

## 1. Find out the top 5 most visited destinations.

### script:

```
REGISTER '/Users/Disha/Downloads/piggybank-0.16.0.jar';

A = load '/Users/Disha/Downloads/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 '/Users/Disha/Downloads/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;
```

**Execution:** pig -x local <script\_name>

**output:**

```

job_local543284220_0003 -> job_local713114684_0004,
job_local713114684_0004 -> job_local512453837_0005,
job_local512453837_0005

```

```

2017-09-14 13:58:11,590 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,594 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,597 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,606 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,609 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,612 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,617 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,620 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,622 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,630 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,642 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,650 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,665 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,671 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,681 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 13:58:11,687 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-14 13:58:11,708 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-09-14 13:58:11,708 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-14 13:58:11,747 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2017-09-14 13:58:11,748 [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)
2017-09-14 13:58:11,884 [main] INFO org.apache.pig.Main - Pig script completed in 38 seconds and 549 milliseconds (38549 ms)

```

task2:

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

**script:**

```
REGISTER '/Users/Disha/Downloads/piggybank-0.16.0.jar';
```

```
A = load '/Users/Disha/Downloads/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;
```

**Execution:** pig -x local <script\_name>

**output:**

```
java -Xmx1000m -Djava....16.0-core-h2.jar -x local ~ -bash ~ -bash +

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_local1029258389_0001    ->    job_local1373794402_0002,
job_local1373794402_0002    ->    job_local1609293669_0003,
job_local1609293669_0003    ->    job_local1802329328_0004,
job_local1802329328_0004

2017-09-14 14:02:48,950 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,952 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,954 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,966 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,968 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,971 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,978 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,987 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,989 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,995 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,997 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:48,999 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-09-14 14:02:49,004 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-14 14:02:49,006 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2017-09-14 14:02:49,007 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-14 14:02:49,026 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2017-09-14 14:02:49,026 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
2017-09-14 14:02:49,091 [main] INFO org.apache.pig.Main - Pig script completed in 23 seconds and 636 milliseconds (23636 ms)
Chaitanyas-MacBook-Pro:~ Disha$
```

**task3:**

**Top ten origins with the highest AVG departure delay**

**script:**

```
REGISTER '/Users/Disha/Downloads/piggybank-0.16.0.jar';
```

```
A = load '/Users/Disha/Downloads/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 '/Users/Disha/Downloads/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;
```

