FINAL PROJECT

INFO7250- Engg Of Big Data

Ankit Yadav

NUID- 001271369

Northeastern University

CONTENTS

0. Overview about the dataset

I am using the data on **Airline On-Time Statistics and Delay Causes from**

http://stat-computing.org/dataexpo/2009/the-data.html

This is dataset containing information about airline schedule with following columns:

Variable descriptions

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code

19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

The reason of selections this data set is that it has many numbers of columns which will enable me to use various MapReduce algorithms studies in the course for different types of analysis.

Also, the data is evenly segregated in yearly basis. So, in case If I can am unable to load complete data in my computer then too I can do the same analysis on small portion of same data more easily.

In this project I will try to answer following questions :-

1. Which month, time or day of week contributed in maximum delay in airline departure and/or arrivals ?
2. At what time during the day the airlines are most busy?
3. What were various causes of delay ?
4. Depending on departure and arrival time, which destination is best efficient from which starting city?

Apart from above analysis, I will try to do more analysis using Apache Hive and Apache Pig.

1. Analysis of Flight Data using MapReduce on Hadoop

1- Getting total count of all the data:

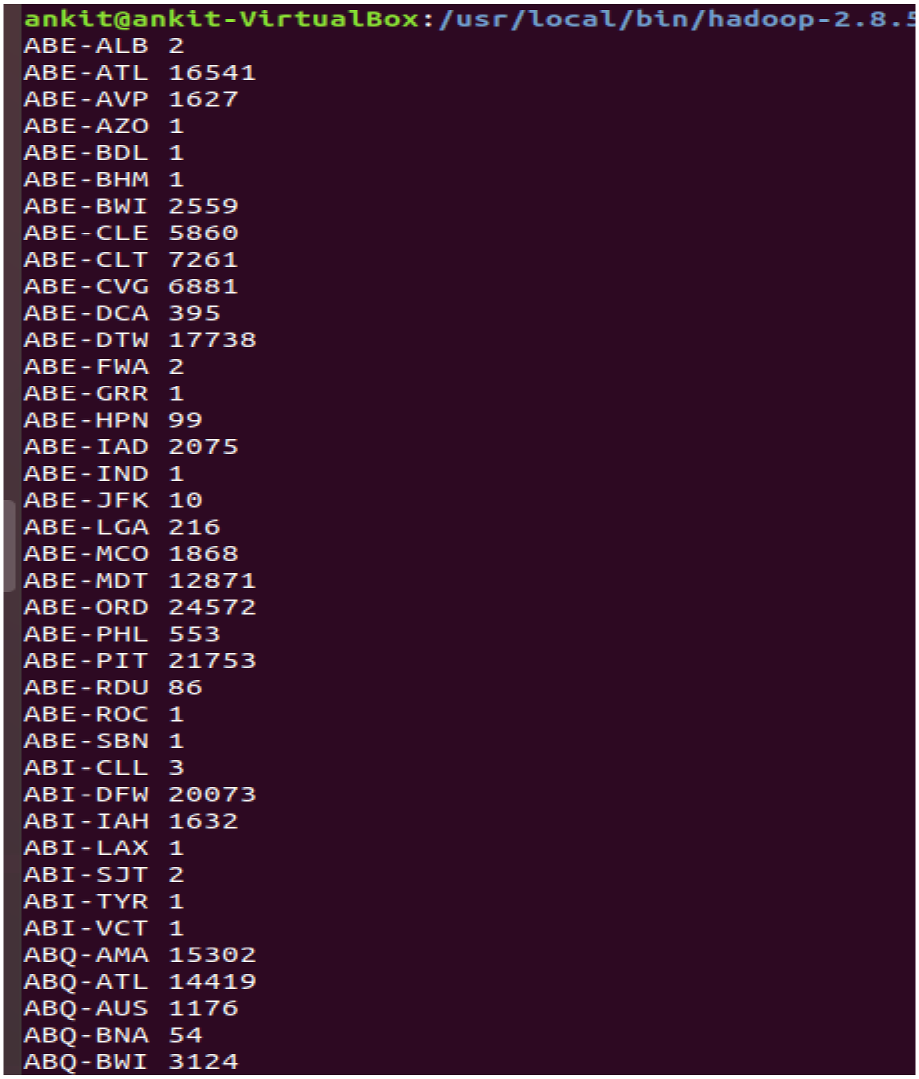
This is a very basic map reduce use case in which we count the whole data to get a sense of how many total records are there:

```
hadoop jar ~/Downloads/ProjectJars/count.jar
hadoop.project.total_count.MRCount /flight-data /FinalProjectMROutput/2-
Total-Data-Count
```

The final count is: **123534970**

2- Getting the total flights from all source destinations pairs in from 1987 to 2008:

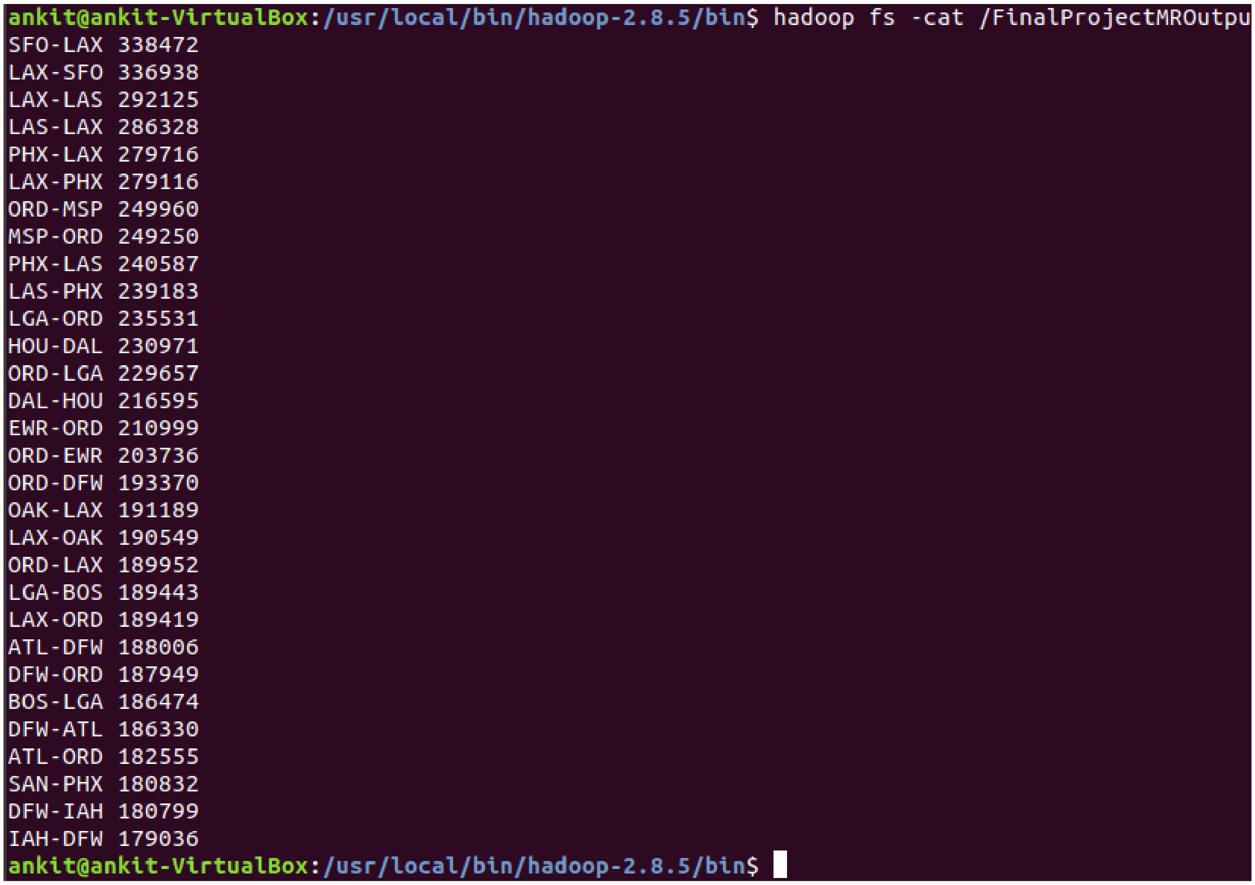
This was a huge data and MapReduce made this analysis quite simple and fast:



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5
ABE-ALB 2
ABE-ATL 16541
ABE-AVP 1627
ABE-AZO 1
ABE-BDL 1
ABE-BHM 1
ABE-BWI 2559
ABE-CLE 5860
ABE-CLT 7261
ABE-CVG 6881
ABE-DCA 395
ABE-DTW 17738
ABE-FWA 2
ABE-GRR 1
ABE-HPN 99
ABE-IAD 2075
ABE-IND 1
ABE-JFK 10
ABE-LGA 216
ABE-MCO 1868
ABE-MDT 12871
ABE-ORD 24572
ABE-PHL 553
ABE-PIT 21753
ABE-RDU 86
ABE-ROC 1
ABE-SBN 1
ABI-CLL 3
ABI-DFW 20073
ABI-IAH 1632
ABI-LAX 1
ABI-SJT 2
ABI-TYR 1
ABI-VCT 1
ABQ-AMA 15302
ABQ-ATL 14419
ABQ-AUS 1176
ABQ-BNA 54
ABQ-BWI 3124
```

3: Top 30 source destination pairs

Sorting the above data to get top 30 most busy Source Destination pair:



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutpu
SFO-LAX 338472
LAX-SFO 336938
LAX-LAS 292125
LAS-LAX 286328
PHX-LAX 279716
LAX-PHX 279116
ORD-MSP 249960
MSP-ORD 249250
PHX-LAS 240587
LAS-PHX 239183
LGA-ORD 235531
HOU-DAL 230971
ORD-LGA 229657
DAL-HOU 216595
EWR-ORD 210999
ORD-EWR 203736
ORD-DFW 193370
OAK-LAX 191189
LAX-OAK 190549
ORD-LAX 189952
LGA-BOS 189443
LAX-ORD 189419
ATL-DFW 188006
DFW-ORD 187949
BOS-LGA 186474
DFW-ATL 186330
ATL-ORD 182555
SAN-PHX 180832
DFW-IAH 180799
IAH-DFW 179036
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

4: Delay in flight percentage

We considered the delay greater than or equal to 15 minutes as delay . Now we need to count those flights which had delay greater than or equal to 15 minutes.

<u>Delayed flight Count:</u>

Percentage of departure delayed flights: Total Flight Count/ Departure Delayed Flight count

= (19690422/123534970) * 100 = 15.94 %

So, this shows that the actual delay greater that 15 minutes is very less and generally flights depart on time.

Let’s check the same for arrival delay

Percentage of departure delayed flights: Total Flight Count/ Delayed Flight count

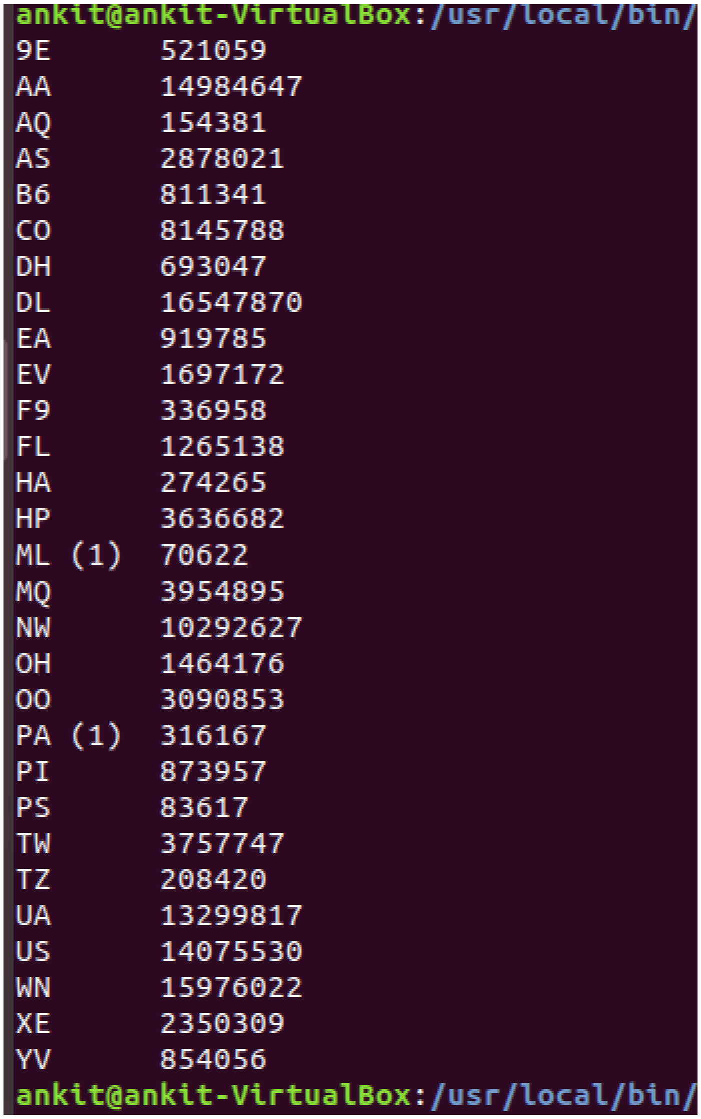
= (24627925/123534970) * 100 = 19.9 %

So, the delay in departure and arrival is between 15 to 20 % range.

So, it shows that overall flights are mostly on time from/to all source destination

5- Count of unique carrier’s flights

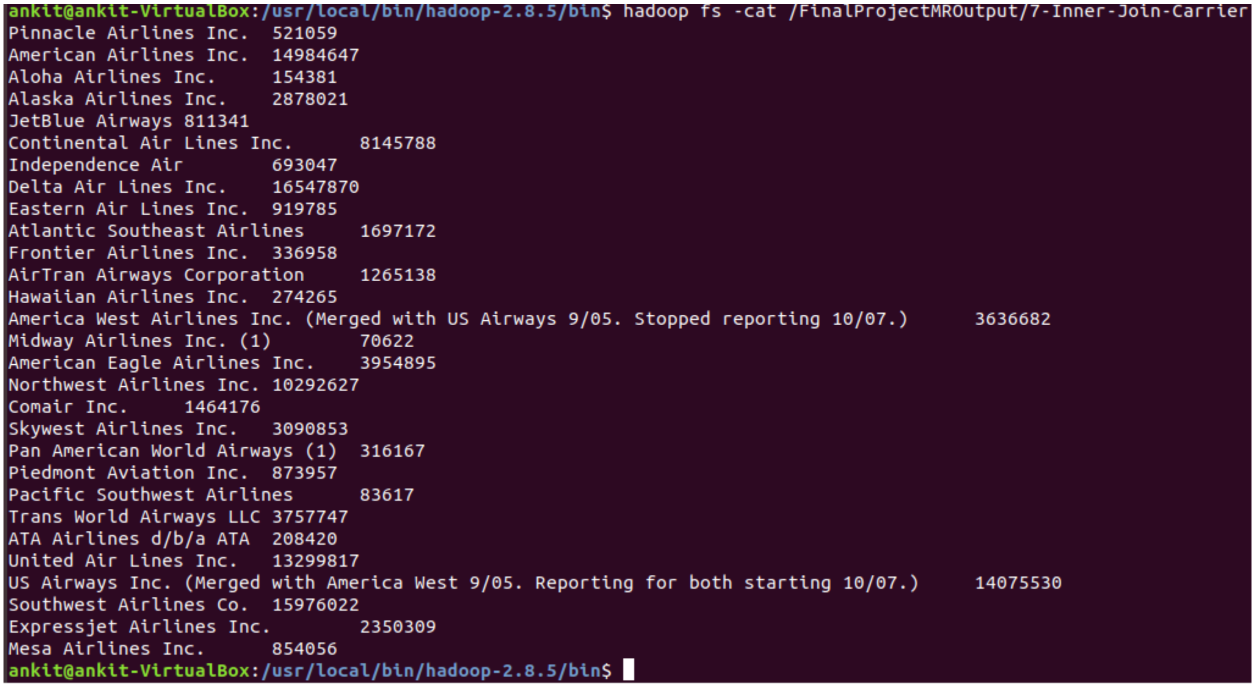
The data for unique carriers are as follows:



```
ankit@ankit-VirtualBox:/usr/local/bin/
9E       521059
AA       14984647
AQ       154381
AS       2878021
B6       811341
CO       8145788
DH       693047
DL       16547870
EA       919785
EV       1697172
F9       336958
FL       1265138
HA       274265
HP       3636682
ML (1)   70622
MQ       3954895
NW       10292627
OH       1464176
OO       3090853
PA (1)   316167
PI       873957
PS       83617
TW       3757747
TZ       208420
UA       13299817
US       14075530
WN       15976022
XE       2350309
YV       854056
ankit@ankit-VirtualBox:/usr/local/bin/
```

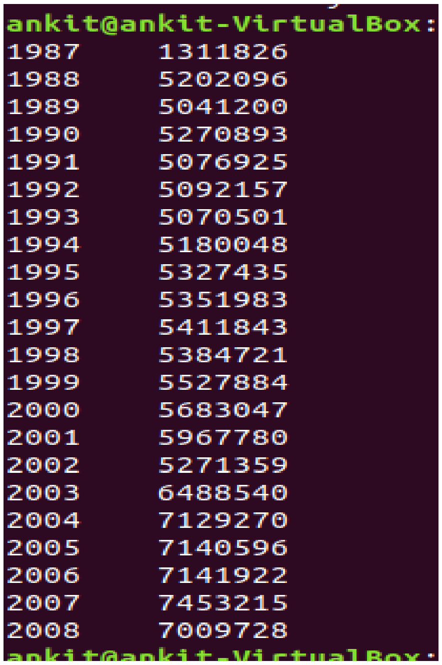
6- Inner Join to get the full name for unique carriers

We did inner join with between two files to get carrier names instead of carrier codes



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/7-Inner-Join-Carrier
Pinnacle Airlines Inc.  521059
American Airlines Inc.  14984647
Aloha Airlines Inc.     154381
Alaska Airlines Inc.    2878021
JetBlue Airways 811341
Continental Air Lines Inc.      8145788
Independence Air        693047
Delta Air Lines Inc.    16547870
Eastern Air Lines Inc.  919785
Atlantic Southeast Airlines     1697172
Frontier Airlines Inc.  336958
AirTran Airways Corporation     1265138
Hawaiian Airlines Inc.  274265
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)      3636682
Midway Airlines Inc. (1)        70622
American Eagle Airlines Inc.    3954895
Northwest Airlines Inc. 10292627
Comair Inc.     1464176
Skywest Airlines Inc.   3090853
Pan American World Airways (1)  316167
Piedmont Aviation Inc.  873957
Pacific Southwest Airlines      83617
Trans World Airways LLC 3757747
ATA Airlines d/b/a ATA  208420
United Air Lines Inc.   13299817
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)     14075530
Southwest Airlines Co.  15976022
Expressjet Airlines Inc.        2350309
Mesa Airlines Inc.      854056
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

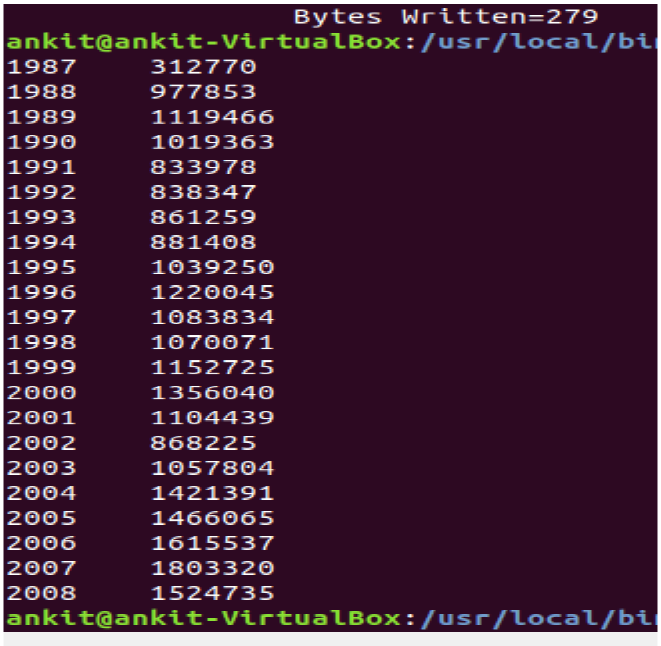
7- Getting Flight data by year



