedflights = load '/home/acadgild/hadoop/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','
SKIP_INPUT_HEADER');
2017-11-06 21:40:19,685 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-ch
ecksum
2017-11-06 21:40:19,685 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> list_year_flightnum_origin_dest = foreach delayedflights generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray) $18
 as dest;
grunt> filtered_dest = filter list_year_flightnum_origin_dest by dest is not null;
grunt> grouped_by_dest = group filtered_dest by dest;
grunt> count_dest = foreach grouped_by_dest generate group, COUNT(filtered_dest.dest);
grunt> order_dest = order count_dest by $1 DESC;
grunt> top_5_most_visted_destinations = LIMIT order_dest 5;
grunt> DUMP top_5_most_visted_destinations;
2017-11-06 21:47:28,946 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER,LIMIT
2017-11-06 21:47:29,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-ch
ecksum
2017-11-06 21:47:29,040 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```



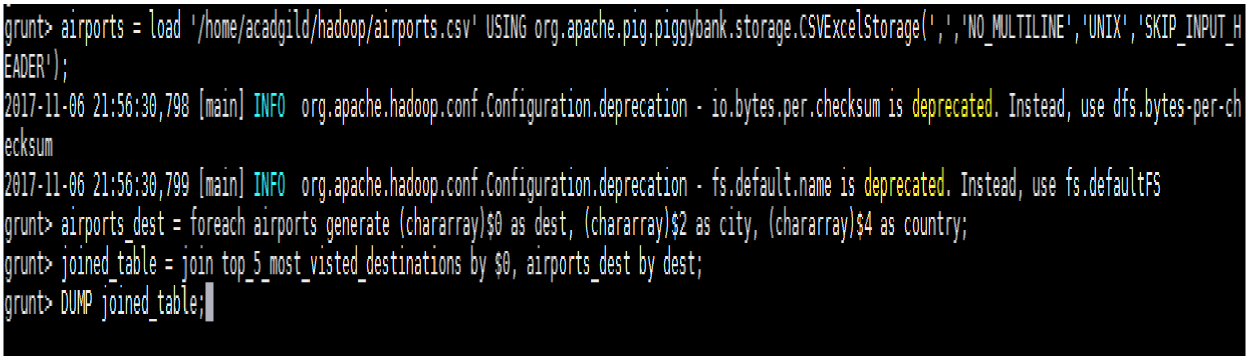
```
2017-11-06 21:48:57,651 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-06 21:48:57,653 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(ORD,108984)
(ATL,106898)
(DFW,70657)
(DEN,63003)
(LAX,59969)
grunt>
```

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

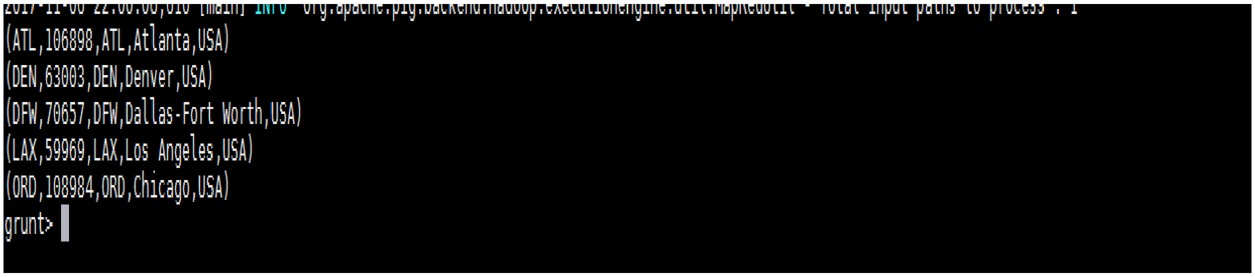
airports_dest = foreach airports generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;

joined_table = join top_5_most_visted_destinations by $0, airports_dest by dest;

DUMP joined_table;



```
grunt> airports = load '/home/acadgild/hadoop/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_H
EADER');
2017-11-06 21:56:30,798 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-ch
ecksum
2017-11-06 21:56:30,799 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> airports_dest = foreach airports generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
grunt> joined_table = join top_5_most_visted_destinations by $0, airports_dest by dest;
grunt> DUMP joined_table;
```



```
2017-11-06 22:00:00,010 [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>
```

Problem Statement 2:-

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

Solution:-

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

filtered_cancel = filter list_flights_cancel by cancelled == 1 AND cancel_code =='B';