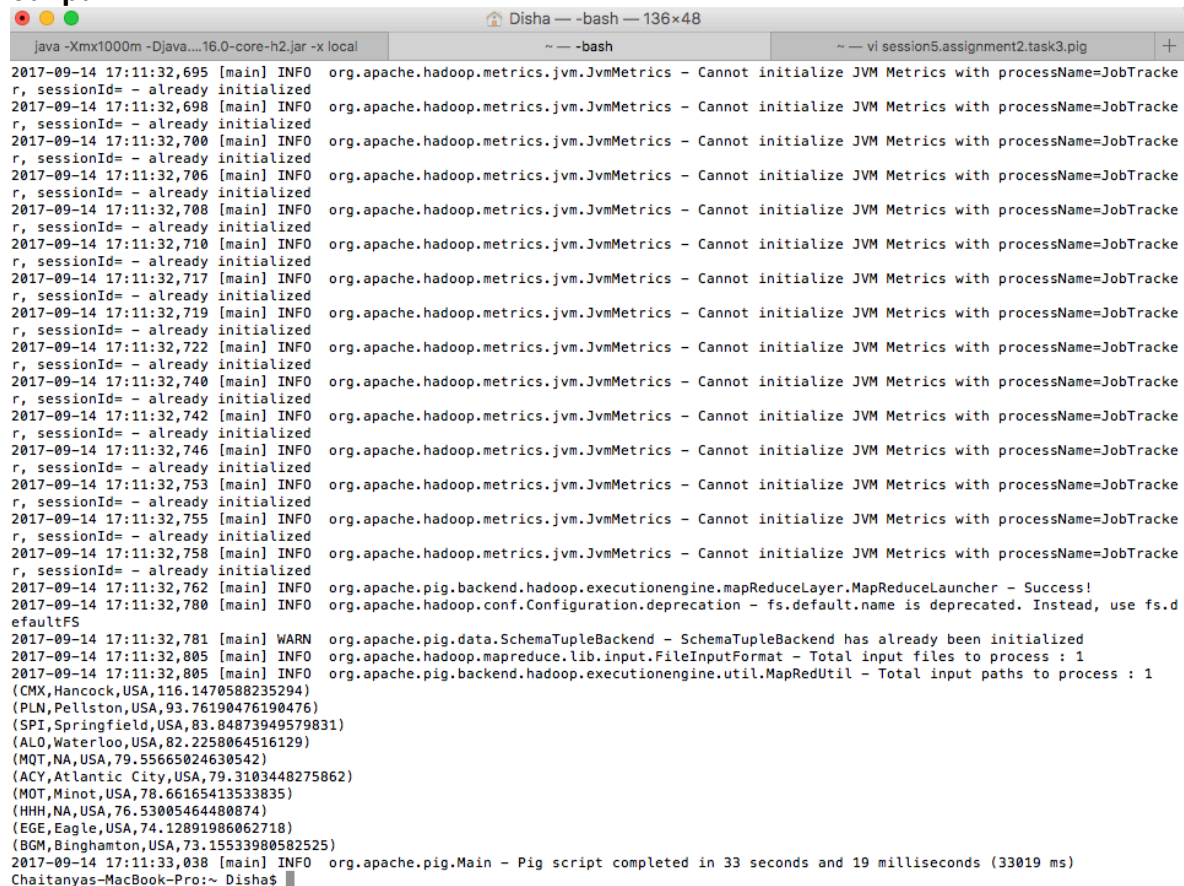
```
Final = foreach Joined generate $0,$1,$2,$4;
```

```
Final_Result = ORDER Final by $3 DESC;
```

```
dump Final_Result;
```

**Execution:** pig -x local <script\_name>

**output:**



```
java -Xmx1000m -Djava....16.0-core-h2.jar -x local ~ -bash ~ -- vi session5.assignment2.task3.pig  
2017-09-14 17:11:32,695 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,698 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,700 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,706 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,708 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,710 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,717 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,719 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,722 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,740 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,742 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,746 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,753 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,755 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,758 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke  
r, sessionId= - already initialized  
2017-09-14 17:11:32,762 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
2017-09-14 17:11:32,780 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d  
efaultFS  
2017-09-14 17:11:32,781 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized  
2017-09-14 17:11:32,805 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1  
2017-09-14 17:11:32,805 [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)  
2017-09-14 17:11:33,038 [main] INFO org.apache.pig.Main - Pig script completed in 33 seconds and 19 milliseconds (33019 ms)  
Chaitanyas-MacBook-Pro:~ Disha$
```