```
ankit@ankit-VirtualBox:
1987    1311826
1988    5202096
1989    5041200
1990    5270893
1991    5076925
1992    5092157
1993    5070501
1994    5180048
1995    5327435
1996    5351983
1997    5411843
1998    5384721
1999    5527884
2000    5683047
2001    5967780
2002    5271359
2003    6488540
2004    7129270
2005    7140596
2006    7141922
2007    7453215
2008    7009728
ankit@ankit-VirtualBox:
```

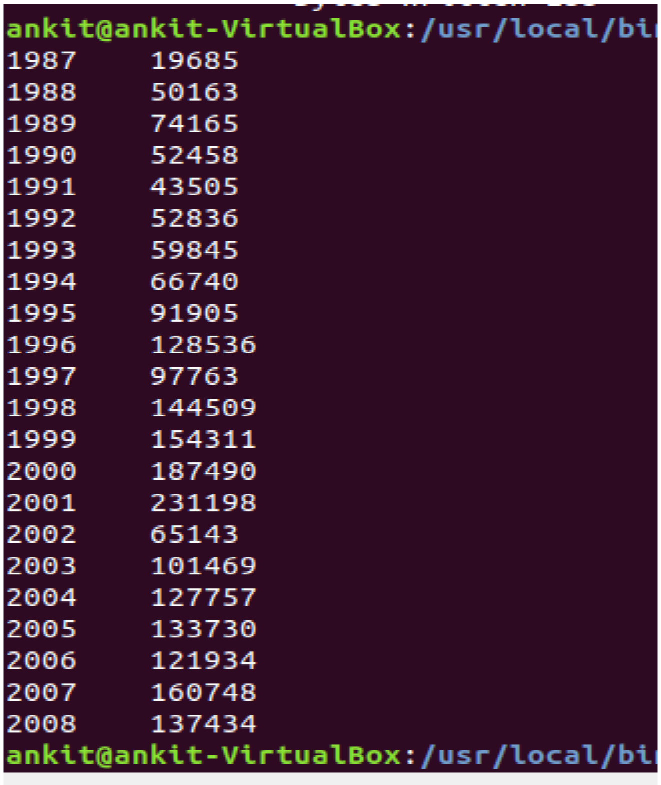
8- Delayed flights per year

In this we will check delayed flights per year(we will count flights as delayed only is the delay time if greater than equal to 15 minutes)



```
                Bytes Written=279
ankit@ankit-VirtualBox:/usr/local/bi
1987    312770
1988    977853
1989    1119466
1990    1019363
1991    833978
1992    838347
1993    861259
1994    881408
1995    1039250
1996    1220045
1997    1083834
1998    1070071
1999    1152725
2000    1356040
2001    1104439
2002    868225
2003    1057804
2004    1421391
2005    1466065
2006    1615537
2007    1803320
2008    1524735
ankit@ankit-VirtualBox:/usr/local/bi
```

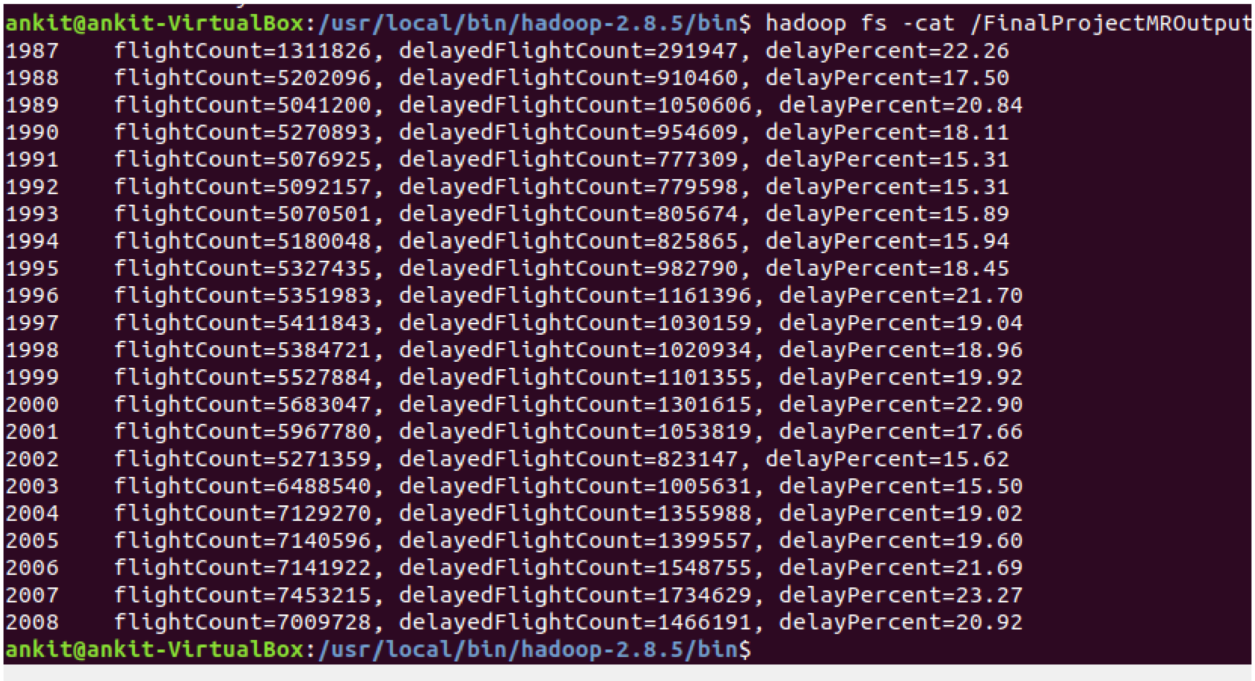
9- Cancelled flights by year



```
ankit@ankit-VirtualBox:/usr/local/bi
1987    19685
1988    50163
1989    74165
1990    52458
1991    43505
1992    52836
1993    59845
1994    66740
1995    91905
1996    128536
1997    97763
1998    144509
1999    154311
2000    187490
2001    231198
2002    65143
2003    101469
2004    127757
2005    133730
2006    121934
2007    160748
2008    137434
ankit@ankit-VirtualBox:/usr/local/bi
```

10- Ratio of delayed flights per year to total flights

We can get percentage of delayed flights per year also



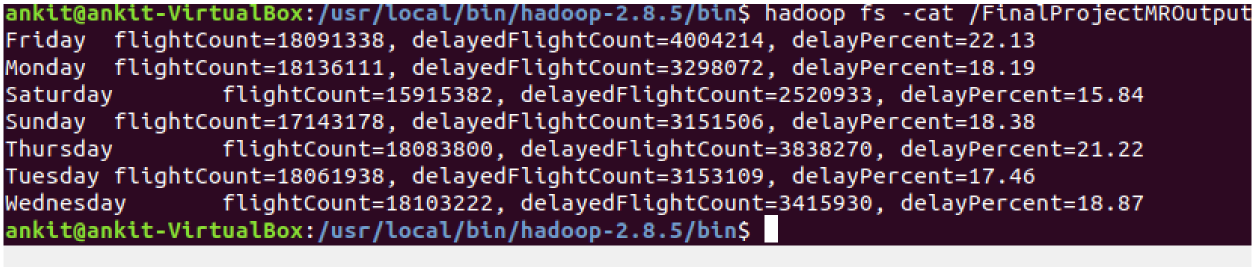
```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput
1987    flightCount=1311826, delayedFlightCount=291947, delayPercent=22.26
1988    flightCount=5202096, delayedFlightCount=910460, delayPercent=17.50
1989    flightCount=5041200, delayedFlightCount=1050606, delayPercent=20.84
1990    flightCount=5270893, delayedFlightCount=954609, delayPercent=18.11
1991    flightCount=5076925, delayedFlightCount=777309, delayPercent=15.31
1992    flightCount=5092157, delayedFlightCount=779598, delayPercent=15.31
1993    flightCount=5070501, delayedFlightCount=805674, delayPercent=15.89
1994    flightCount=5180048, delayedFlightCount=825865, delayPercent=15.94
1995    flightCount=5327435, delayedFlightCount=982790, delayPercent=18.45
1996    flightCount=5351983, delayedFlightCount=1161396, delayPercent=21.70
1997    flightCount=5411843, delayedFlightCount=1030159, delayPercent=19.04
1998    flightCount=5384721, delayedFlightCount=1020934, delayPercent=18.96
1999    flightCount=5527884, delayedFlightCount=1101355, delayPercent=19.92
2000    flightCount=5683047, delayedFlightCount=1301615, delayPercent=22.90
2001    flightCount=5967780, delayedFlightCount=1053819, delayPercent=17.66
2002    flightCount=5271359, delayedFlightCount=823147, delayPercent=15.62
2003    flightCount=6488540, delayedFlightCount=1005631, delayPercent=15.50
2004    flightCount=7129270, delayedFlightCount=1355988, delayPercent=19.02
2005    flightCount=7140596, delayedFlightCount=1399557, delayPercent=19.60
2006    flightCount=7141922, delayedFlightCount=1548755, delayPercent=21.69
2007    flightCount=7453215, delayedFlightCount=1734629, delayPercent=23.27
2008    flightCount=7009728, delayedFlightCount=1466191, delayPercent=20.92
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

It shows that the years 1991-1994, 2002,2003 were best years to fly as they had least delayed flights (less than 16%).

Years with most delays were- 1987, 2000, 2007 with more than 22% flights delayed

11- Total flights by day of week and ratio to delayed

Following is the data of total flights , delayed flights and their ratio.

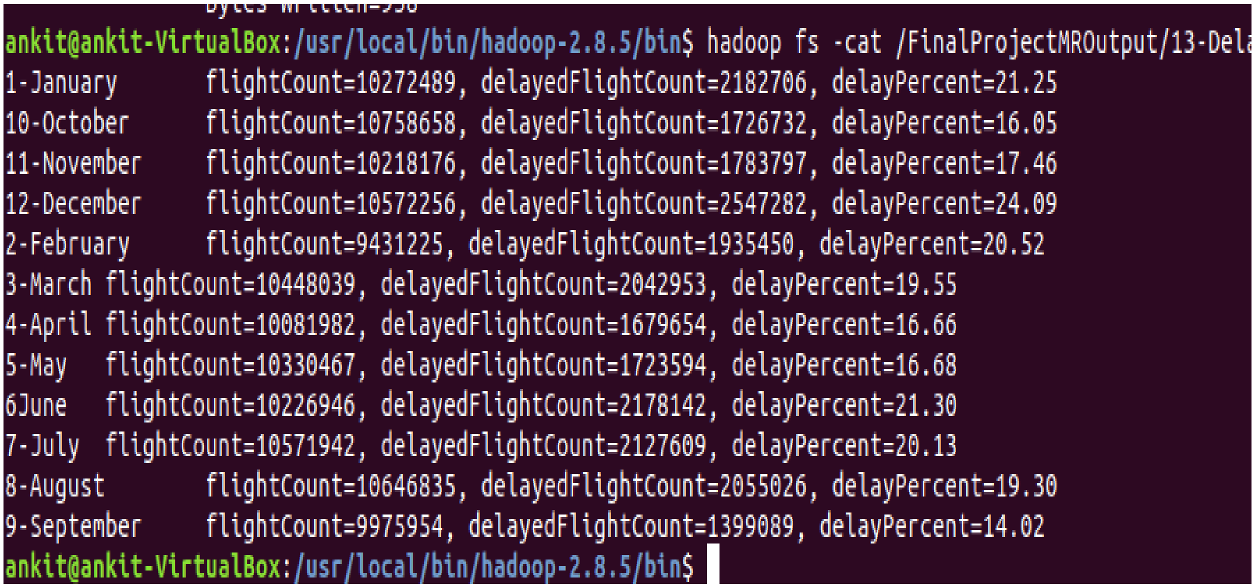


```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput
Friday  flightCount=18091338, delayedFlightCount=4004214, delayPercent=22.13
Monday  flightCount=18136111, delayedFlightCount=3298072, delayPercent=18.19
Saturday        flightCount=15915382, delayedFlightCount=2520933, delayPercent=15.84
Sunday  flightCount=17143178, delayedFlightCount=3151506, delayPercent=18.38
Thursday        flightCount=18083800, delayedFlightCount=3838270, delayPercent=21.22
Tuesday flightCount=18061938, delayedFlightCount=3153109, delayPercent=17.46
Wednesday       flightCount=18103222, delayedFlightCount=3415930, delayPercent=18.87
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

From this data we can infer that maximum delay is on Thursday and Fridays that is when weekends are starting .

Best day to fly are when the weekends ends like Saturday, Sunday or on weekdays

12- Total flights by months of year and ratio to delayed



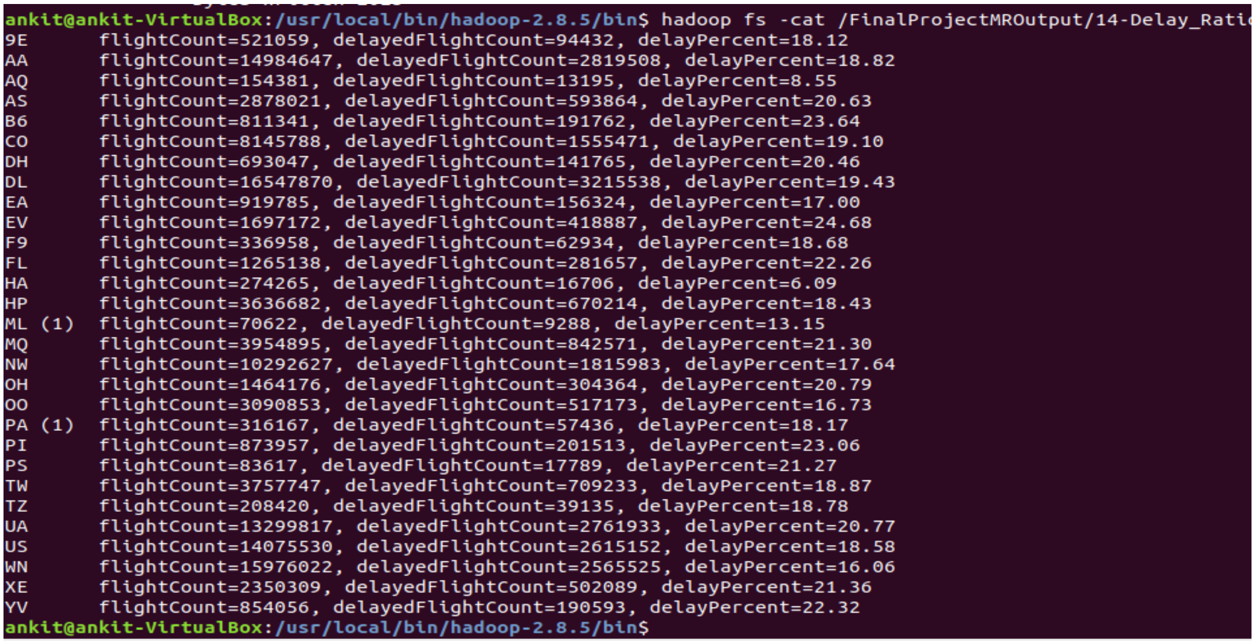
```
                Bytes Written=958
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/13-Dela
1-January       flightCount=10272489, delayedFlightCount=2182706, delayPercent=21.25
10-October      flightCount=10758658, delayedFlightCount=1726732, delayPercent=16.05
11-November     flightCount=10218176, delayedFlightCount=1783797, delayPercent=17.46
12-December     flightCount=10572256, delayedFlightCount=2547282, delayPercent=24.09
2-February      flightCount=9431225, delayedFlightCount=1935450, delayPercent=20.52
3-March flightCount=10448039, delayedFlightCount=2042953, delayPercent=19.55
4-April flightCount=10081982, delayedFlightCount=1679654, delayPercent=16.66
5-May   flightCount=10330467, delayedFlightCount=1723594, delayPercent=16.68
6June   flightCount=10226946, delayedFlightCount=2178142, delayPercent=21.30
7-July  flightCount=10571942, delayedFlightCount=2127609, delayPercent=20.13
8-August        flightCount=10646835, delayedFlightCount=2055026, delayPercent=19.30
9-September     flightCount=9975954, delayedFlightCount=1399089, delayPercent=14.02
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

We can infer from this data that best months to fly was September with least delay- 14 %

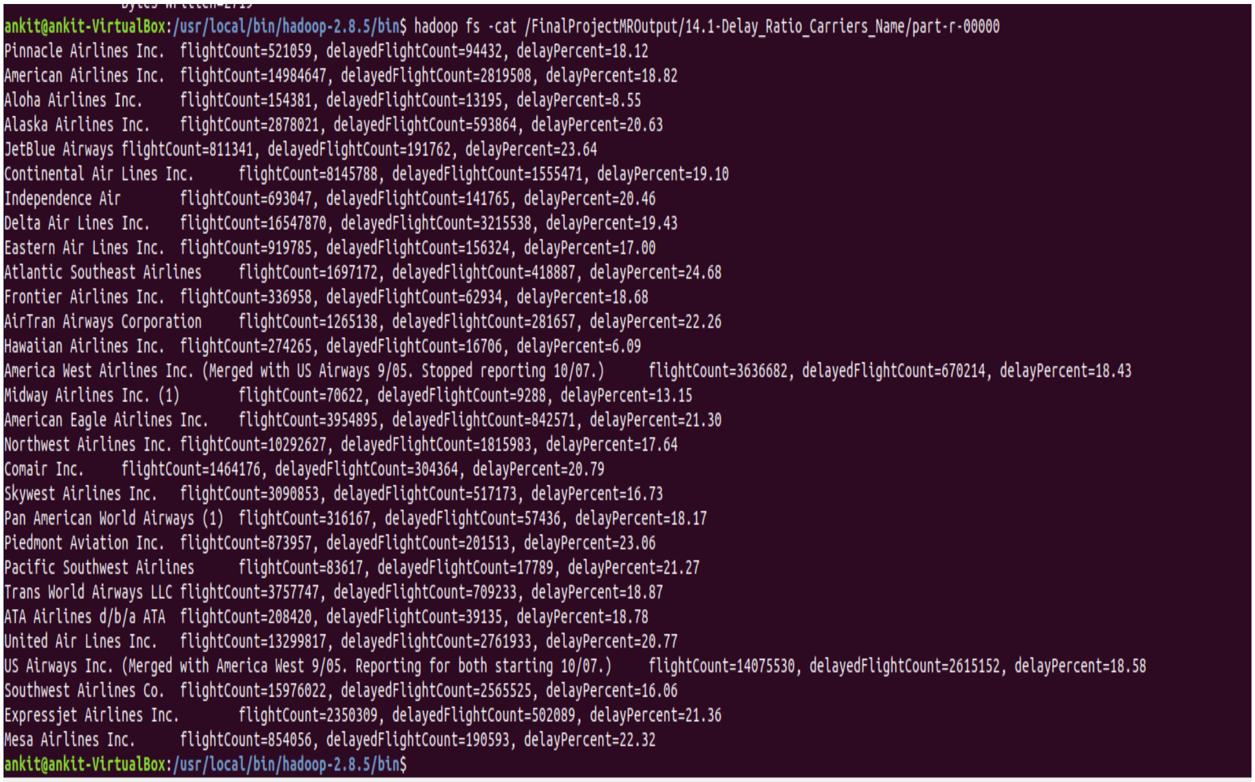
Other good months were- April, May and October with delay – 16%

Worst month was December- 24% flights delayed. It may be due to big holidays season in December.

