



# FINAL PROJECT

INFO7250- Engg Of Big Data

Ankit Yadav

NUID- 001271369

Northeastern University

## **CONTENTS**

## 1. Analysis of Flight Data using MAP REDUCE 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 /FinalProjectMROutput
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.

Delayed flight Count:

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 than 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 /FinalProjectMR/output/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: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMR/output/8-Flight-Data-by-Year
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: /usr/local/bin/hadoop-2.8.5/bin$
```

## 8- Delayed flights per year

In this we will check delayed flights per year( we will count flights as delayed only if the delay time is greater than or 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

```
bytesWritten=550
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /FinalProjectMROutput/13-DelayedFlightsByMonth.txt
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
6-June         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_Ratio
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-Cancelled
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: 2004
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 /FinalProjectMR/output/16/part-r-00000
ABE: BDL AZO AVP ATL SBN ROC RDU PIT PHL ORD MDT ALB MCO LGA JFK IND IAD HPM GRR FMA DTW DCA CVG CLT CLE BWI BHM
ABI: SJT LAX IAH DFW CLL VCT TYR
ABQ: STL TUL TUS TMF TPA AMA ATL AUS BNA BMT CLE COS CVG DAL DEN DFW ELP ENR 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: MCM ATL VLD
ACK: LGA ENR JFK
ACT: SJT CLL TYR DFW IAH ILE
ACV: CEC CLC MFR MRY RDU SFO SJC SMF SLC
ACY: MCO LGA JFK CVG BMT BOS ATL PIT MYR
ADK: ANC AKN
ADQ: BET ANC
AEX: ILE JAN LIT MLU MSY SHV ATL AUS BTR DFW GTR HOU IAH ABI
AGS: CAE ATL SAV LGA JFK IAH ENR CVG CLT CLE CHS CHA
AKN: ANC ADK DLG
ALB: MCO ATL BDL BGR BOS BTM BUF BMT CLE CLT CVG DCA DTW ENR FLL GRR IAD ISP JFK LAS LGA MDW MHT MKE MSP ORD PHL PIT PVD PWM RDU ROC SBN SMF 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 ENR FAI HNL JNU KOA LAS LAX MSP OGG OME ORD OTZ PDX PHX SCC SEA SFO SIT SLC STL YAK
ANI: KSM ANC
ATL: MIA TPA
APP: ATL MIA TPA
ASE: GJT ATL DEN LAX HSN ORD PHX RFD SFO SLC
ATL: HON ABE ABQ ABY ACY AEX AGS ALB ANC APP ASE ATM AUS AVL AVP AZO BDL BGR BHM BMT BNA BOI BOS BPT BQK BQN BTR BTM 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 DNH DSM DTW EGE ELP ERI EVV ENR ENR EYW FAY FCA FLL FLO FNT FSD FSM FMA GNV GPT GRB GRK GRR GSO GSP GTR GUC HHK HKY HNL HOU HPM HSV HTS IAD IAH ICT ILG ILM IND ISO ISP JAC JAN JAX JF
K LAN LAS LAM LAX LEB LFT LGA LGB LIT LNB LYM MCI MCM MCO MDT MDW MEI MEM MFE MGH 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 REC RND ROK ROR RSM SAN SAT SAV SBN SCE SDF SEA SFO SGF SHV SJC SJU SLC SMF SNA SOP SRQ STL STT STX SMF SYR TLM TOL TPA TRI TTN TUL TUP TUS TVC VLD VPS VNA
ATH: DTW ATL GRB GRK LEX MKE MSP ORD XNA CHS CVG CMA DSM
AUS: PIH ABQ ATL BHM BNA BOS BTR BMT CID CLE CLT CMH COS CRP CVG DAL DEN DFW DSM DTW ELP ENR FLL GJT GPT HOU HRL IAD IAH IND JAX JFK LAS LAX LBB LGB MAF MCI MCM MCO MDW MEM MFE MIA MSP MSY OAK OKC ONT O
RD ORF PHL PHX PIA PIT PWM RDU RFD RND SAN SAT SBN SEA SFO SHV SJC SLC SNA SPI STL TPA TUL TUS
AVL: ATL BMT CLT CVG DTW ENR IAH LEX LGA AGS MCO MSP PIT TRI
AVP: CVG ABE ALB ATL AZO BDL BGR BUF BMT CLT DTW ELM HPM JFK LAN LGA MDT ORD PHL PIT
AZO: ATL CVG DAY DTW FAR GRB LAN MKE MSP ORD PIT SBN XNA
BDL: BMT ALB ATL BNA BOS BTM BUF CLE CLT CMH CVG DAY DCA DEN DFW DTW ENR FLL GRB HPM IAD IAH IND ISP JFK LAS LAX LGA MCO MDW MEM MHT MIA MKE MSP ORD PBI PHL PHX PIT PVD PWM RDU ROC RSM SAT SFO SJU S
LC STL SMF SYR TPA
BET: ANI ANC ARA 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 -c
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- Hierarchical data for all source, destination and carrier information