**Task4:**

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

## script:

```
REGISTER '/Users/Disha/Downloads/piggybank-0.16.0.jar';

A = load '/Users/Disha/Downloads/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);

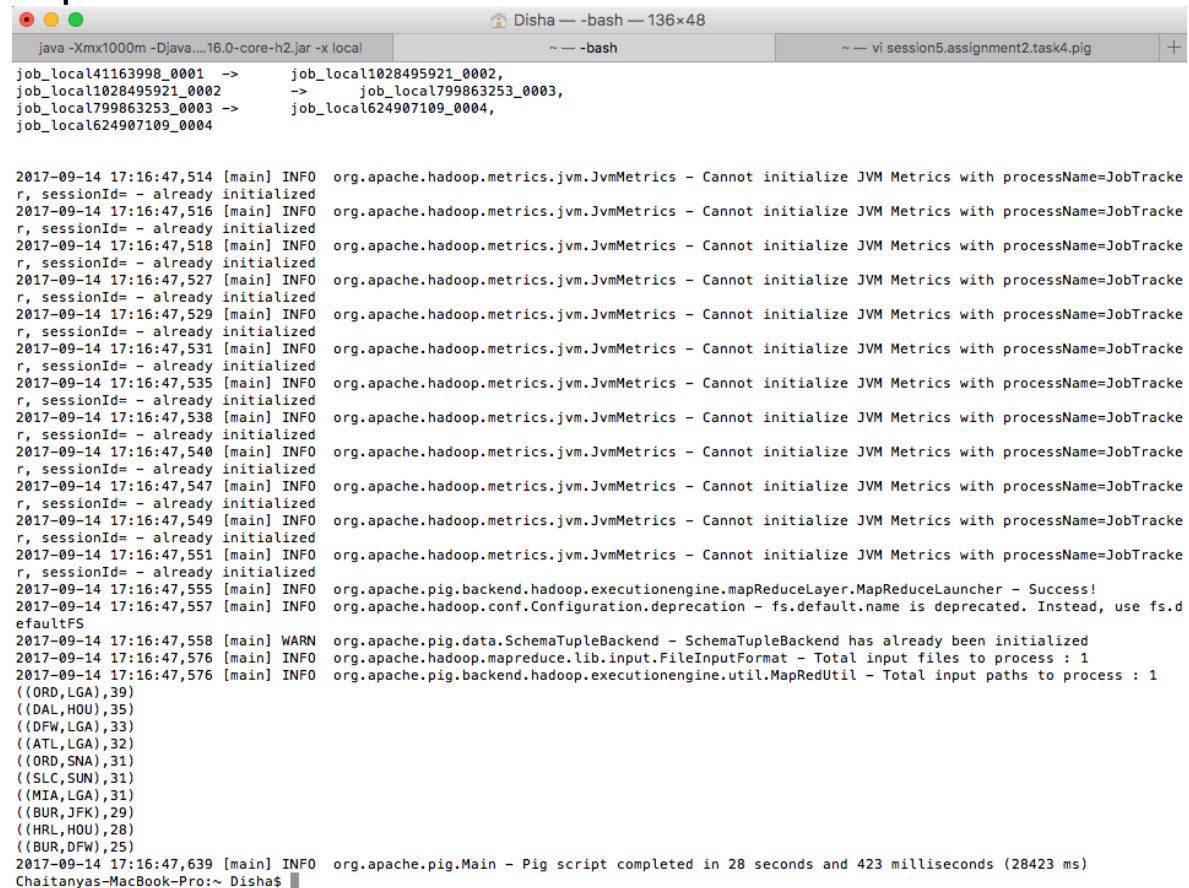
F = ORDER E BY $1 DESC;

Result = limit F 10;

dump Result;
```

**Execution:** pig -x local <script\_name>

## output:



```
java -Xmx1000m -Djava....16.0-core-h2.jar -x local
~ -- bash
~ -- vi session5.assignment2.task4.pig

job_local141163998_0001 -> job_local1028495921_0002,
job_local1028495921_0002 -> job_local799863253_0003,
job_local799863253_0003 -> job_local624907109_0004,
job_local624907109_0004

2017-09-14 17:16:47,514 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,516 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,518 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,527 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,529 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,531 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,535 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,538 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,540 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,547 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,549 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,551 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-09-14 17:16:47,555 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-14 17:16:47,557 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-09-14 17:16:47,558 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-14 17:16:47,576 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2017-09-14 17:16:47,576 [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)
2017-09-14 17:16:47,639 [main] INFO org.apache.pig.Main - Pig script completed in 28 seconds and 423 milliseconds (28423 ms)
Chaitanyas-MacBook-Pro:~ Disha$
```