13- Total delayed flights by flight carriers and ratio of delayed flights



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/14-Delay_Rati
9E      flightCount=521059, delayedFlightCount=94432, delayPercent=18.12
AA      flightCount=14984647, delayedFlightCount=2819508, delayPercent=18.82
AQ      flightCount=154381, delayedFlightCount=13195, delayPercent=8.55
AS      flightCount=2878021, delayedFlightCount=593864, delayPercent=20.63
B6      flightCount=811341, delayedFlightCount=191762, delayPercent=23.64
CO      flightCount=8145788, delayedFlightCount=1555471, delayPercent=19.10
DH      flightCount=693047, delayedFlightCount=141765, delayPercent=20.46
DL      flightCount=16547870, delayedFlightCount=3215538, delayPercent=19.43
EA      flightCount=919785, delayedFlightCount=156324, delayPercent=17.00
EV      flightCount=1697172, delayedFlightCount=418887, delayPercent=24.68
F9      flightCount=336958, delayedFlightCount=62934, delayPercent=18.68
FL      flightCount=1265138, delayedFlightCount=281657, delayPercent=22.26
HA      flightCount=274265, delayedFlightCount=16706, delayPercent=6.09
HP      flightCount=3636682, delayedFlightCount=670214, delayPercent=18.43
ML (1)  flightCount=70622, delayedFlightCount=9288, delayPercent=13.15
MQ      flightCount=3954895, delayedFlightCount=842571, delayPercent=21.30
NW      flightCount=10292627, delayedFlightCount=1815983, delayPercent=17.64
OH      flightCount=1464176, delayedFlightCount=304364, delayPercent=20.79
OO      flightCount=3090853, delayedFlightCount=517173, delayPercent=16.73
PA (1)  flightCount=316167, delayedFlightCount=57436, delayPercent=18.17
PI      flightCount=873957, delayedFlightCount=201513, delayPercent=23.06
PS      flightCount=83617, delayedFlightCount=17789, delayPercent=21.27
TW      flightCount=3757747, delayedFlightCount=709233, delayPercent=18.87
TZ      flightCount=208420, delayedFlightCount=39135, delayPercent=18.78
UA      flightCount=13299817, delayedFlightCount=2761933, delayPercent=20.77
US      flightCount=14075530, delayedFlightCount=2615152, delayPercent=18.58
WN      flightCount=15976022, delayedFlightCount=2565525, delayPercent=16.06
XE      flightCount=2350309, delayedFlightCount=502089, delayPercent=21.36
YV      flightCount=854056, delayedFlightCount=190593, delayPercent=22.32
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```



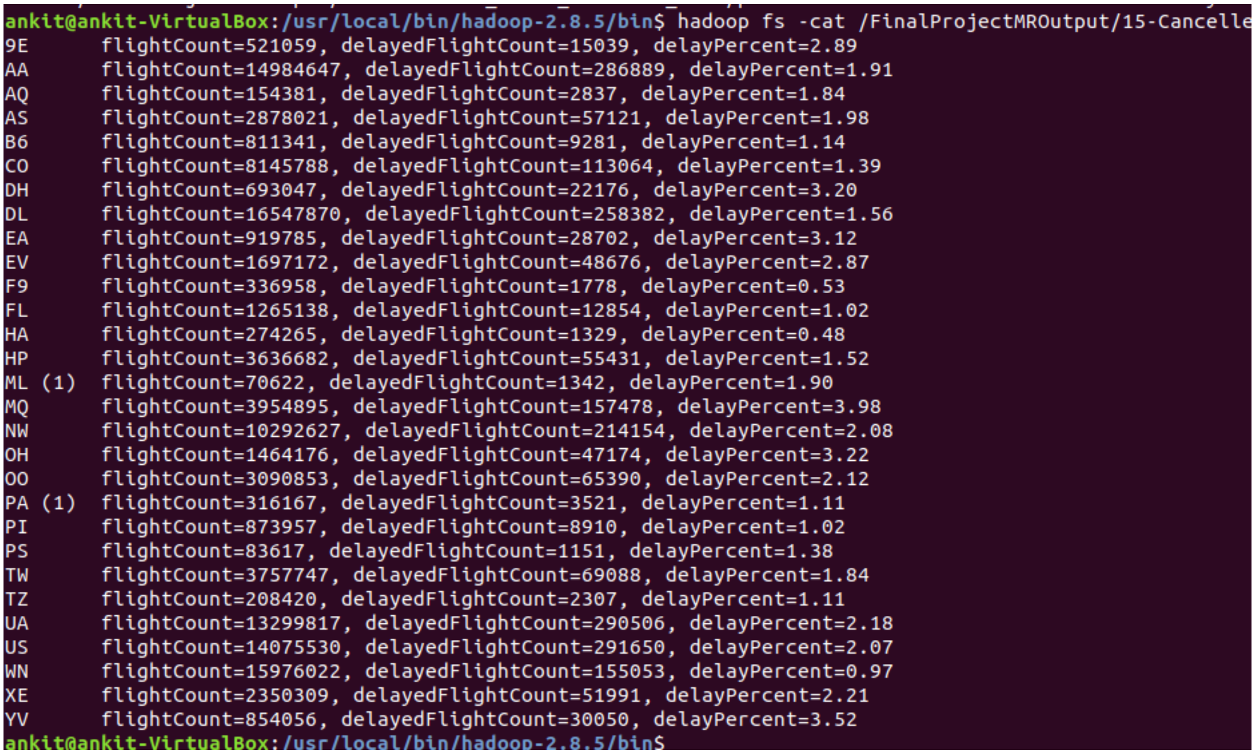
```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/14.1-Delay_Ratio_Carriers_Name/part-r-00000
Pinnacle Airlines Inc.  flightCount=521059, delayedFlightCount=94432, delayPercent=18.12
American Airlines Inc.  flightCount=14984647, delayedFlightCount=2819508, delayPercent=18.82
Aloha Airlines Inc.     flightCount=154381, delayedFlightCount=13195, delayPercent=8.55
Alaska Airlines Inc.    flightCount=2878021, delayedFlightCount=593864, delayPercent=20.63
JetBlue Airways flightCount=811341, delayedFlightCount=191762, delayPercent=23.64
Continental Air Lines Inc.      flightCount=8145788, delayedFlightCount=1555471, delayPercent=19.10
Independence Air        flightCount=693047, delayedFlightCount=141765, delayPercent=20.46
Delta Air Lines Inc.    flightCount=16547870, delayedFlightCount=3215538, delayPercent=19.43
Eastern Air Lines Inc.  flightCount=919785, delayedFlightCount=156324, delayPercent=17.00
Atlantic Southeast Airlines     flightCount=1697172, delayedFlightCount=418887, delayPercent=24.68
Frontier Airlines Inc.  flightCount=336958, delayedFlightCount=62934, delayPercent=18.68
AirTran Airways Corporation     flightCount=1265138, delayedFlightCount=281657, delayPercent=22.26
Hawaiian Airlines Inc.  flightCount=274265, delayedFlightCount=16706, delayPercent=6.09
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)      flightCount=3636682, delayedFlightCount=670214, delayPercent=18.43
Midway Airlines Inc. (1)        flightCount=70622, delayedFlightCount=9288, delayPercent=13.15
American Eagle Airlines Inc.    flightCount=3954895, delayedFlightCount=842571, delayPercent=21.30
Northwest Airlines Inc. flightCount=10292627, delayedFlightCount=1815983, delayPercent=17.64
Comair Inc.     flightCount=1464176, delayedFlightCount=304364, delayPercent=20.79
Skywest Airlines Inc.   flightCount=3090853, delayedFlightCount=517173, delayPercent=16.73
Pan American World Airways (1)  flightCount=316167, delayedFlightCount=57436, delayPercent=18.17
Piedmont Aviation Inc.  flightCount=873957, delayedFlightCount=201513, delayPercent=23.06
Pacific Southwest Airlines      flightCount=83617, delayedFlightCount=17789, delayPercent=21.27
Trans World Airways LLC flightCount=3757747, delayedFlightCount=709233, delayPercent=18.87
ATA Airlines d/b/a ATA  flightCount=208420, delayedFlightCount=39135, delayPercent=18.78
United Air Lines Inc.   flightCount=13299817, delayedFlightCount=2761933, delayPercent=20.77
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)     flightCount=14075530, delayedFlightCount=2615152, delayPercent=18.58
Southwest Airlines Co.  flightCount=15976022, delayedFlightCount=2565525, delayPercent=16.06
Expressjet Airlines Inc.        flightCount=2350309, delayedFlightCount=502089, delayPercent=21.36
Mesa Airlines Inc.      flightCount=854056, delayedFlightCount=190593, delayPercent=22.32
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

By using this analysis we can check which carriers are more prone to delays and can plan flights with those carriers who are less prone to delays.

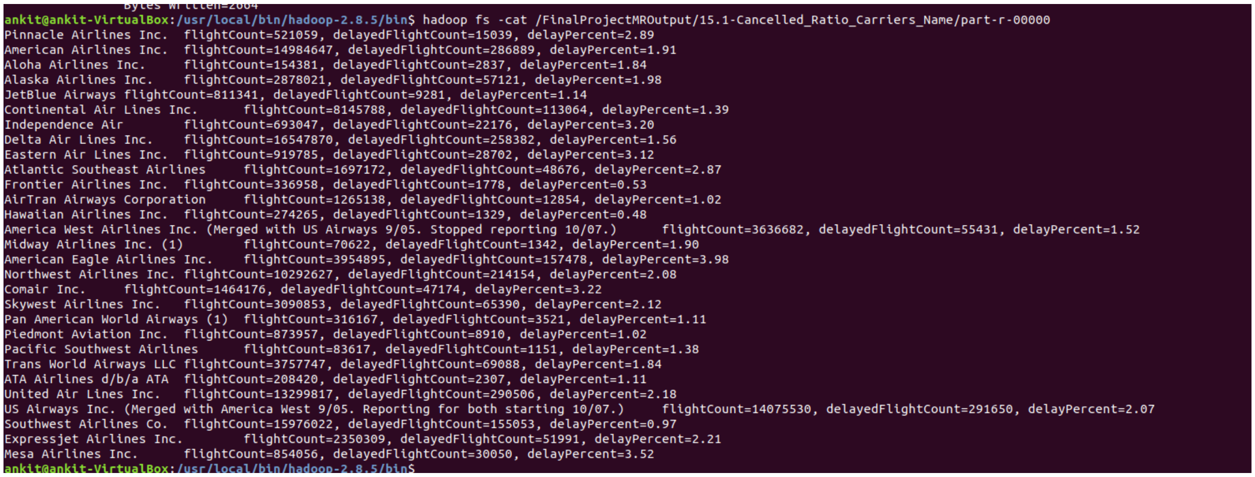
Carriers with least delays- **Hawaiian Airlines, Aloha Airlines** with **6%** and **8%** flights delayed respectively.

Carriers with most delays- **JetBlue Airways, Atlantic Southeast Airlines** with around **24%** flights delayed.

14- Total cancelled flights by flight carriers and ratio of cancelled flights



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/15-Cancelle
9E      flightCount=521059, delayedFlightCount=15039, delayPercent=2.89
AA      flightCount=14984647, delayedFlightCount=286889, delayPercent=1.91
AQ      flightCount=154381, delayedFlightCount=2837, delayPercent=1.84
AS      flightCount=2878021, delayedFlightCount=57121, delayPercent=1.98
B6      flightCount=811341, delayedFlightCount=9281, delayPercent=1.14
CO      flightCount=8145788, delayedFlightCount=113064, delayPercent=1.39
DH      flightCount=693047, delayedFlightCount=22176, delayPercent=3.20
DL      flightCount=16547870, delayedFlightCount=258382, delayPercent=1.56
EA      flightCount=919785, delayedFlightCount=28702, delayPercent=3.12
EV      flightCount=1697172, delayedFlightCount=48676, delayPercent=2.87
F9      flightCount=336958, delayedFlightCount=1778, delayPercent=0.53
FL      flightCount=1265138, delayedFlightCount=12854, delayPercent=1.02
HA      flightCount=274265, delayedFlightCount=1329, delayPercent=0.48
HP      flightCount=3636682, delayedFlightCount=55431, delayPercent=1.52
ML (1)  flightCount=70622, delayedFlightCount=1342, delayPercent=1.90
MQ      flightCount=3954895, delayedFlightCount=157478, delayPercent=3.98
NW      flightCount=10292627, delayedFlightCount=214154, delayPercent=2.08
OH      flightCount=1464176, delayedFlightCount=47174, delayPercent=3.22
OO      flightCount=3090853, delayedFlightCount=65390, delayPercent=2.12
PA (1)  flightCount=316167, delayedFlightCount=3521, delayPercent=1.11
PI      flightCount=873957, delayedFlightCount=8910, delayPercent=1.02
PS      flightCount=83617, delayedFlightCount=1151, delayPercent=1.38
TW      flightCount=3757747, delayedFlightCount=69088, delayPercent=1.84
TZ      flightCount=208420, delayedFlightCount=2307, delayPercent=1.11
UA      flightCount=13299817, delayedFlightCount=290506, delayPercent=2.18
US      flightCount=14075530, delayedFlightCount=291650, delayPercent=2.07
WN      flightCount=15976022, delayedFlightCount=155053, delayPercent=0.97
XE      flightCount=2350309, delayedFlightCount=51991, delayPercent=2.21
YV      flightCount=854056, delayedFlightCount=30050, delayPercent=3.52
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```



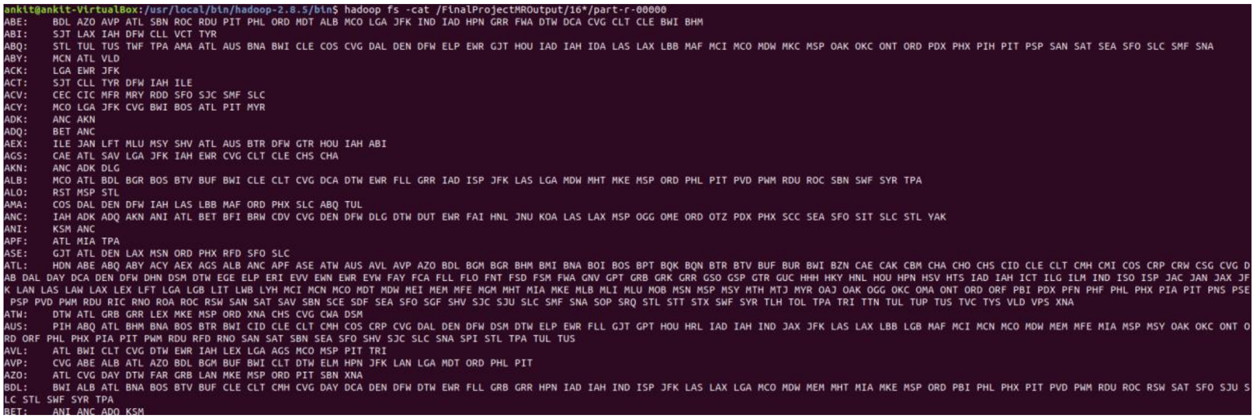
```
                Bytes Written=2664
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/15.1-Cancelled_Ratio_Carriers_Name/part-r-00000
Pinnacle Airlines Inc.  flightCount=521059, delayedFlightCount=15039, delayPercent=2.89
American Airlines Inc.  flightCount=14984647, delayedFlightCount=286889, delayPercent=1.91
Aloha Airlines Inc.     flightCount=154381, delayedFlightCount=2837, delayPercent=1.84
Alaska Airlines Inc.    flightCount=2878021, delayedFlightCount=57121, delayPercent=1.98
JetBlue Airways flightCount=811341, delayedFlightCount=9281, delayPercent=1.14
Continental Air Lines Inc.      flightCount=8145788, delayedFlightCount=113064, delayPercent=1.39
Independence Air        flightCount=693047, delayedFlightCount=22176, delayPercent=3.20
Delta Air Lines Inc.    flightCount=16547870, delayedFlightCount=258382, delayPercent=1.56
Eastern Air Lines Inc.  flightCount=919785, delayedFlightCount=28702, delayPercent=3.12
Atlantic Southeast Airlines     flightCount=1697172, delayedFlightCount=48676, delayPercent=2.87
Frontier Airlines Inc.  flightCount=336958, delayedFlightCount=1778, delayPercent=0.53
AirTran Airways Corporation     flightCount=1265138, delayedFlightCount=12854, delayPercent=1.02
Hawaiian Airlines Inc.  flightCount=274265, delayedFlightCount=1329, delayPercent=0.48
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)      flightCount=3636682, delayedFlightCount=55431, delayPercent=1.52
Midway Airlines Inc. (1)        flightCount=70622, delayedFlightCount=1342, delayPercent=1.90
American Eagle Airlines Inc.    flightCount=3954895, delayedFlightCount=157478, delayPercent=3.98
Northwest Airlines Inc. flightCount=10292627, delayedFlightCount=214154, delayPercent=2.08
Comair Inc.     flightCount=1464176, delayedFlightCount=47174, delayPercent=3.22
Skywest Airlines Inc.   flightCount=3090853, delayedFlightCount=65390, delayPercent=2.12
Pan American World Airways (1)  flightCount=316167, delayedFlightCount=3521, delayPercent=1.11
Piedmont Aviation Inc.  flightCount=873957, delayedFlightCount=8910, delayPercent=1.02
Pacific Southwest Airlines      flightCount=83617, delayedFlightCount=1151, delayPercent=1.38
Trans World Airways LLC flightCount=3757747, delayedFlightCount=69088, delayPercent=1.84
ATA Airlines d/b/a ATA  flightCount=208420, delayedFlightCount=2307, delayPercent=1.11
United Air Lines Inc.   flightCount=13299817, delayedFlightCount=290506, delayPercent=2.18
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)     flightCount=14075530, delayedFlightCount=291650, delayPercent=2.07
Southwest Airlines Co.  flightCount=15976022, delayedFlightCount=155053, delayPercent=0.97
Expressjet Airlines Inc.        flightCount=2350309, delayedFlightCount=51991, delayPercent=2.21
Mesa Airlines Inc.      flightCount=854056, delayedFlightCount=30050, delayPercent=3.52
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

The number of cancelled flights are very less for almost all the carriers less than 4%.