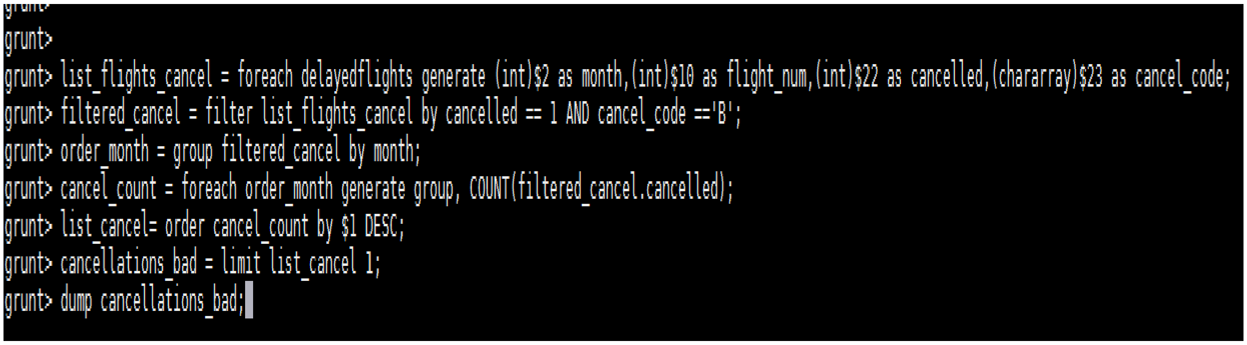
order_month = group filtered_cancel by month;

cancel_count = foreach order_month generate group, COUNT(filtered_cancel.cancelled);

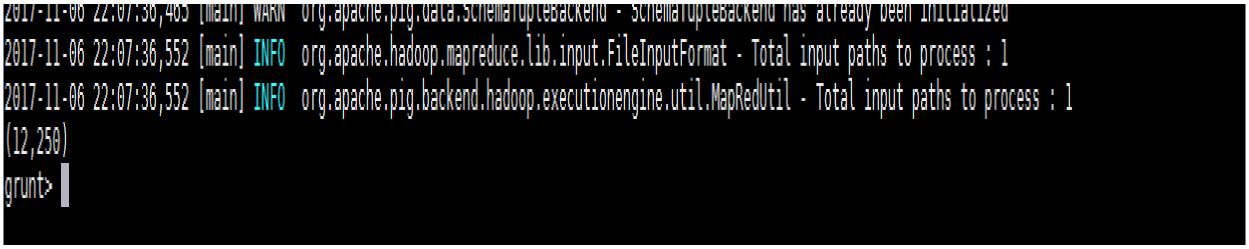
list_cancel= order cancel_count by $1 DESC;

cancellations_bad = limit list_cancel 1;

dump cancellations_bad;



```
grunt>
grunt>
grunt> list_flights_cancel = foreach delayedflights generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as cancelled,(chararray)$23 as cancel_code;
grunt> filtered_cancel = filter list_flights_cancel by cancelled == 1 AND cancel_code =='B';
grunt> order_month = group filtered_cancel by month;
grunt> cancel_count = foreach order_month generate group, COUNT(filtered_cancel.cancelled);
grunt> list_cancel= order cancel_count by $1 DESC;
grunt> cancellations_bad = limit list_cancel 1;
grunt> dump cancellations_bad;
```



```
2017-11-06 22:07:36,485 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-06 22:07:36,552 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-06 22:07:36,552 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
grunt>
```

Problem Statement 3:-

Top ten origins with the highest AVG departure delay

Solution:-

list_dep_delay = foreach delayedflights generate (int)$16 as dep_delay, (chararray)$17 as origin;

filtered_dep_delay = filter list_dep_delay by (dep_delay is not null) AND (origin is not null);

group_by_origin = group filtered_dep_delay by origin;

average_dep_delay = foreach group_by_origin generate group, AVG(filtered_dep_delay.dep_delay);

origins_highest_avg_dep_delay = order average_dep_delay by $1 DESC;

Top_ten_origins = limit origins_highest_avg_dep_delay 10;

