FINAL PROJECT REPORT

INFO 7250: ENGG. OF BIG DATA

REEMA DUTTA
NEU ID: 001890899

Content

1. Introduction of Data Set

For this project the data on Airline On-Time Statistics and Delay Causes is exported 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.

2. Map Reduce Algorithms

1. Total Count of all the data in the dataset

Total data- 123534991

```
cat: /FinalProjectOutput/MROutput/1-TotalCount/part-r-0000 : No such file or directory
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/1-TotalCount/part-r-00000
19/12/11 00:34:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
123534991
Reemas-MacBook-Pro:bin reemadutta$
```

2. Total number of flights for each source and destination

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/2.1-SrcDestCount/part-r-00000|head
19/12/11 00:41:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
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
```

3. Top 25 source and destination based on flight count

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/2.2-Top25SrcDest/part-r-00000
19/12/11 00:50:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
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
Reemas-MacBook-Pro:bin reemadutta$
```

4. Number of all flights grouped by flight carriers

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/3.1-UniqueCarCount/part-r-00000
19/12/11 01:04:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
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
```

5. Inner join to get the carrier full name from carrier's code

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/3.2-InnerJoinUniqueCarCount/part-r-00000
19/12/11 01:07:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
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
Reemas-MacBook-Pro:bin reemadutta$
```

7. Number of flights per year

This count shows that number of flights per year gradually increased over time from 1987 to 2008 with maximum flights in 2007 and minimum in 1987 with steep increase in 1988 and 2003 and then gradual increases.

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/4-FlightsPerYear/part-r-00000
19/12/11 01:13:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
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
```

8. Percentage delay of flight per year

This analysis shows that most delayed flights were in year 1987 and 1996, while least delay was observed from 1991 to 1993

```
[Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/5-FlightsDelayPerYear/part-r-00000|head
19/12/11 01:20:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
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
```

9. Percentage delay of flight calculated day wise

This analysis shows that best days to fly are Tuesday and Saturday, while worst days are Friday and Thursday.
This may be due to weekend starting during those days.

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/6-FlightsDelayPerDay/part-r-00000|head
19/12/11 01:35:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
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
Reemas-MacBook-Pro:bin reemadutta$
```

10. Percentage delay of flight calculated month wise

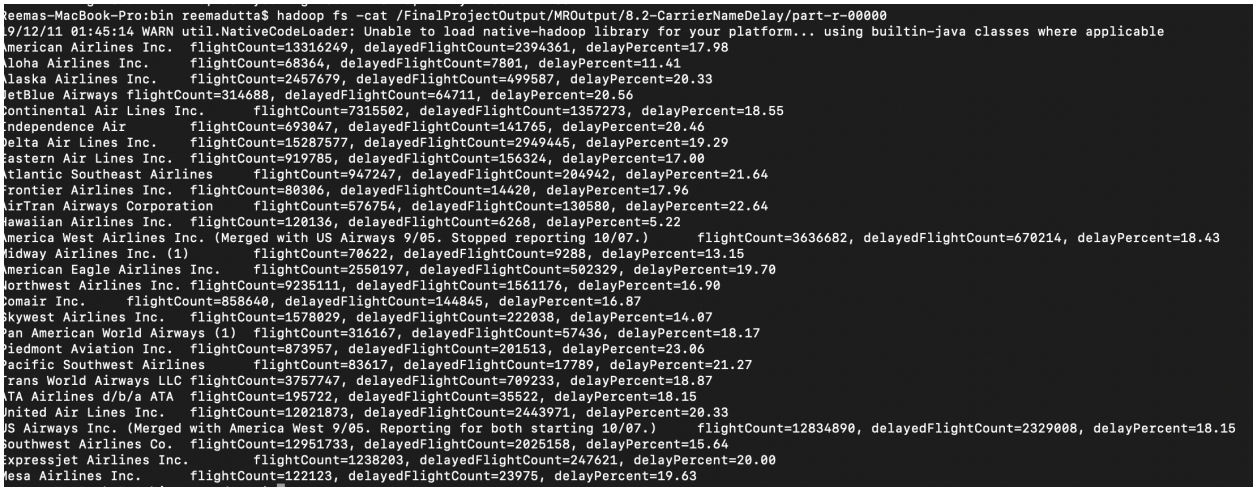
Most flights are delayed during the month of December – 24%. This may be because December is a holiday season.
While months with least delay are September, October, April and May.