Among them best are **Frontier Airlines, Hawaiian Airlines** with **0.5%** cancelled flights and worst are **American Eagle Airlines, Mesa Airlines** with more than **3.5%** cancelled flights.

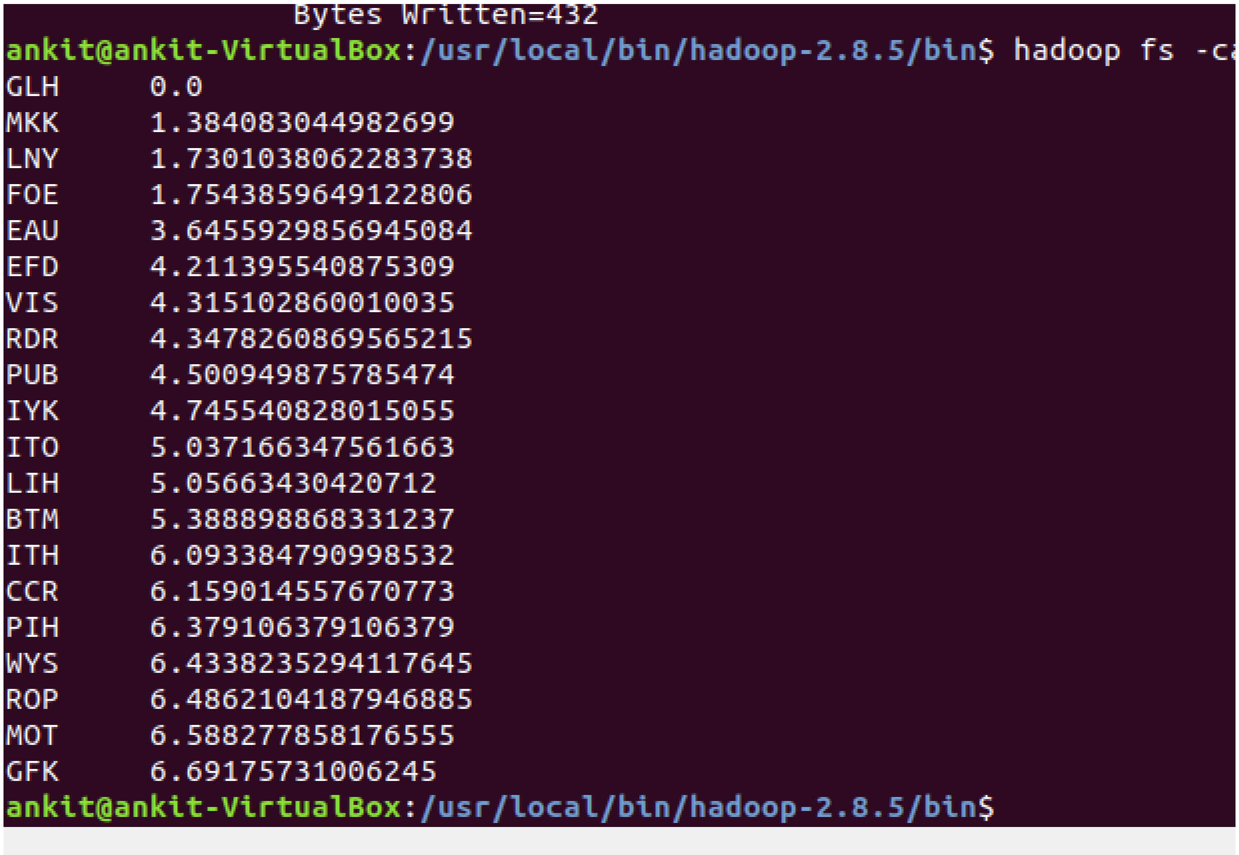
15- Inverted index for all source and destination

This data can help to search for all the destination stations from a particular source stations.



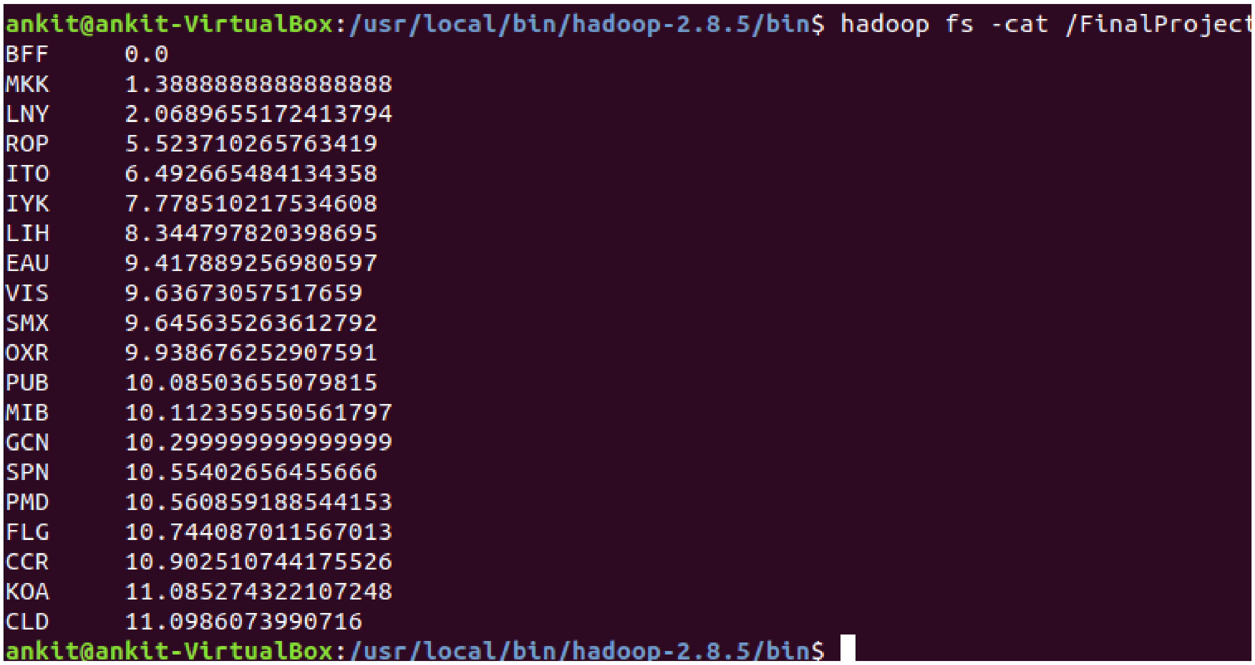
```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/16*/part-r-00000
ABE:    BDL AZO AVP ATL SBN ROC RDU PIT PHL ORD MDT ALB MCO LGA JFK IND IAD HPN GRR FWA DTW DCA CVG CLT CLE BWI BHM
ABI:    SJT LAX IAH DFW CLL VCT TYR
ABQ:    STL TUL TUS TWF TPA AMA ATL AUS BNA BWI CLE COS CVG DAL DEN DFW ELP EWR GJT HOU IAD IAH IDA LAS LAX LBB MAF MCI MCO MDW MKC MSP OAK OKC ONT ORD PDX PHX PIH PIT PSP SAN SAT SEA SFO SLC SMF SNA
ABY:    MCN ATL VLD
ACK:    LGA EWR JFK
ACT:    SJT CLL TYR DFW IAH ILE
ACV:    CEC CIC MFR MRY RDD SFO SJC SMF SLC
ACY:    MCO LGA JFK CVG BWI BOS ATL PIT MYR
ADK:    ANC AKN
ADQ:    BET ANC
AEX:    ILE JAN LFT MLU MSY SHV ATL AUS BTR DFW GTR HOU IAH ABI
AGS:    CAE ATL SAV LGA JFK IAH EWR CVG CLT CLE CHS CHA
AKN:    ANC ADK DLG
ALB:    MCO ATL BDL BGR BOS BTV BUF BWI CLE CLT CVG DCA DTW EWR FLL GRR IAD ISP JFK LAS LGA MDW MHT MKE MSP ORD PHL PIT PVD PWM RDU ROC SBN SWF SYR TPA
ALO:    RST MSP STL
AMA:    COS DAL DEN DFW IAH LAS LBB MAF ORD PHX SLC ABQ TUL
ANC:    IAH ADK ADQ AKN ANI ATL BET BFI BRW CDV CVG DEN DFW DLG DTW DUT EWR FAI HNL JNU KOA LAS LAX MSP OGG OME ORD OTZ PDX PHX SCC SEA SFO SIT SLC STL YAK
ANI:    KSM ANC
APF:    ATL MIA TPA
ASE:    GJT ATL DEN LAX MSN ORD PHX RFD SFO SLC
ATL:    HDN ABE ABQ ABY ACY AEX AGS ALB ANC APF ASE ATW AUS AVL AVP AZO BDL BGM BGR BHM BMI BNA BOI BOS BPT BQK BQN BTR BTV BUF BUR BWI BZN CAE CAK CBM CHA CHO CHS CID CLE CLT CMH CMI COS CRP CRW CSG CVG D
AB DAL DAY DCA DEN DFW DHN DSM DTW EGE ELP ERI EVV EWN EWR EYW FAY FCA FLL FLO FNT FSD FSM FWA GNV GPT GRB GRK GRR GSO GSP GTR GUC HHH HKY HNL HOU HPN HSV HTS IAD IAH ICT ILG ILM IND ISO ISP JAC JAN JAX JF
K LAN LAS LAW LAX LEX LFT LGA LGB LIT LWB LYH MCI MCN MCO MDT MDW MEI MEM MFE MGM MHT MIA MKE MLB MLI MLU MOB MSN MSP MSY MTH MTJ MYR OAJ OAK OGG OKC OMA ONT ORD ORF PBI PDX PFN PHF PHL PHX PIA PIT PNS PSE
 PSP PVD PWM RDU RIC RNO ROA ROC RSW SAN SAT SAV SBN SCE SDF SEA SFO SGF SHV SJC SJU SLC SMF SNA SOP SRQ STL STT STX SWF SYR TLH TOL TPA TRI TTN TUL TUP TUS TVC TYS VLD VPS XNA
ATW:    DTW ATL GRB GRR LEX MKE MSP ORD XNA CHS CVG CWA DSM
AUS:    PIH ABQ ATL BHM BNA BOS BTR BWI CID CLE CLT CMH COS CRP CVG DAL DEN DFW DSM DTW ELP EWR FLL GJT GPT HOU HRL IAD IAH IND JAX JFK LAS LAX LBB LGB MAF MCI MCN MCO MDW MEM MFE MIA MSP MSY OAK OKC ONT O
RD ORF PHL PHX PIA PIT PWM RDU RFD RNO SAN SAT SBN SEA SFO SHV SJC SLC SNA SPI STL TPA TUL TUS
AVL:    ATL BWI CLT CVG DTW EWR IAH LEX LGA AGS MCO MSP PIT TRI
AVP:    CVG ABE ALB ATL AZO BDL BGM BUF BWI CLT DTW ELM HPN JFK LAN LGA MDT ORD PHL PIT
AZO:    ATL CVG DAY DTW FAR GRB LAN MKE MSP ORD PIT SBN XNA
BDL:    BWI ALB ATL BNA BOS BTV BUF CLE CLT CMH CVG DAY DCA DEN DFW DTW EWR FLL GRB GRR HPN IAD IAH IND ISP JFK LAS LAX LGA MCO MDW MEM MHT MIA MKE MSP ORD PBI PHL PHX PIT PVD PWM RDU ROC RSW SAT SFO SJU S
LC STL SWF SYR TPA
BET:    ANI ANC ADQ KSM
```

16- Top 20 best source station with least departure delayed flight percent



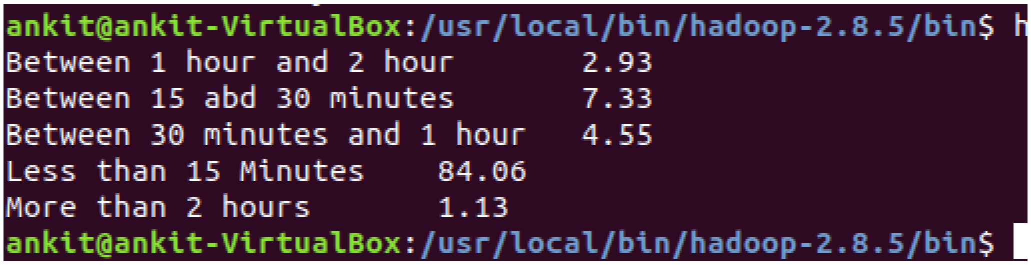
```
                    Bytes Written=432
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ca
GLH     0.0
MKK     1.384083044982699
LNY     1.7301038062283738
FOE     1.7543859649122806
EAU     3.6455929856945084
EFD     4.211395540875309
VIS     4.315102860010035
RDR     4.3478260869565215
PUB     4.500949875785474
IYK     4.745540828015055
ITO     5.037166347561663
LIH     5.05663430420712
BTM     5.388898868331237
ITH     6.093384790998532
CCR     6.159014557670773
PIH     6.379106379106379
WYS     6.4338235294117645
ROP     6.4862104187946885
MOT     6.588277858176555
GFK     6.69175731006245
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

17- Top 20 best destination station with least arrival delayed flight percent



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProject
BFF     0.0
MKK     1.3888888888888888
LNY     2.0689655172413794
ROP     5.523710265763419
ITO     6.492665484134358
IYK     7.778510217534608
LIH     8.344797820398695
EAU     9.417889256980597
VIS     9.63673057517659
SMX     9.645635263612792
OXR     9.938676252907591
PUB     10.08503655079815
MIB     10.112359550561797
GCN     10.299999999999999
SPN     10.55402656455666
PMD     10.560859188544153
FLG     10.744087011567013
CCR     10.902510744175526
KOA     11.085274322107248
CLD     11.0986073990716
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

18- Delay groups- grouping amount of flights per delay groups



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ h
Between 1 hour and 2 hour        2.93
Between 15 abd 30 minutes        7.33
Between 30 minutes and 1 hour    4.55
Less than 15 Minutes    84.06
More than 2 hours       1.13
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

Between 1 hour and 2 hour	**2.93**
Between 15 and 30 minutes	**7.33**
Between 30 minutes and 1 hour	**4.55**
Less than 15 Minutes	**84.06**
More than 2 hours	**1.13**

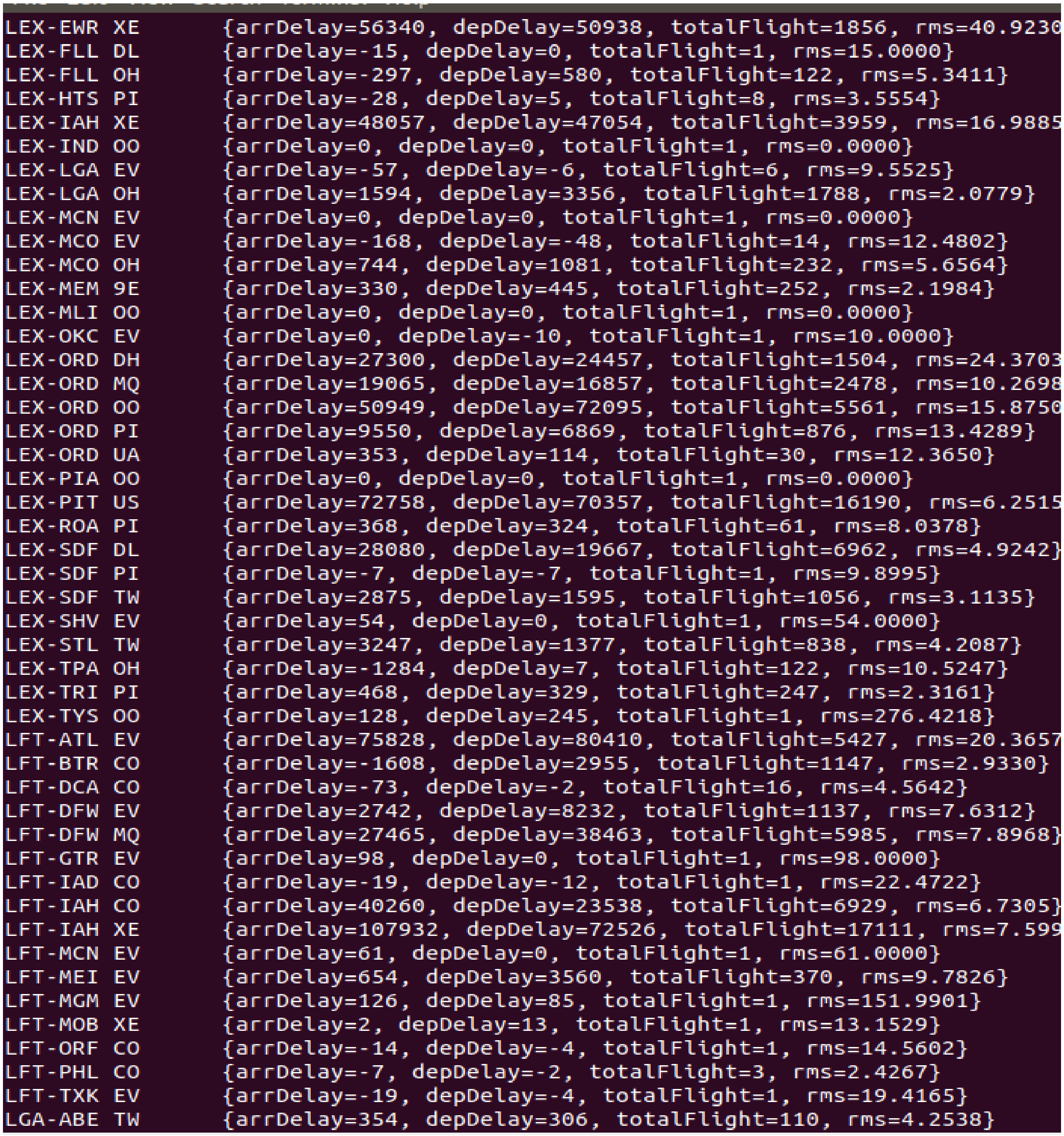
19- Hierarchical data for all source, destination and carrier information

20- Recommendation system- Best carrier for a source destination route

I first calculate average arrival delay and average departure delay for each source destination pair for each carrier.

After this I calculate root mean square value for average arrival delay and average departure delay.

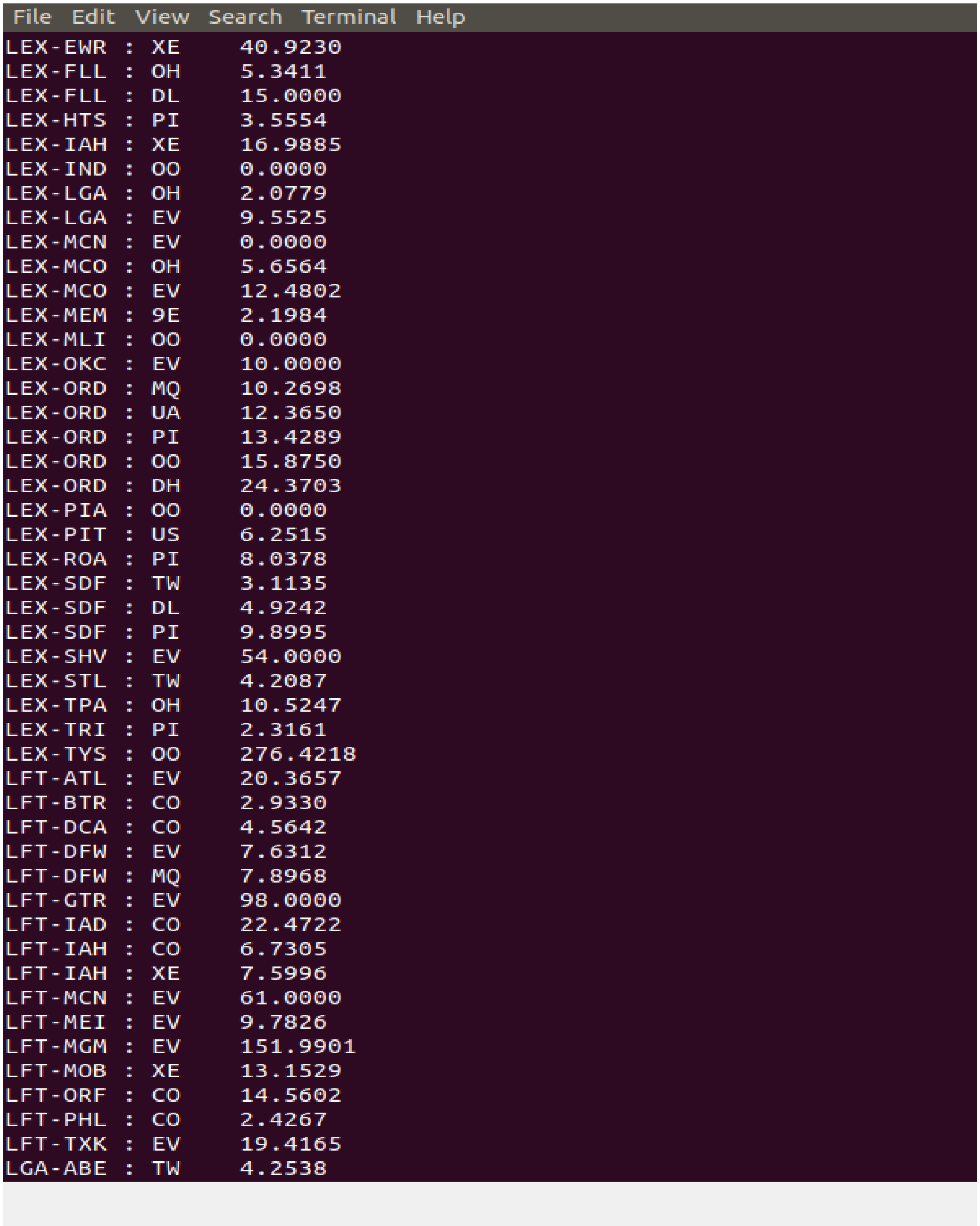
$$\textbf{Rms} = \frac{\sqrt{AvgArrivalDelay^2+AvgDepartureDelay^2}}{}$$