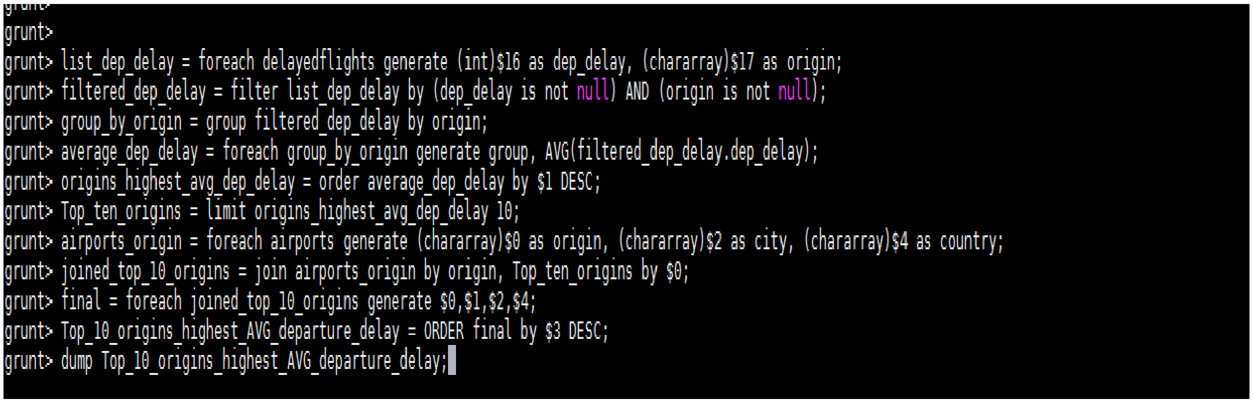
airports_origin = foreach airports generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;

joined_top_10_origins = join airports_origin by origin, Top_ten_origins by $0;

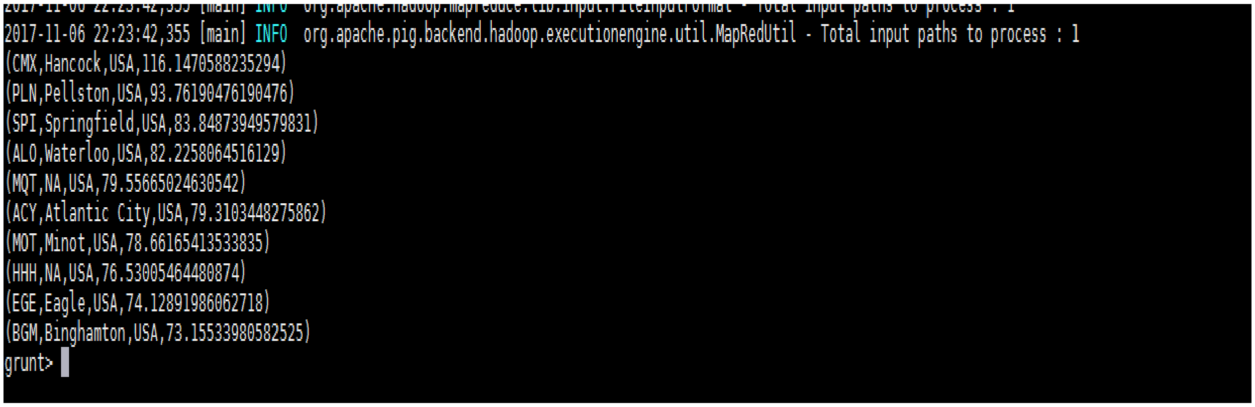
final = foreach joined_top_10_origins generate $0,$1,$2,$4;

Top_10_origins_highest_AVG_departure_delay = ORDER final by $3 DESC;