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/7-FlightsDelayPerMonth/part-r-00000
19/12/11 01:37:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
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
```

11. Percentage delay of flight calculated carrier wise

Least delay was observed on Hawaiian Airlines- 5%.
Also, SkyWest airlines and Southwest Airlines had low delay percentage.

Higher delay percentage was observed with Piedmont Aviation Inc, Alaska Airlines and United Airlines.



12. Inverted index to find all destination routes from each source airports

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/9-InvertedIndex/part-r-00000
19/12/11 01:48:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ABE:    ORD DTW GRR BHM HPN CLT BDL JFK FWA BWI PHL MDT SBN RDU DCA LGA ROC IAD CVG AZO ALB MCO ATL PIT AVP CLE IND
ABI:    CLL LAX VCT DFW SJT TYR IAH
ABQ:    ORD SAN LAX TPA SNA SAT IDA ELP STL SLC LBB GJT DFW ONT AMA BWI SFO AUS PHX MDW EWR TWF MKC PDX COS OKC TUL MSP DEN PSP IAD SMF DAL SEA PIH HOU IAH MCI MAF TUS BNA CVG OAK MCO ATL PIT CL
ABY:    VLD ATL MCN
ACK:    JFK EWR LGA
ACT:    CLL ILE DFW SJT TYR IAH
ACV:    SLC SJC RDD MRY MFR CIC SFO CEC SMF
ACY:    JFK BWI LGA BOS MYR CVG MCO ATL PIT
ADK:    AKN ANC
ADQ:    BET ANC
AEX:    BTR ILE GTR ABI DFW MLU LFT AUS SHV HOU IAH MSY JAN ATL
AGS:    SAV CLT JFK CHS EWR CAE LGA IAH CVG ATL CLE CHA
AKN:    ADK ANC DLG
ALB:    ORD TPA BTV DTW GRR FLL CLT BDL JFK BUF PHL BWI MHT SBN EWR MDW RDU DCA SWF MKE BOS PWM ISP LGA ROC MSP BGR IAD SYR CVG MCO ATL PIT CLE LAS PVD
ALO:    STL MSP RST
AMA:    ORD SLC LBB ABQ DFW PHX COS TUL DEN DAL IAH MAF LAS
ANC:    ORD LAX YAK BFI DTW BRW HNL ADK STL SLC ADQ JNU CDV OTZ DFW KOA SFO PHX EWR DUT AKN PDX OME MSP DEN SEA OGG FAI BET IAH SCC CVG ATL ANI DLG LAS SIT
ANI:    KSM ANC
APF:    TPA MIA ATL
ASE:    ORD LAX SLC GJT RFD SFO PHX MSN DEN ATL
ATL:    JAX TVC HNL STL XNA STT SDF PNS DFW LWB STX MHT LFT EYW OMA MIA LGB MYR LGA IAD SEA HOU IAH CVG RSW BTR ABE BTV HPN FLL FLO DHN BDL GNV VLD ABQ BUF ASE ONT ABY LYH SFO BUR RDU SWF ISO MG
SGF LIT ACY ICT CHA MLB ATW ELP GPT CHO MLI FNT JFK CHS BWI PBI MLU AUS CID SHV PSE BGM HSV BGR PSP SYR AEX GRB AVL OAK OAJ AVP ORD GRK ORF BHM GRR HTS SJC TLH HDN VPS SJU MOB GSO PDX GSP RIC AG
PVD FAY GTR CLT SLC GUC FSD PFN CMI CMH FSM FCA PWM SMF TOL AZO ERI TPA SNA HHH PHF PHL BMI PHX DAB COS MSN PIA MSP OGG DAL MCI MSY BNA SOP ALB LAN MCO MCN PIT DAY LAS MTH LAX MTJ LAW ILG TRI IL
MDT BOI DSM MDW DCA ROA CAE ROC BOS CAK MEI MEM ANC HKY IND CRP SAN EVV DTW SAT BPT CRW SAV MFE TTN CBM SRQ CSG BQK EWN BQN SBN EWR EGE OKC TUL DEN JAC TUP MGM TUS SCE JAN APF LEX
ATW:    XNA CVG CWA MKE ATL CHS GRB ORD GRR MSP DSM DTW LEX
AUS:    JAX STL DFW MIA LGB IAD SEA HOU IAH CVG BTR FLL ABQ ONT SFO RDU ATL HRL ELP GPT JFK RFD BWI CID SHV OAK ORD ORF BHM SJC CLE CLT SLC CMH PWM MAF TPA SNA PHL PHX COS PIA MSP DAL PIH MSY MG
MCN PIT LAS LAX LBB SPI RNO DSM MDW BOS MEM IND CRP SAN DTW SAT MFE GJT SBN EWR OKC TUL DEN TUS
```

13. Average flight carrier Distance

Carriers with higher airtime were- Pan American World Airways, Frontier Airlines, Eastern Airlines
Carriers with least airtime were- Continental Airlines, SkyWest Airlines etc.