```
LEX-EWR XE      {arrDelay=56340, depDelay=50938, totalFlight=1856, rms=40.9230
LEX-FLL DL      {arrDelay=-15, depDelay=0, totalFlight=1, rms=15.0000}
LEX-FLL OH      {arrDelay=-297, depDelay=580, totalFlight=122, rms=5.3411}
LEX-HTS PI      {arrDelay=-28, depDelay=5, totalFlight=8, rms=3.5554}
LEX-IAH XE      {arrDelay=48057, depDelay=47054, totalFlight=3959, rms=16.9885
LEX-IND OO      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
LEX-LGA EV      {arrDelay=-57, depDelay=-6, totalFlight=6, rms=9.5525}
LEX-LGA OH      {arrDelay=1594, depDelay=3356, totalFlight=1788, rms=2.0779}
LEX-MCN EV      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
LEX-MCO EV      {arrDelay=-168, depDelay=-48, totalFlight=14, rms=12.4802}
LEX-MCO OH      {arrDelay=744, depDelay=1081, totalFlight=232, rms=5.6564}
LEX-MEM 9E      {arrDelay=330, depDelay=445, totalFlight=252, rms=2.1984}
LEX-MLI OO      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
LEX-OKC EV      {arrDelay=0, depDelay=-10, totalFlight=1, rms=10.0000}
LEX-ORD DH      {arrDelay=27300, depDelay=24457, totalFlight=1504, rms=24.3703
LEX-ORD MQ      {arrDelay=19065, depDelay=16857, totalFlight=2478, rms=10.2698
LEX-ORD OO      {arrDelay=50949, depDelay=72095, totalFlight=5561, rms=15.8750
LEX-ORD PI      {arrDelay=9550, depDelay=6869, totalFlight=876, rms=13.4289}
LEX-ORD UA      {arrDelay=353, depDelay=114, totalFlight=30, rms=12.3650}
LEX-PIA OO      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
LEX-PIT US      {arrDelay=72758, depDelay=70357, totalFlight=16190, rms=6.2515
LEX-ROA PI      {arrDelay=368, depDelay=324, totalFlight=61, rms=8.0378}
LEX-SDF DL      {arrDelay=28080, depDelay=19667, totalFlight=6962, rms=4.9242}
LEX-SDF PI      {arrDelay=-7, depDelay=-7, totalFlight=1, rms=9.8995}
LEX-SDF TW      {arrDelay=2875, depDelay=1595, totalFlight=1056, rms=3.1135}
LEX-SHV EV      {arrDelay=54, depDelay=0, totalFlight=1, rms=54.0000}
LEX-STL TW      {arrDelay=3247, depDelay=1377, totalFlight=838, rms=4.2087}
LEX-TPA OH      {arrDelay=-1284, depDelay=7, totalFlight=122, rms=10.5247}
LEX-TRI PI      {arrDelay=468, depDelay=329, totalFlight=247, rms=2.3161}
LEX-TYS OO      {arrDelay=128, depDelay=245, totalFlight=1, rms=276.4218}
LFT-ATL EV      {arrDelay=75828, depDelay=80410, totalFlight=5427, rms=20.3657
LFT-BTR CO      {arrDelay=-1608, depDelay=2955, totalFlight=1147, rms=2.9330}
LFT-DCA CO      {arrDelay=-73, depDelay=-2, totalFlight=16, rms=4.5642}
LFT-DFW EV      {arrDelay=2742, depDelay=8232, totalFlight=1137, rms=7.6312}
LFT-DFW MQ      {arrDelay=27465, depDelay=38463, totalFlight=5985, rms=7.8968}
LFT-GTR EV      {arrDelay=98, depDelay=0, totalFlight=1, rms=98.0000}
LFT-IAD CO      {arrDelay=-19, depDelay=-12, totalFlight=1, rms=22.4722}
LFT-IAH CO      {arrDelay=40260, depDelay=23538, totalFlight=6929, rms=6.7305}
LFT-IAH XE      {arrDelay=107932, depDelay=72526, totalFlight=17111, rms=7.599
LFT-MCN EV      {arrDelay=61, depDelay=0, totalFlight=1, rms=61.0000}
LFT-MEI EV      {arrDelay=654, depDelay=3560, totalFlight=370, rms=9.7826}
LFT-MGM EV      {arrDelay=126, depDelay=85, totalFlight=1, rms=151.9901}
LFT-MOB XE      {arrDelay=2, depDelay=13, totalFlight=1, rms=13.1529}
LFT-ORF CO      {arrDelay=-14, depDelay=-4, totalFlight=1, rms=14.5602}
LFT-PHL CO      {arrDelay=-7, depDelay=-2, totalFlight=3, rms=2.4267}
LFT-TXK EV      {arrDelay=-19, depDelay=-4, totalFlight=1, rms=19.4165}
LGA-ABE TW      {arrDelay=354, depDelay=306, totalFlight=110, rms=4.2538}
```

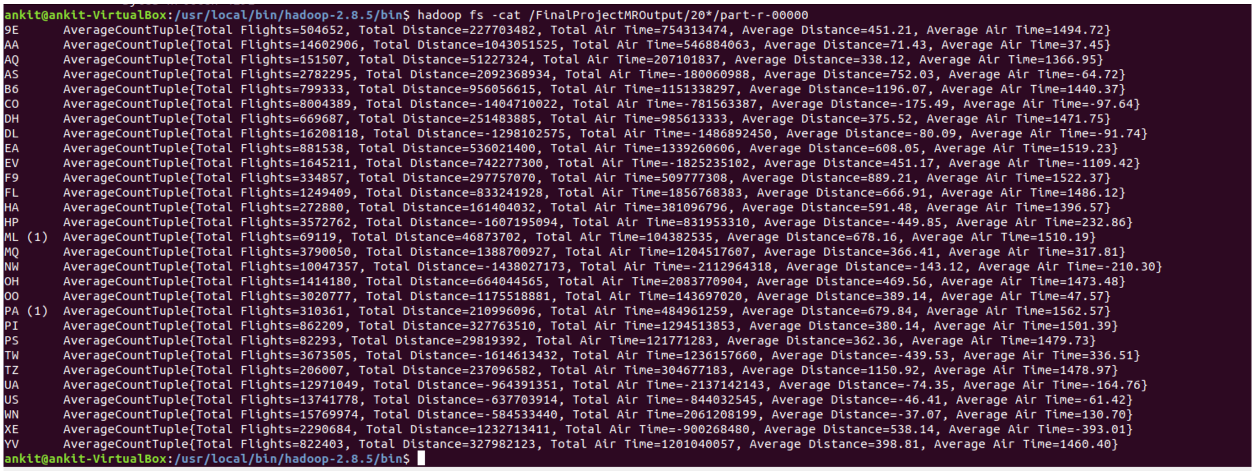
After this I sorted the carriers for all source destination pair in ascending order by RMS value.

It gives user a recommendation for choosing a carrier between a source destination pair with least arrival and/or departure delay.

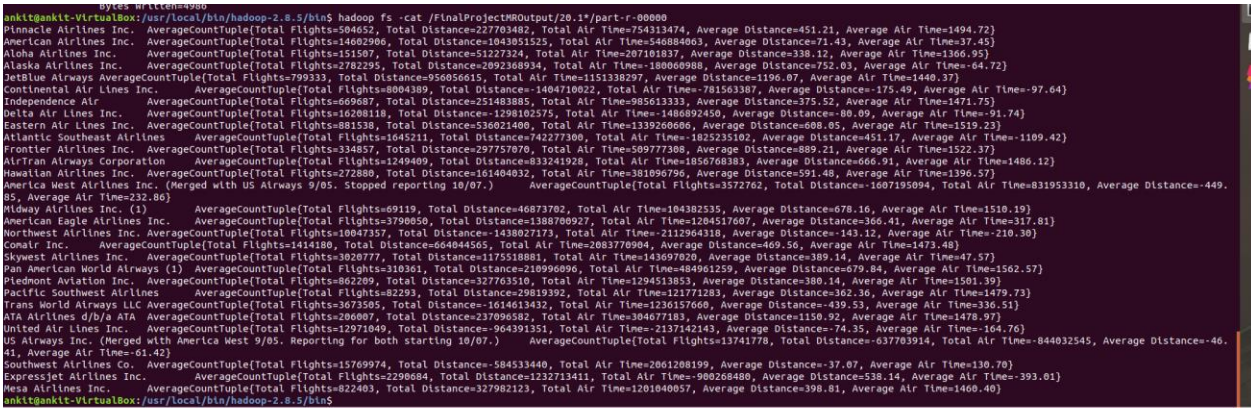


```
LEX-EWR : XE     40.9230
LEX-FLL : OH     5.3411
LEX-FLL : DL     15.0000
LEX-HTS : PI     3.5554
LEX-IAH : XE     16.9885
LEX-IND : OO     0.0000
LEX-LGA : OH     2.0779
LEX-LGA : EV     9.5525
LEX-MCN : EV     0.0000
LEX-MCO : OH     5.6564
LEX-MCO : EV     12.4802
LEX-MEM : 9E     2.1984
LEX-MLI : OO     0.0000
LEX-OKC : EV     10.0000
LEX-ORD : MQ     10.2698
LEX-ORD : UA     12.3650
LEX-ORD : PI     13.4289
LEX-ORD : OO     15.8750
LEX-ORD : DH     24.3703
LEX-PIA : OO     0.0000
LEX-PIT : US     6.2515
LEX-ROA : PI     8.0378
LEX-SDF : TW     3.1135
LEX-SDF : DL     4.9242
LEX-SDF : PI     9.8995
LEX-SHV : EV     54.0000
LEX-STL : TW     4.2087
LEX-TPA : OH     10.5247
LEX-TRI : PI     2.3161
LEX-TYS : OO     276.4218
LFT-ATL : EV     20.3657
LFT-BTR : CO     2.9330
LFT-DCA : CO     4.5642
LFT-DFW : EV     7.6312
LFT-DFW : MQ     7.8968
LFT-GTR : EV     98.0000
LFT-IAD : CO     22.4722
LFT-IAH : CO     6.7305
LFT-IAH : XE     7.5996
LFT-MCN : EV     61.0000
LFT-MEI : EV     9.7826
LFT-MGM : EV     151.9901
LFT-MOB : XE     13.1529
LFT-ORF : CO     14.5602
LFT-PHL : CO     2.4267
LFT-TXK : EV     19.4165
LGA-ABE : TW     4.2538
```

21- Average flying distance per carrier

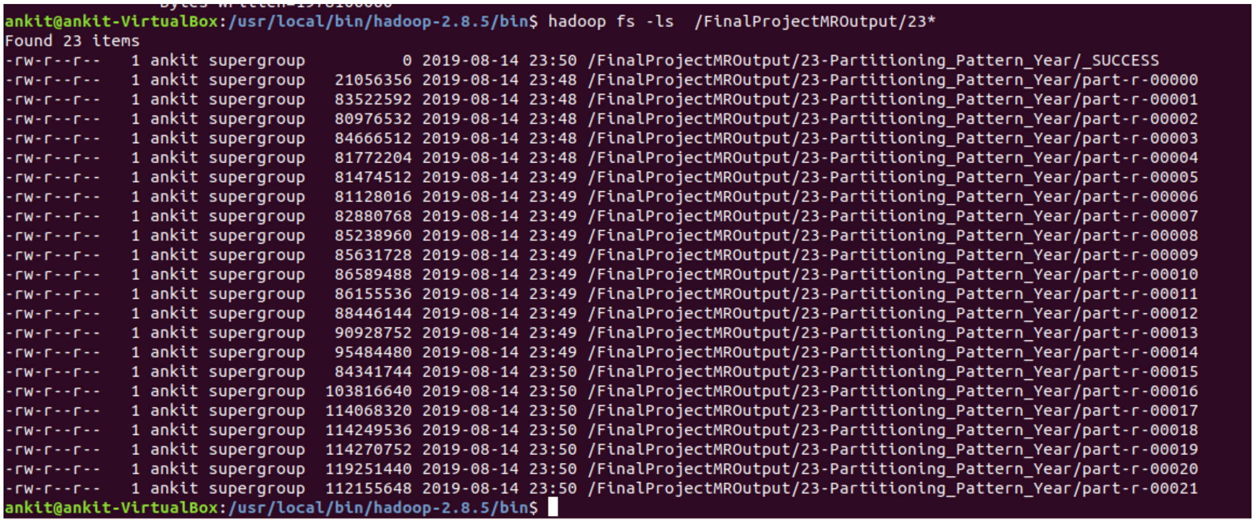