dump Top_10_origins_highest_AVG_departure_delay;



```
grunt>
grunt>
grunt> list_dep_delay = foreach delayedflights generate (int)$16 as dep_delay, (chararray)$17 as origin;
grunt> filtered_dep_delay = filter list_dep_delay by (dep_delay is not null) AND (origin is not null);
grunt> group_by_origin = group filtered_dep_delay by origin;
grunt> average_dep_delay = foreach group_by_origin generate group, AVG(filtered_dep_delay.dep_delay);
grunt> origins_highest_avg_dep_delay = order average_dep_delay by $1 DESC;
grunt> Top_ten_origins = limit origins_highest_avg_dep_delay 10;
grunt> airports_origin = foreach airports generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country;
grunt> joined_top_10_origins = join airports_origin by origin, Top_ten_origins by $0;
grunt> final = foreach joined_top_10_origins generate $0,$1,$2,$4;
grunt> Top_10_origins_highest_AVG_departure_delay = ORDER final by $3 DESC;
grunt> dump Top_10_origins_highest_AVG_departure_delay;
```



```
2017-11-06 22:23:42,355 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-06 22:23:42,355 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(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

Solution:-

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

filtered_diversion = FILTER list_diversion BY (origin is not null) AND (dest is not null) AND (diversion == 1);

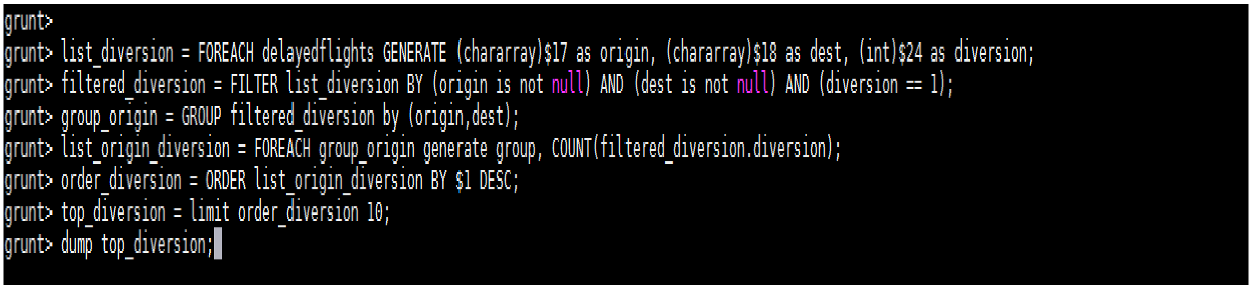
group_origin = GROUP filtered_diversion by (origin,dest);

list_origin_diversion = FOREACH group_origin generate group, COUNT(filtered_diversion.diversion);

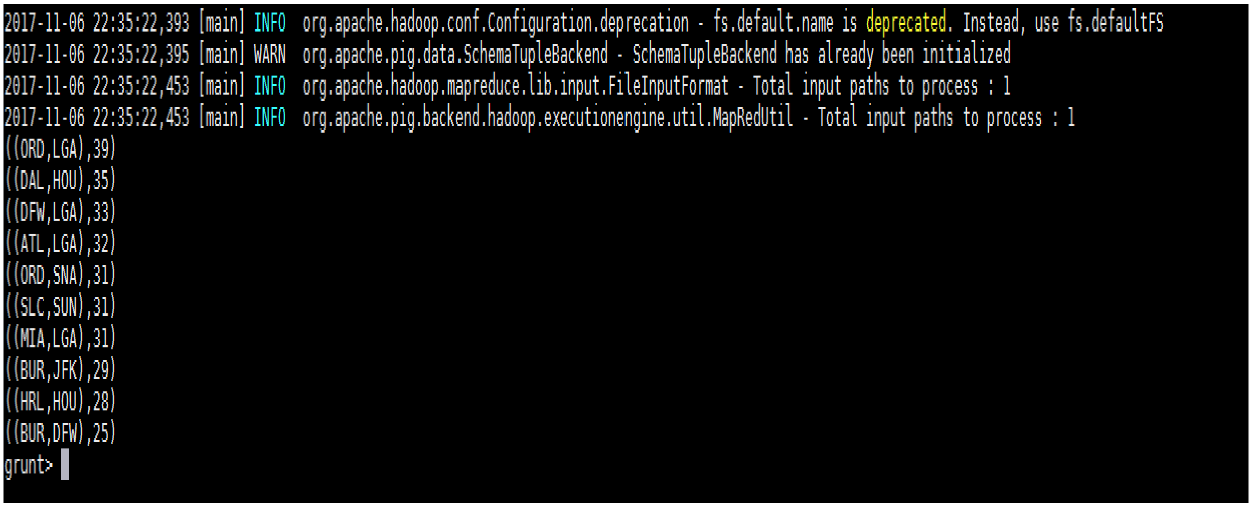
order_diversion = ORDER list_origin_diversion BY $1 DESC;

top_diversion = limit order_diversion 10;

dump top_diversion;



```
grunt>
grunt> list_diversion = FOREACH delayedflights GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
grunt> filtered_diversion = FILTER list_diversion BY (origin is not null) AND (dest is not null) AND (diversion == 1);
grunt> group_origin = GROUP filtered_diversion by (origin,dest);
grunt> list_origin_diversion = FOREACH group_origin generate group, COUNT(filtered_diversion.diversion);
grunt> order_diversion = ORDER list_origin_diversion BY $1 DESC;
grunt> top_diversion = limit order_diversion 10;
grunt> dump top_diversion;
```



```
2017-11-06 22:35:22,393 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-06 22:35:22,395 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-06 22:35:22,453 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-06 22:35:22,453 [main] INFO  org.apache.pig.backend.hadoop.executionengine.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>
```

Submitted By:-

Hardik Kaushik