```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/20*/part-r-00000
9E      AverageCountTuple{Total Flights=504652, Total Distance=227703482, Total Air Time=754313474, Average Distance=451.21, Average Air Time=1494.72}
AA      AverageCountTuple{Total Flights=14602906, Total Distance=1043051525, Total Air Time=546884063, Average Distance=71.43, Average Air Time=37.45}
AQ      AverageCountTuple{Total Flights=151507, Total Distance=51227324, Total Air Time=207101837, Average Distance=338.12, Average Air Time=1366.95}
AS      AverageCountTuple{Total Flights=2782295, Total Distance=2092368934, Total Air Time=-180060988, Average Distance=752.03, Average Air Time=-64.72}
B6      AverageCountTuple{Total Flights=799333, Total Distance=956056615, Total Air Time=1151338297, Average Distance=1196.07, Average Air Time=1440.37}
CO      AverageCountTuple{Total Flights=8004389, Total Distance=-1404710022, Total Air Time=-781563387, Average Distance=-175.49, Average Air Time=-97.64}
DH      AverageCountTuple{Total Flights=669687, Total Distance=251483885, Total Air Time=985613333, Average Distance=375.52, Average Air Time=1471.75}
DL      AverageCountTuple{Total Flights=16208118, Total Distance=-1298102575, Total Air Time=-1486892450, Average Distance=-80.09, Average Air Time=-91.74}
EA      AverageCountTuple{Total Flights=881538, Total Distance=536021400, Total Air Time=1339260606, Average Distance=608.05, Average Air Time=1519.23}
EV      AverageCountTuple{Total Flights=1645211, Total Distance=742277300, Total Air Time=-1825235102, Average Distance=451.17, Average Air Time=-1109.42}
F9      AverageCountTuple{Total Flights=334857, Total Distance=297757070, Total Air Time=509777308, Average Distance=889.21, Average Air Time=1522.37}
FL      AverageCountTuple{Total Flights=1249409, Total Distance=833241928, Total Air Time=1856768383, Average Distance=666.91, Average Air Time=1486.12}
HA      AverageCountTuple{Total Flights=272880, Total Distance=161404032, Total Air Time=381096796, Average Distance=591.48, Average Air Time=1396.57}
HP      AverageCountTuple{Total Flights=3572762, Total Distance=-1607195094, Total Air Time=831953310, Average Distance=-449.85, Average Air Time=232.86}
ML (1)  AverageCountTuple{Total Flights=69119, Total Distance=46873702, Total Air Time=104382535, Average Distance=678.16, Average Air Time=1510.19}
MQ      AverageCountTuple{Total Flights=3790050, Total Distance=1388700927, Total Air Time=1204517607, Average Distance=366.41, Average Air Time=317.81}
NW      AverageCountTuple{Total Flights=10047357, Total Distance=-1438027173, Total Air Time=-2112964318, Average Distance=-143.12, Average Air Time=-210.30}
OH      AverageCountTuple{Total Flights=1414180, Total Distance=664044565, Total Air Time=2083770904, Average Distance=469.56, Average Air Time=1473.48}
OO      AverageCountTuple{Total Flights=3020777, Total Distance=1175518881, Total Air Time=143697020, Average Distance=389.14, Average Air Time=47.57}
PA (1)  AverageCountTuple{Total Flights=310361, Total Distance=210996096, Total Air Time=484961259, Average Distance=679.84, Average Air Time=1562.57}
PI      AverageCountTuple{Total Flights=862209, Total Distance=327763510, Total Air Time=1294513853, Average Distance=380.14, Average Air Time=1501.39}
PS      AverageCountTuple{Total Flights=82293, Total Distance=29819392, Total Air Time=121771283, Average Distance=362.36, Average Air Time=1479.73}
TW      AverageCountTuple{Total Flights=3673505, Total Distance=-1614613432, Total Air Time=1236157660, Average Distance=-439.53, Average Air Time=336.51}
TZ      AverageCountTuple{Total Flights=206007, Total Distance=237096582, Total Air Time=304677183, Average Distance=1150.92, Average Air Time=1478.97}
UA      AverageCountTuple{Total Flights=12971049, Total Distance=-964391351, Total Air Time=-2137142143, Average Distance=-74.35, Average Air Time=-164.76}
US      AverageCountTuple{Total Flights=13741778, Total Distance=-637703914, Total Air Time=-844032545, Average Distance=-46.41, Average Air Time=-61.42}
WN      AverageCountTuple{Total Flights=15769974, Total Distance=-584533440, Total Air Time=2061208199, Average Distance=-37.07, Average Air Time=130.70}
XE      AverageCountTuple{Total Flights=2290684, Total Distance=1232713411, Total Air Time=-900268480, Average Distance=538.14, Average Air Time=-393.01}
YV      AverageCountTuple{Total Flights=822403, Total Distance=327982123, Total Air Time=1201040057, Average Distance=398.81, Average Air Time=1460.40}
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```



```
        Bytes Written=4986
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/20.1*/part-r-00000
Pinnacle Airlines Inc.  AverageCountTuple{Total Flights=504652, Total Distance=227703482, Total Air Time=754313474, Average Distance=451.21, Average Air Time=1494.72}
American Airlines Inc.  AverageCountTuple{Total Flights=14602906, Total Distance=1043051525, Total Air Time=546884063, Average Distance=71.43, Average Air Time=37.45}
Aloha Airlines Inc.     AverageCountTuple{Total Flights=151507, Total Distance=51227324, Total Air Time=207101837, Average Distance=338.12, Average Air Time=1366.95}
Alaska Airlines Inc.    AverageCountTuple{Total Flights=2782295, Total Distance=2092368934, Total Air Time=-180060988, Average Distance=752.03, Average Air Time=-64.72}
JetBlue Airways AverageCountTuple{Total Flights=799333, Total Distance=956056615, Total Air Time=1151338297, Average Distance=1196.07, Average Air Time=1440.37}
Continental Air Lines Inc.      AverageCountTuple{Total Flights=8004389, Total Distance=-1404710022, Total Air Time=-781563387, Average Distance=-175.49, Average Air Time=-97.64}
Independence Air        AverageCountTuple{Total Flights=669687, Total Distance=251483885, Total Air Time=985613333, Average Distance=375.52, Average Air Time=1471.75}
Delta Air Lines Inc.    AverageCountTuple{Total Flights=16208118, Total Distance=-1298102575, Total Air Time=-1486892450, Average Distance=-80.09, Average Air Time=-91.74}
Eastern Air Lines Inc.  AverageCountTuple{Total Flights=881538, Total Distance=536021400, Total Air Time=1339260606, Average Distance=608.05, Average Air Time=1519.23}
Atlantic Southeast Airlines     AverageCountTuple{Total Flights=1645211, Total Distance=742277300, Total Air Time=-1825235102, Average Distance=451.17, Average Air Time=-1109.42}
Frontier Airlines Inc.  AverageCountTuple{Total Flights=334857, Total Distance=297757070, Total Air Time=509777308, Average Distance=889.21, Average Air Time=1522.37}
AirTran Airways Corporation     AverageCountTuple{Total Flights=1249409, Total Distance=833241928, Total Air Time=1856768383, Average Distance=666.91, Average Air Time=1486.12}
Hawaiian Airlines Inc.  AverageCountTuple{Total Flights=272880, Total Distance=161404032, Total Air Time=381096796, Average Distance=591.48, Average Air Time=1396.57}
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)      AverageCountTuple{Total Flights=3572762, Total Distance=-1607195094, Total Air Time=831953310, Average Distance=-449.
85, Average Air Time=232.86}
Midway Airlines Inc. (1)        AverageCountTuple{Total Flights=69119, Total Distance=46873702, Total Air Time=104382535, Average Distance=678.16, Average Air Time=1510.19}
American Eagle Airlines Inc.    AverageCountTuple{Total Flights=3790050, Total Distance=1388700927, Total Air Time=1204517607, Average Distance=366.41, Average Air Time=317.81}
Northwest Airlines Inc. AverageCountTuple{Total Flights=10047357, Total Distance=-1438027173, Total Air Time=-2112964318, Average Distance=-143.12, Average Air Time=-210.30}
Comair Inc.     AverageCountTuple{Total Flights=1414180, Total Distance=664044565, Total Air Time=2083770904, Average Distance=469.56, Average Air Time=1473.48}
Skywest Airlines Inc.   AverageCountTuple{Total Flights=3020777, Total Distance=1175518881, Total Air Time=143697020, Average Distance=389.14, Average Air Time=47.57}
Pan American World Airways (1)  AverageCountTuple{Total Flights=310361, Total Distance=210996096, Total Air Time=484961259, Average Distance=679.84, Average Air Time=1562.57}
Piedmont Aviation Inc.  AverageCountTuple{Total Flights=862209, Total Distance=327763510, Total Air Time=1294513853, Average Distance=380.14, Average Air Time=1501.39}
Pacific Southwest Airlines      AverageCountTuple{Total Flights=82293, Total Distance=29819392, Total Air Time=121771283, Average Distance=362.36, Average Air Time=1479.73}
Trans World Airways LLC AverageCountTuple{Total Flights=3673505, Total Distance=-1614613432, Total Air Time=1236157660, Average Distance=-439.53, Average Air Time=336.51}
ATA Airlines d/b/a ATA  AverageCountTuple{Total Flights=206007, Total Distance=237096582, Total Air Time=304677183, Average Distance=1150.92, Average Air Time=1478.97}
United Air Lines Inc.   AverageCountTuple{Total Flights=12971049, Total Distance=-964391351, Total Air Time=-2137142143, Average Distance=-74.35, Average Air Time=-164.76}
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)     AverageCountTuple{Total Flights=13741778, Total Distance=-637703914, Total Air Time=-844032545, Average Distance=-46.
41, Average Air Time=-61.42}
Southwest Airlines Co.  AverageCountTuple{Total Flights=15769974, Total Distance=-584533440, Total Air Time=2061208199, Average Distance=-37.07, Average Air Time=130.70}
Expressjet Airlines Inc.        AverageCountTuple{Total Flights=2290684, Total Distance=1232713411, Total Air Time=-900268480, Average Distance=538.14, Average Air Time=-393.01}
Mesa Airlines Inc.      AverageCountTuple{Total Flights=822403, Total Distance=327982123, Total Air Time=1201040057, Average Distance=398.81, Average Air Time=1460.40}
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

22- Using partitioning pattern on the basis of year

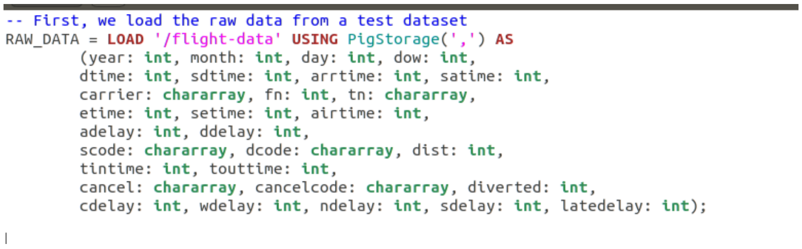


```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls  /FinalProjectMROutput/23*
Found 23 items
-rw-r--r--   1 ankit supergroup          0 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/_SUCCESS
-rw-r--r--   1 ankit supergroup   21056356 2019-08-14 23:48 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00000
-rw-r--r--   1 ankit supergroup   83522592 2019-08-14 23:48 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00001
-rw-r--r--   1 ankit supergroup   80976532 2019-08-14 23:48 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00002
-rw-r--r--   1 ankit supergroup   84666512 2019-08-14 23:48 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00003
-rw-r--r--   1 ankit supergroup   81772204 2019-08-14 23:48 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00004
-rw-r--r--   1 ankit supergroup   81474512 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00005
-rw-r--r--   1 ankit supergroup   81128016 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00006
-rw-r--r--   1 ankit supergroup   82880768 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00007
-rw-r--r--   1 ankit supergroup   85238960 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00008
-rw-r--r--   1 ankit supergroup   85631728 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00009
-rw-r--r--   1 ankit supergroup   86589488 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00010
-rw-r--r--   1 ankit supergroup   86155536 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00011
-rw-r--r--   1 ankit supergroup   88446144 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00012
-rw-r--r--   1 ankit supergroup   90928752 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00013
-rw-r--r--   1 ankit supergroup   95484480 2019-08-14 23:49 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00014
-rw-r--r--   1 ankit supergroup   84341744 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00015
-rw-r--r--   1 ankit supergroup  103816640 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00016
-rw-r--r--   1 ankit supergroup  114068320 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00017
-rw-r--r--   1 ankit supergroup  114249536 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00018
-rw-r--r--   1 ankit supergroup  114270752 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00019
-rw-r--r--   1 ankit supergroup  119251440 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00020
-rw-r--r--   1 ankit supergroup  112155648 2019-08-14 23:50 /FinalProjectMROutput/23-Partitioning_Pattern_Year/part-r-00021
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

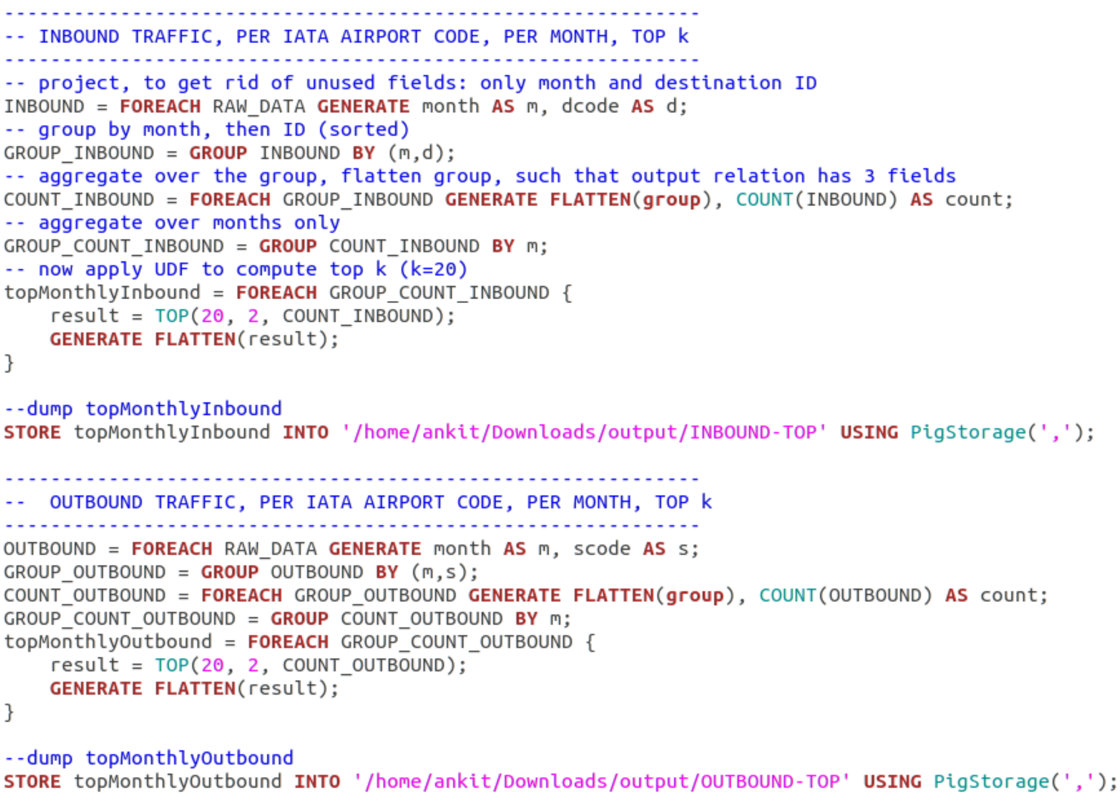
2. Analysis of Flight Data using Apache PIG on Hadoop

Query 1: Top 20 cities by total volume of flights

What are the busiest cities by total flight traffic? For each airport code compute the number of inbound, outbound and all flights.

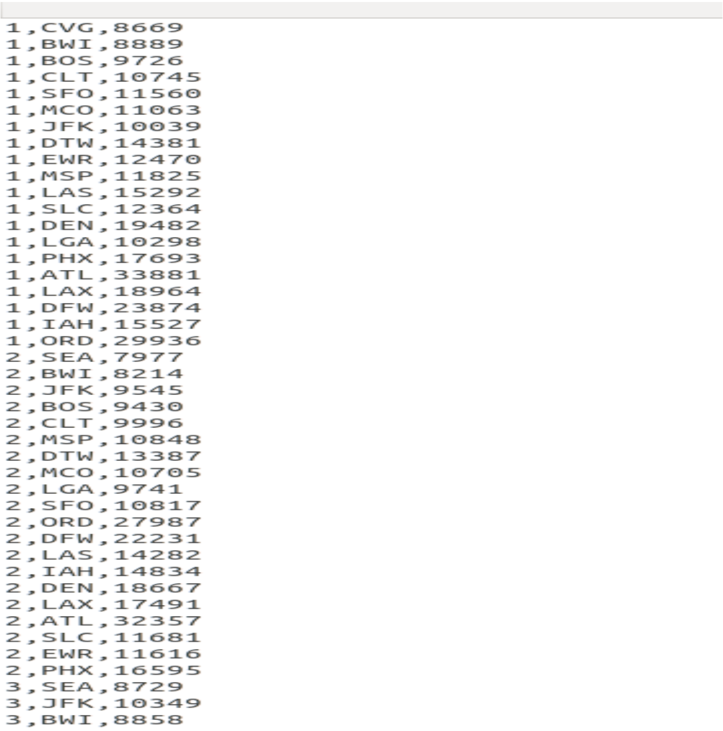


```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/flight-data' USING PigStorage(',') AS
        (year: int, month: int, day: int, dow: int,
        dtime: int, sdtime: int, arrtime: int, satime: int,
        carrier: chararray, fn: int, tn: chararray,
        etime: int, setime: int, airtime: int,
        adelay: int, ddelay: int,
        scode: chararray, dcode: chararray, dist: int,
        tintime: int, touttime: int,
        cancel: chararray, cancelcode: chararray, diverted: int,
        cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);
```



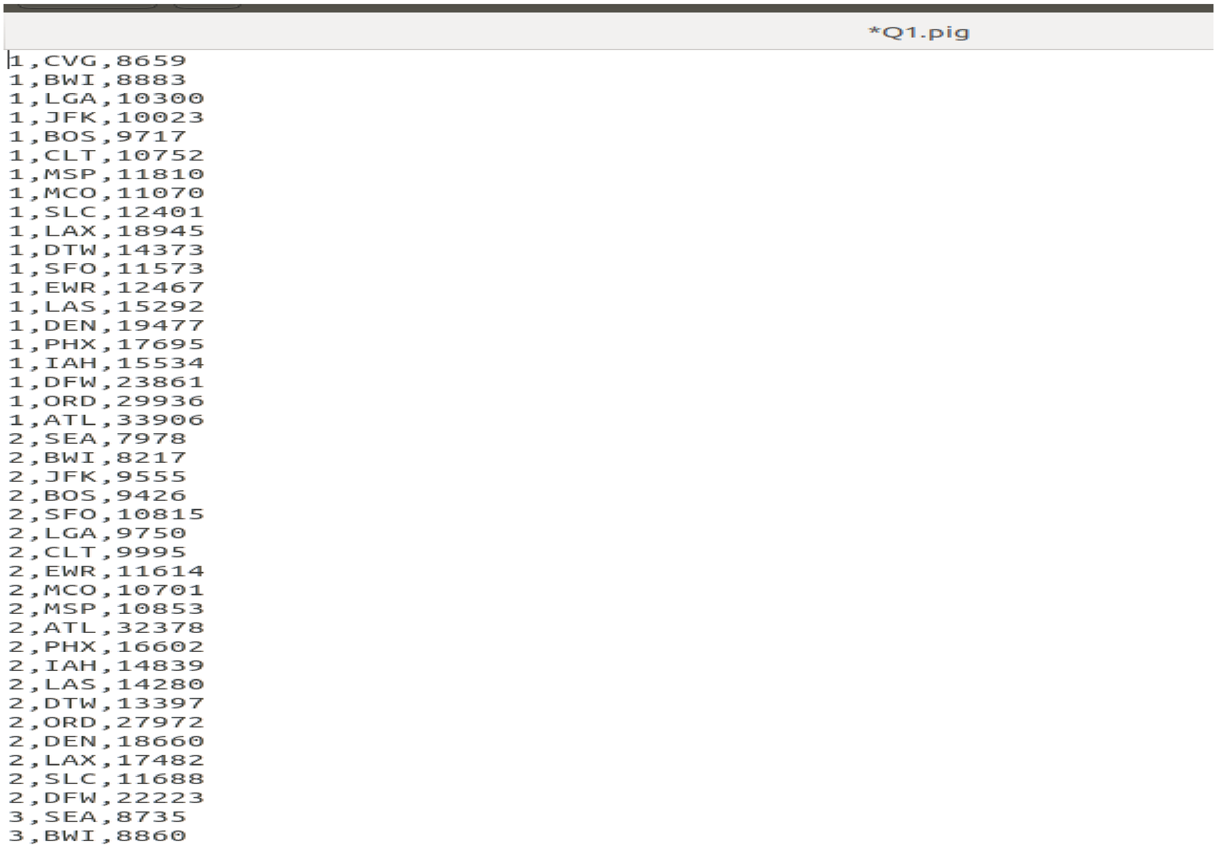
```
------------------------------------------------------------
-- INBOUND TRAFFIC, PER IATA AIRPORT CODE, PER MONTH, TOP k
------------------------------------------------------------
-- project, to get rid of unused fields: only month and destination ID
INBOUND = FOREACH RAW_DATA GENERATE month AS m, dcode AS d;
-- group by month, then ID (sorted)
GROUP_INBOUND = GROUP INBOUND BY (m,d);
-- aggregate over the group, flatten group, such that output relation has 3 fields
COUNT_INBOUND = FOREACH GROUP_INBOUND GENERATE FLATTEN(group), COUNT(INBOUND) AS count;
-- aggregate over months only
GROUP_COUNT_INBOUND = GROUP COUNT_INBOUND BY m;
-- now apply UDF to compute top k (k=20)
topMonthlyInbound = FOREACH GROUP_COUNT_INBOUND {
    result = TOP(20, 2, COUNT_INBOUND);
    GENERATE FLATTEN(result);
}

--dump topMonthlyInbound
STORE topMonthlyInbound INTO '/home/ankit/Downloads/output/INBOUND-TOP' USING PigStorage(',');

------------------------------------------------------------
--  OUTBOUND TRAFFIC, PER IATA AIRPORT CODE, PER MONTH, TOP k
------------------------------------------------------------
OUTBOUND = FOREACH RAW_DATA GENERATE month AS m, scode AS s;
GROUP_OUTBOUND = GROUP OUTBOUND BY (m,s);
COUNT_OUTBOUND = FOREACH GROUP_OUTBOUND GENERATE FLATTEN(group), COUNT(OUTBOUND) AS count;
GROUP_COUNT_OUTBOUND = GROUP COUNT_OUTBOUND BY m;
topMonthlyOutbound = FOREACH GROUP_COUNT_OUTBOUND {
    result = TOP(20, 2, COUNT_OUTBOUND);
    GENERATE FLATTEN(result);
}

--dump topMonthlyOutbound
STORE topMonthlyOutbound INTO '/home/ankit/Downloads/output/OUTBOUND-TOP' USING PigStorage(',');
```

OUTPUT: INBOUNT_TOP



```
1,CVG,8669
1,BWI,8889
1,BOS,9726
1,CLT,10745
1,SFO,11560
1,MCO,11063
1,JFK,10039
1,DTW,14381
1,EWR,12470
1,MSP,11825
1,LAS,15292
1,SLC,12364
1,DEN,19482
1,LGA,10298
1,PHX,17693
1,ATL,33881
1,LAX,18964
1,DFW,23874
1,IAH,15527
1,ORD,29936
2,SEA,7977
2,BWI,8214
2,JFK,9545
2,BOS,9430
2,CLT,9996
2,MSP,10848
2,DTW,13387
2,MCO,10705
2,LGA,9741
2,SFO,10817
2,ORD,27987
2,DFW,22231
2,LAS,14282
2,IAH,14834
2,DEN,18667
2,LAX,17491
2,ATL,32357
2,SLC,11681
2,EWR,11616
2,PHX,16595
3,SEA,8729
3,JFK,10349
3,BWI,8858
```

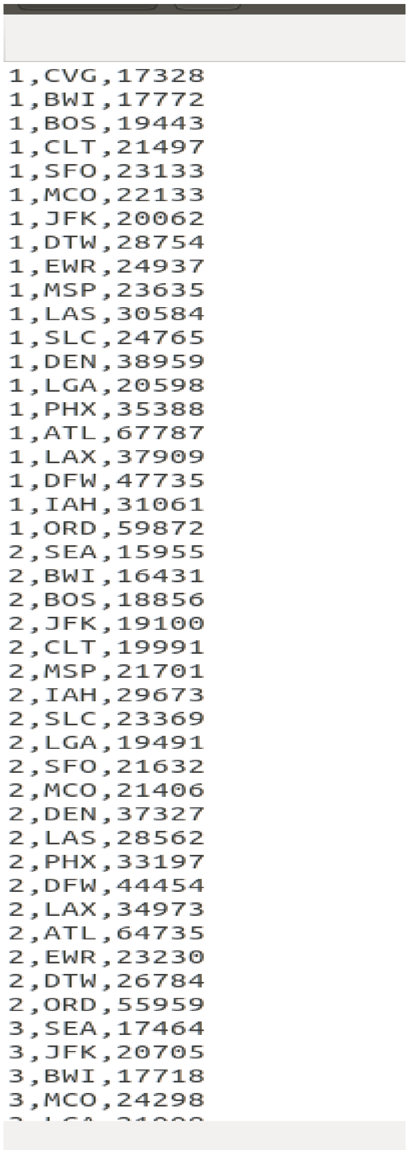
OUTPUT: OUTBOUND TOP



```
1,CVG,8659
1,BWI,8883
1,LGA,10300
1,JFK,10023
1,BOS,9717
1,CLT,10752
1,MSP,11810
1,MCO,11070
1,SLC,12401
1,LAX,18945
1,DTW,14373
1,SFO,11573
1,EWR,12467
1,LAS,15292
1,DEN,19477
1,PHX,17695
1,IAH,15534
1,DFW,23861
1,ORD,29936
1,ATL,33906
2,SEA,7978
2,BWI,8217
2,JFK,9555
2,BOS,9426
2,SFO,10815
2,LGA,9750
2,CLT,9995
2,EWR,11614
2,MCO,10701
2,MSP,10853
2,ATL,32378
2,PHX,16602
2,IAH,14839
2,LAS,14280
2,DTW,13397
2,ORD,27972
2,DEN,18660
2,LAX,17482
2,SLC,11688
2,DFW,22223
3,SEA,8735
3,BWI,8860
```

```
------------------------------------------------------------
-- TOTAL TRAFFIC, PER IATA AIRPORT CODE, PER MONTH, TOP k
------------------------------------------------------------
UNION_TRAFFIC = UNION COUNT_INBOUND, COUNT_OUTBOUND;
GROUP_UNION_TRAFFIC = GROUP UNION_TRAFFIC BY (m,d);
TOTAL_TRAFFIC = FOREACH GROUP_UNION_TRAFFIC GENERATE FLATTEN(group) AS (m,code), SUM(UNION_TRAFFIC.count) AS total;
TOTAL_MONTHLY = GROUP TOTAL_TRAFFIC BY m;

topMonthlyTraffic = FOREACH TOTAL_MONTHLY {
    result = TOP(20, 2, TOTAL_TRAFFIC);
    GENERATE FLATTEN(result) AS (month, iata, traffic);
}


STORE topMonthlyTraffic INTO '/home/ankit/Downloads/output/MONTHLY-TRAFFIC-TOP/' USING PigStorage(',');

explain -brief -dot -out ./ topMonthlyTraffic
```

Output Monthly Traffic Top

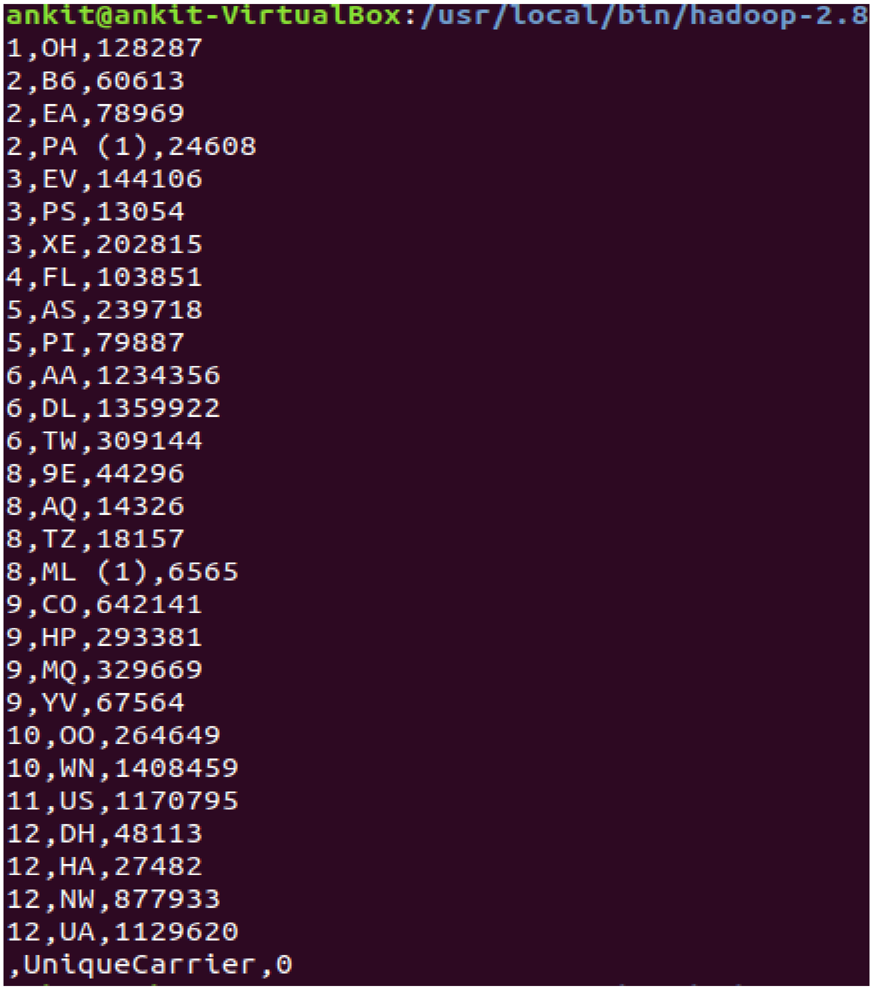


```
1,CVG,17328
1,BWI,17772
1,BOS,19443
1,CLT,21497
1,SFO,23133
1,MCO,22133
1,JFK,20062
1,DTW,28754
1,EWR,24937
1,MSP,23635
1,LAS,30584
1,SLC,24765
1,DEN,38959
1,LGA,20598
1,PHX,35388
1,ATL,67787
1,LAX,37909
1,DFW,47735
1,IAH,31061
1,ORD,59872
2,SEA,15955
2,BWI,16431
2,BOS,18856
2,JFK,19100
2,CLT,19991
2,MSP,21701
2,IAH,29673
2,SLC,23369
2,LGA,19491
2,SFO,21632
2,MCO,21406
2,DEN,37327
2,LAS,28562
2,PHX,33197
2,DFW,44454
2,LAX,34973
2,ATL,64735
2,EWR,23230
2,DTW,26784
2,ORD,55959
3,SEA,17464
3,JFK,20705
3,BWI,17718
3,MCO,24298
```

Query 2: Carrier Popularity

Computing the volume -- total flights -- over each year, by carrier. The carriers are ranked by their median volume (over the 10-year span).

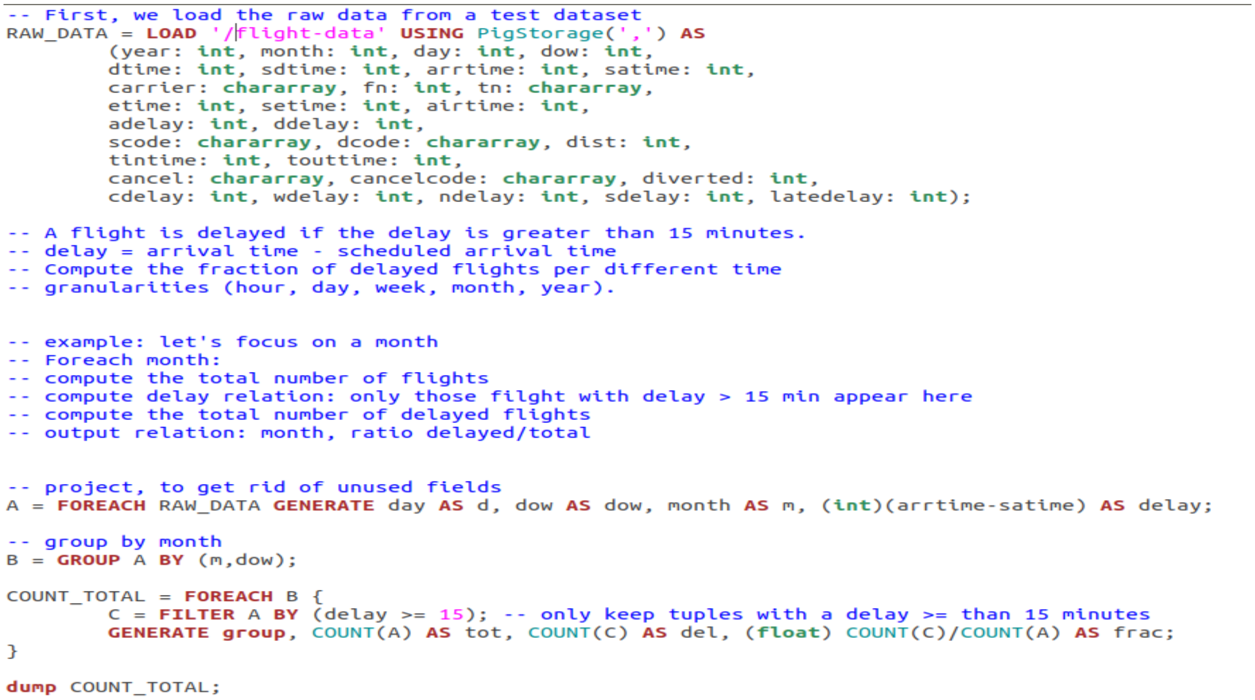
```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/flight-data' USING PigStorage(',') AS
        (year: int, month: int, day: int, dow: int,
        dtime: int, sdtime: int, arrtime: int, satime: int,
        carrier: chararray, fn: int, tn: chararray,
        etime: int, setime: int, airtime: int,
        adelay: int, ddelay: int,
        scode: chararray, dcode: chararray, dist: int,
        tintime: int, touttime: int,
        cancel: chararray, cancelcode: chararray, diverted: int,
        cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

CARRIER_DATA = FOREACH RAW_DATA GENERATE month AS m, carrier AS cname;
GROUP_CARRIERS = GROUP CARRIER_DATA BY (m,cname);
COUNT_CARRIERS = FOREACH GROUP_CARRIERS GENERATE FLATTEN(group), (COUNT(CARRIER_DATA)) AS popularity;

STORE COUNT_CARRIERS INTO '/PIG-OUTPUT-FULL/Q2/COUNT_CARRIERS' USING PigStorage(',');
--dump COUNT_CARRIERS
```



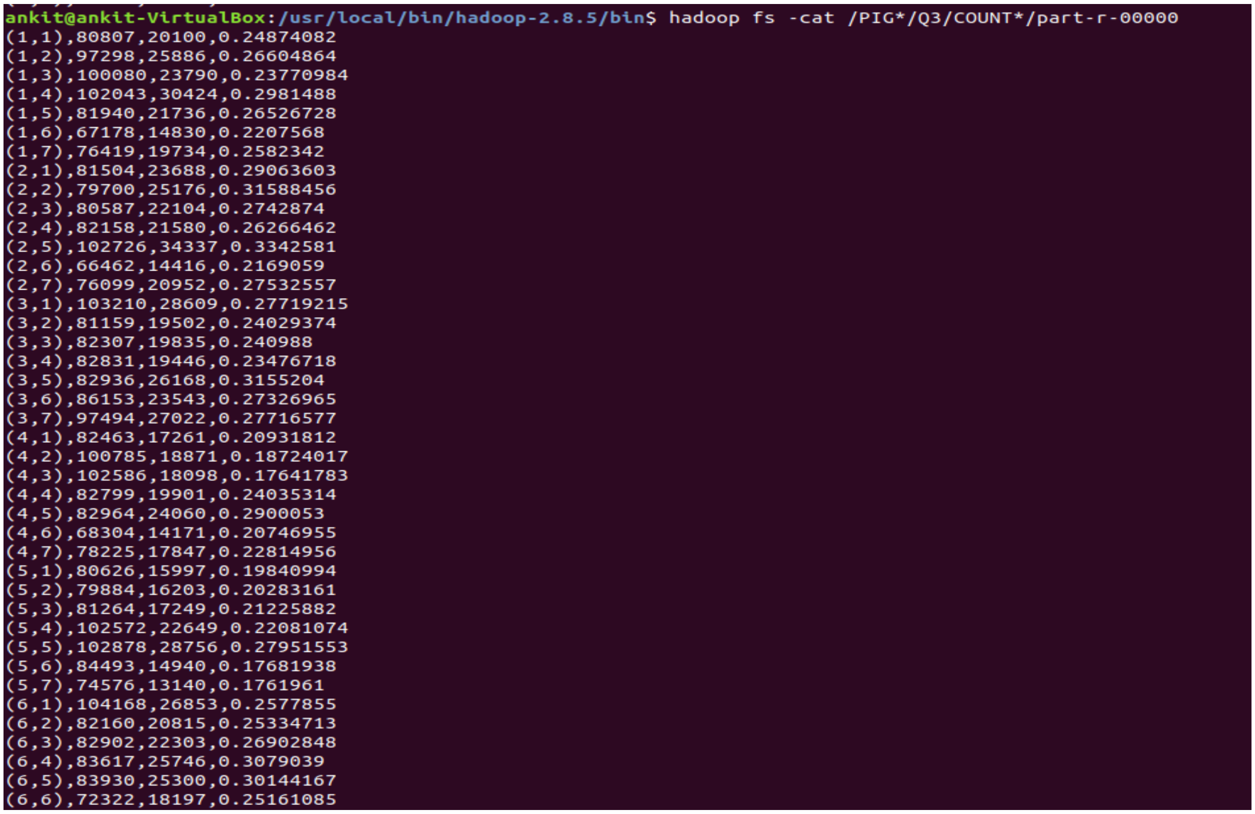
```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8
1,OH,128287
2,B6,60613
2,EA,78969
2,PA (1),24608
3,EV,144106
3,PS,13054
3,XE,202815
4,FL,103851
5,AS,239718
5,PI,79887
6,AA,1234356
6,DL,1359922
6,TW,309144
8,9E,44296
8,AQ,14326
8,TZ,18157
8,ML (1),6565
9,CO,642141
9,HP,293381
9,MQ,329669
9,YV,67564
10,OO,264649
10,WN,1408459
11,US,1170795
12,DH,48113
12,HA,27482
12,NW,877933
12,UA,1129620
,UniqueCarrier,0
```

Query 3: Proportion of Flights Delayed

A flight is delayed if the delay is greater than 15 minutes. I am calculating the fraction of delayed flights per different time limits (hour, day, week, month, year).



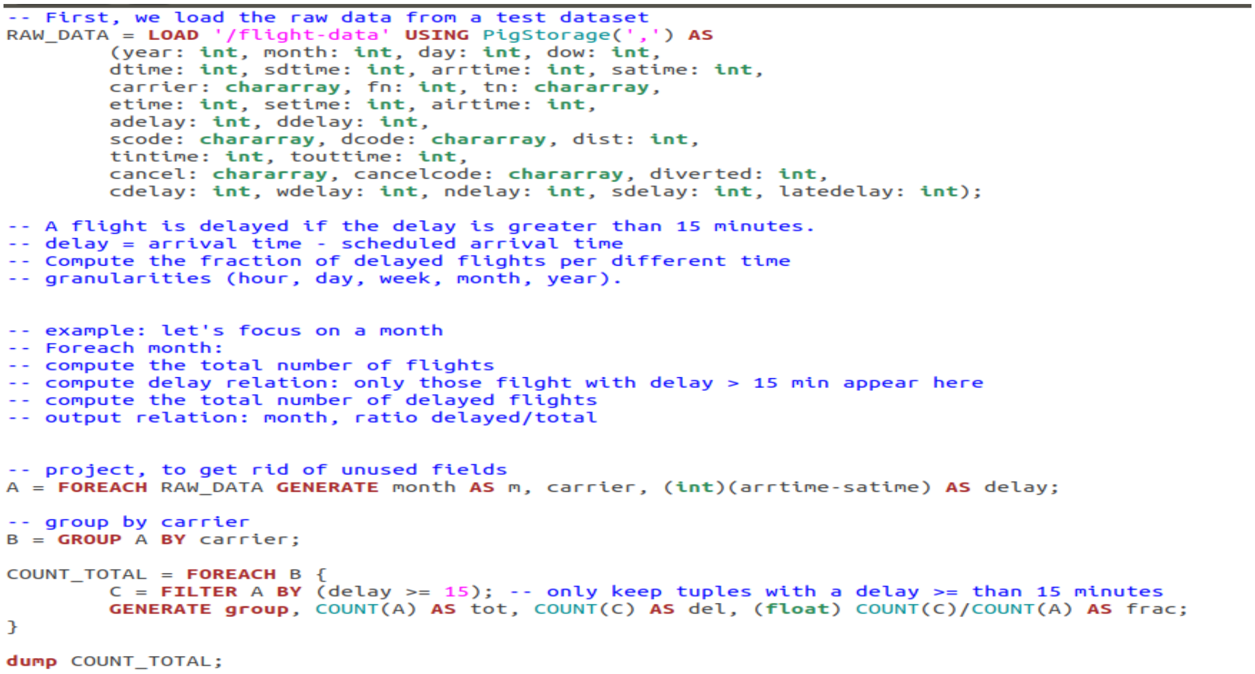
```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/flight-data' USING PigStorage(',') AS
        (year: int, month: int, day: int, dow: int,
        dtime: int, sdtime: int, arrtime: int, satime: int,
        carrier: chararray, fn: int, tn: chararray,
        etime: int, setime: int, airtime: int,
        adelay: int, ddelay: int,
        scode: chararray, dcode: chararray, dist: int,
        tintime: int, touttime: int,
        cancel: chararray, cancelcode: chararray, diverted: int,
        cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

-- A flight is delayed if the delay is greater than 15 minutes.
-- delay = arrival time - scheduled arrival time
-- Compute the fraction of delayed flights per different time
-- granularities (hour, day, week, month, year).


-- example: let's focus on a month
-- Foreach month:
-- compute the total number of flights
-- compute delay relation: only those filght with delay > 15 min appear here
-- compute the total number of delayed flights
-- output relation: month, ratio delayed/total


-- project, to get rid of unused fields
A = FOREACH RAW_DATA GENERATE day AS d, dow AS dow, month AS m, (int)(arrtime-satime) AS delay;

-- group by month
B = GROUP A BY (m,dow);

COUNT_TOTAL = FOREACH B {
        C = FILTER A BY (delay >= 15); -- only keep tuples with a delay >= than 15 minutes
        GENERATE group, COUNT(A) AS tot, COUNT(C) AS del, (float) COUNT(C)/COUNT(A) AS frac;
}

dump COUNT_TOTAL;
```



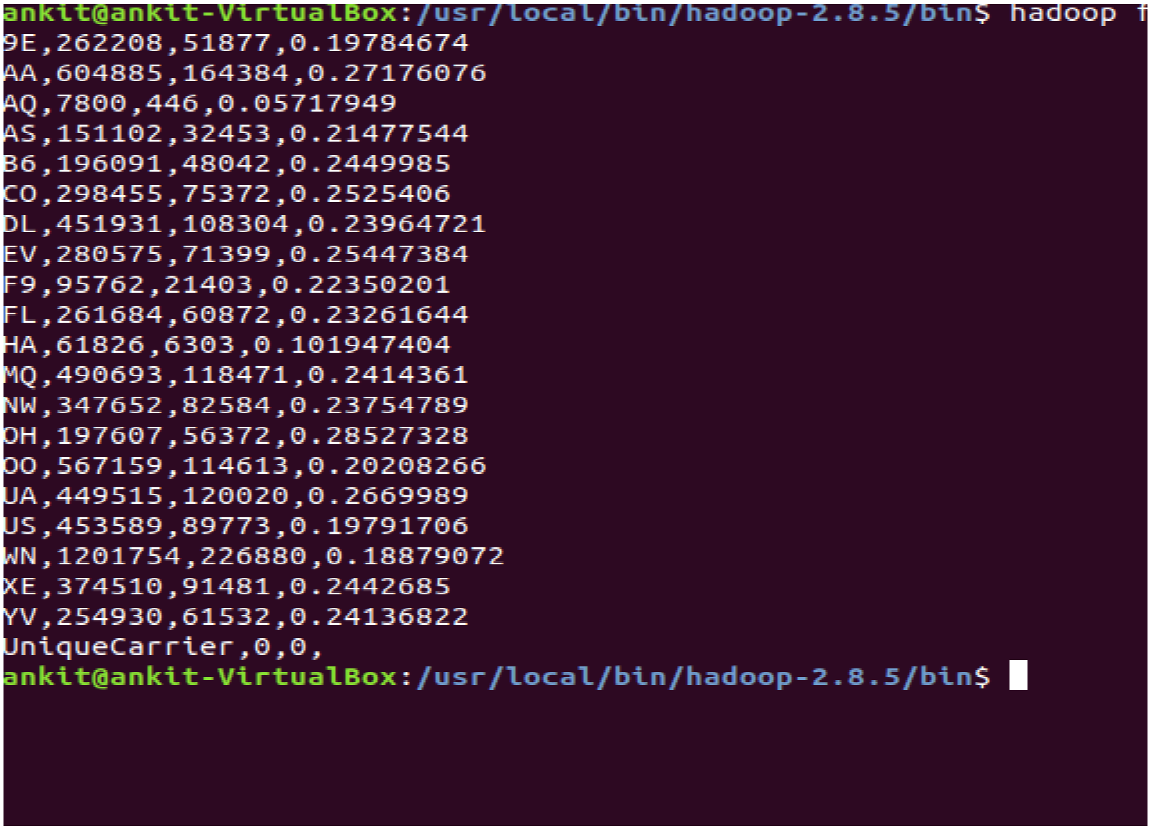
```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /PIG*/Q3/COUNT*/part-r-00000
(1,1),80807,20100,0.24874082
(1,2),97298,25886,0.26604864
(1,3),100080,23790,0.23770984
(1,4),102043,30424,0.2981488
(1,5),81940,21736,0.26526728
(1,6),67178,14830,0.2207568
(1,7),76419,19734,0.2582342
(2,1),81504,23688,0.29063603
(2,2),79700,25176,0.31588456
(2,3),80587,22104,0.2742874
(2,4),82158,21580,0.26266462
(2,5),102726,34337,0.3342581
(2,6),66462,14416,0.2169059
(2,7),76099,20952,0.27532557
(3,1),103210,28609,0.27719215
(3,2),81159,19502,0.24029374
(3,3),82307,19835,0.240988
(3,4),82831,19446,0.23476718
(3,5),82936,26168,0.3155204
(3,6),86153,23543,0.27326965
(3,7),97494,27022,0.27716577
(4,1),82463,17261,0.20931812
(4,2),100785,18871,0.18724017
(4,3),102586,18098,0.17641783
(4,4),82799,19901,0.24035314
(4,5),82964,24060,0.2900053
(4,6),68304,14171,0.20746955
(4,7),78225,17847,0.22814956
(5,1),80626,15997,0.19840994
(5,2),79884,16203,0.20283161
(5,3),81264,17249,0.21225882
(5,4),102572,22649,0.22081074
(5,5),102878,28756,0.27951553
(5,6),84493,14940,0.17681938
(5,7),74576,13140,0.1761961
(6,1),104168,26853,0.2577855
(6,2),82160,20815,0.25334713
(6,3),82902,22303,0.26902848
(6,4),83617,25746,0.3079039
(6,5),83930,25300,0.30144167
(6,6),72322,18197,0.25161085
```

Query 4: Carrier Delays

Calculating the proportion of delayed flights by carrier, ranked by carrier, at different time (hour, day, week, month year). Again, a flight is delayed if the delay is greater than 15 minutes.



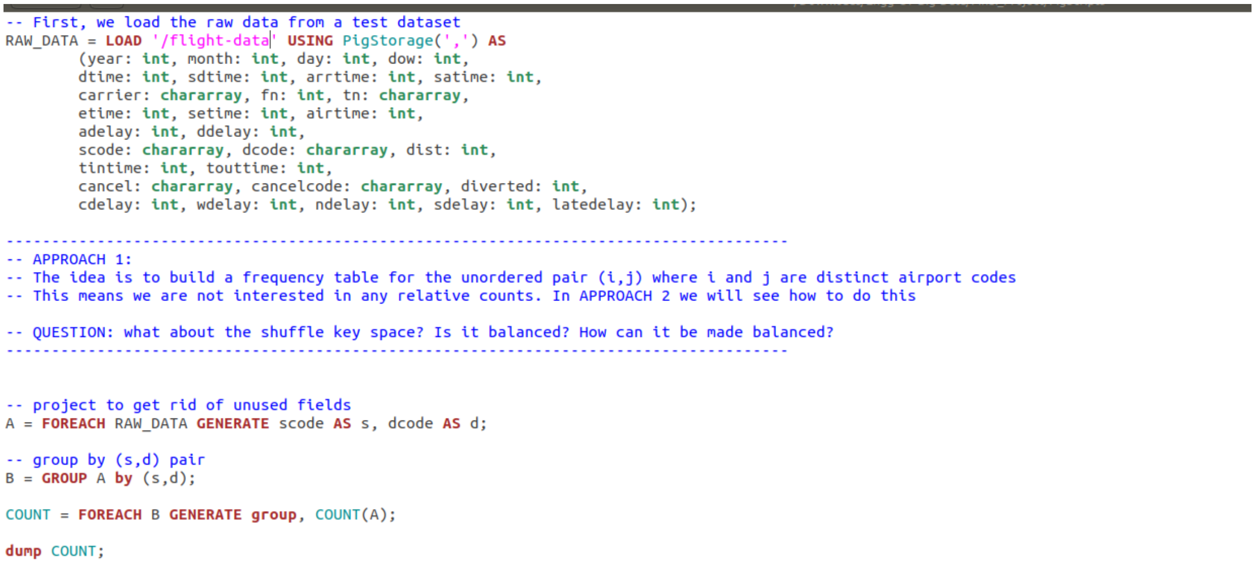
```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/flight-data' USING PigStorage(',') AS
        (year: int, month: int, day: int, dow: int,
        dtime: int, sdtime: int, arrtime: int, satime: int,
        carrier: chararray, fn: int, tn: chararray,
        etime: int, setime: int, airtime: int,
        adelay: int, ddelay: int,
        scode: chararray, dcode: chararray, dist: int,
        tintime: int, touttime: int,
        cancel: chararray, cancelcode: chararray, diverted: int,
        cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

-- A flight is delayed if the delay is greater than 15 minutes.
-- delay = arrival time - scheduled arrival time
-- Compute the fraction of delayed flights per different time
-- granularities (hour, day, week, month, year).


-- example: let's focus on a month
-- Foreach month:
-- compute the total number of flights
-- compute delay relation: only those filght with delay > 15 min appear here
-- compute the total number of delayed flights
-- output relation: month, ratio delayed/total


-- project, to get rid of unused fields
A = FOREACH RAW_DATA GENERATE month AS m, carrier, (int)(arrtime-satime) AS delay;

-- group by carrier
B = GROUP A BY carrier;

COUNT_TOTAL = FOREACH B {
        C = FILTER A BY (delay >= 15); -- only keep tuples with a delay >= than 15 minutes
        GENERATE group, COUNT(A) AS tot, COUNT(C) AS del, (float) COUNT(C)/COUNT(A) AS frac;
}

dump COUNT_TOTAL;
```



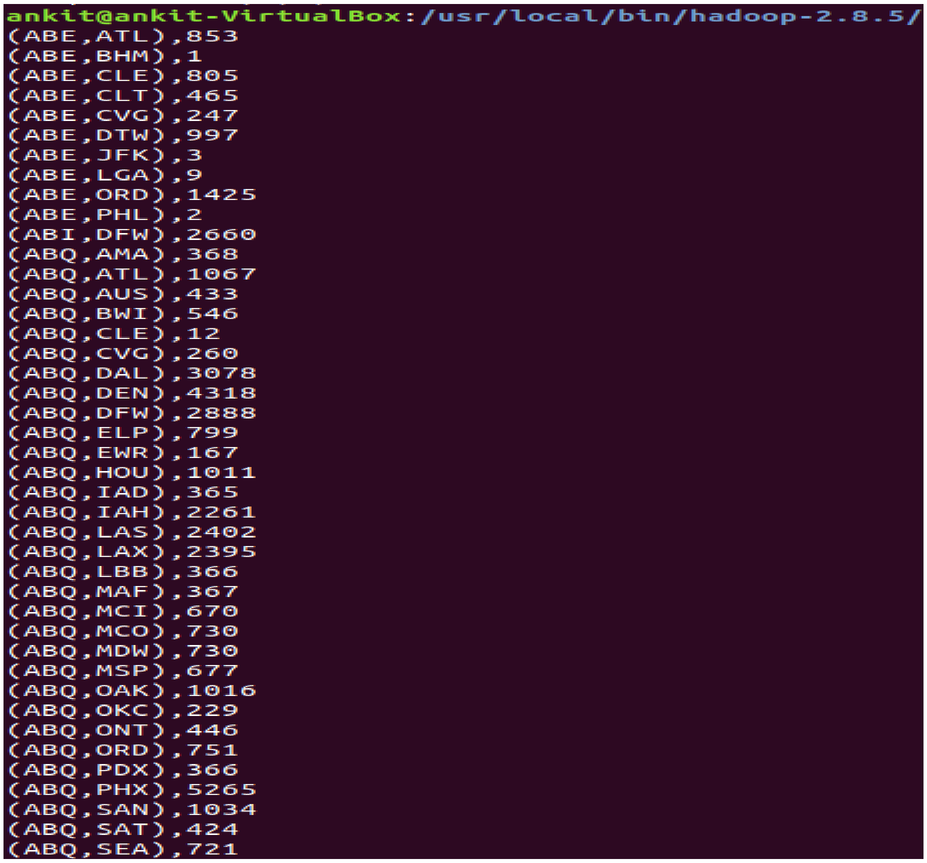
```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop f
9E,262208,51877,0.19784674
AA,604885,164384,0.27176076
AQ,7800,446,0.05717949
AS,151102,32453,0.21477544
B6,196091,48042,0.2449985
CO,298455,75372,0.2525406
DL,451931,108304,0.23964721
EV,280575,71399,0.25447384
F9,95762,21403,0.22350201
FL,261684,60872,0.23261644
HA,61826,6303,0.101947404
MQ,490693,118471,0.2414361
NW,347652,82584,0.23754789
OH,197607,56372,0.28527328
OO,567159,114613,0.20208266
UA,449515,120020,0.2669989
US,453589,89773,0.19791706
WN,1201754,226880,0.18879072
XE,374510,91481,0.2442685
YV,254930,61532,0.24136822
UniqueCarrier,0,0,
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

Query 5: Busy Routes

Which are busy the routes? A simple first approach is to create a frequency table for the unordered pair (i,j) where i and j are distinct airport codes.



```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/flight-data' USING PigStorage(',') AS
        (year: int, month: int, day: int, dow: int,
        dtime: int, sdtime: int, arrtime: int, satime: int,
        carrier: chararray, fn: int, tn: chararray,
        etime: int, setime: int, airtime: int,
        adelay: int, ddelay: int,
        scode: chararray, dcode: chararray, dist: int,
        tintime: int, touttime: int,
        cancel: chararray, cancelcode: chararray, diverted: int,
        cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

-------------------------------------------------------------------------------------
-- APPROACH 1:
-- The idea is to build a frequency table for the unordered pair (i,j) where i and j are distinct airport codes
-- This means we are not interested in any relative counts. In APPROACH 2 we will see how to do this

-- QUESTION: what about the shuffle key space? Is it balanced? How can it be made balanced?
-------------------------------------------------------------------------------------


-- project to get rid of unused fields
A = FOREACH RAW_DATA GENERATE scode AS s, dcode AS d;

-- group by (s,d) pair
B = GROUP A by (s,d);

COUNT = FOREACH B GENERATE group, COUNT(A);

dump COUNT;
```



```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/
(ABE,ATL),853
(ABE,BHM),1
(ABE,CLE),805
(ABE,CLT),465
(ABE,CVG),247
(ABE,DTW),997
(ABE,JFK),3
(ABE,LGA),9
(ABE,ORD),1425
(ABE,PHL),2
(ABI,DFW),2660
(ABQ,AMA),368
(ABQ,ATL),1067
(ABQ,AUS),433
(ABQ,BWI),546
(ABQ,CLE),12
(ABQ,CVG),260
(ABQ,DAL),3078
(ABQ,DEN),4318
(ABQ,DFW),2888
(ABQ,ELP),799
(ABQ,EWR),167
(ABQ,HOU),1011
(ABQ,IAD),365
(ABQ,IAH),2261
(ABQ,LAS),2402
(ABQ,LAX),2395
(ABQ,LBB),366
(ABQ,MAF),367
(ABQ,MCI),670
(ABQ,MCO),730
(ABQ,MDW),730
(ABQ,MSP),677
(ABQ,OAK),1016
(ABQ,OKC),229
(ABQ,ONT),446
(ABQ,ORD),751
(ABQ,PDX),366
(ABQ,PHX),5265
(ABQ,SAN),1034
(ABQ,SAT),424
(ABQ,SEA),721
```

3. References

1. https://learning.oreilly.com/library/view/mapreduce-design-patterns/9781449341954/
2. https://gitlab.eurecom.fr/yonghui.feng/clouds-lab
3. https://learning.oreilly.com/library/view/data-algorithms/9781491906170/ch01.html
4. http://cs229.stanford.edu/proj2013/MathurNagaoNg-PredictingFlightOnTimePerformance.pdf

4. APPENDIX

The code of this project can be found at GitHub repository for this project at

https://github.com/ankit08015/Engg-Of-Big-Data/tree/master/Final_Project