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/10.1-AvgCarrFlightDist/part-r-00000
19/12/11 01:51:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
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}
Reemas-MacBook-Pro:bin reemadutta$
```

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/MROutput/10.2-AvgCarrFlightDist/part-r-00000
19/12/11 01:53:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
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, Aver
.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}
```

3. Recommendation System

One of the main use case of analyzing this huge data set consisting of airlines data over 20 years was to create a recommendation system which can leverage the observations made taking in account all the variables present in the dataset.

I tried to create a recommendation system taking into account total flights, arrival delay and departure delay. I grouped this data for each carrier between all the source and destination combinations.

While working with this I tried to come up with best metric to compare and analyze this dataset. I tried normal difference, mean and some other statistical method to get a best comparing metric between the source destination for each carrier.

The metric that I used is **Root Mean Square** of the absolute delay difference between each source and destination for all the flights of an individual carrier.

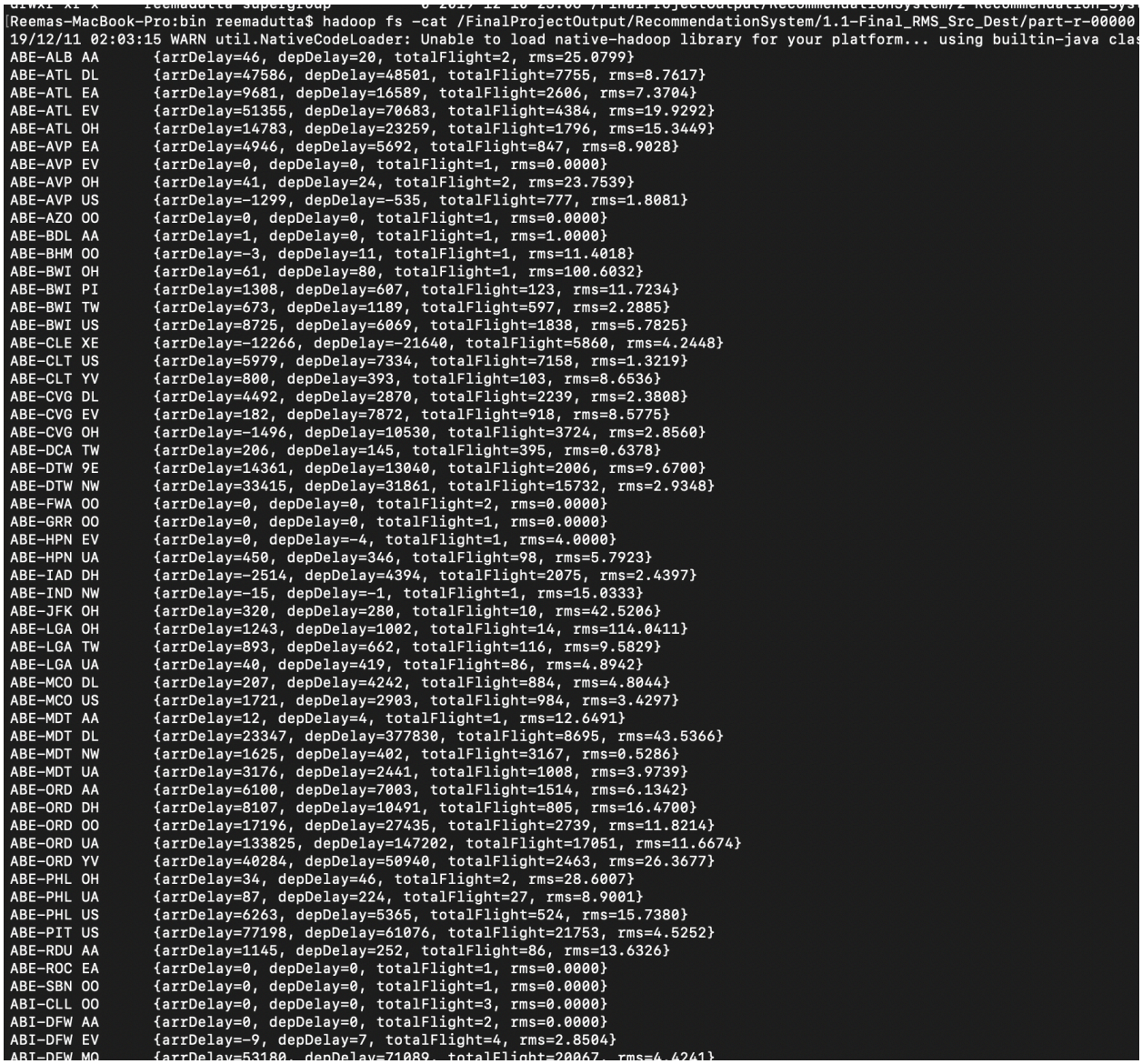
Reason for selecting RMS:

The **root-mean-square deviation (RMSD)** or **root-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences.

RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}.$$

Final RMS between Source and Destination for each carrier



```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/RecommendationSystem/1.1-Final_RMS_Src_Dest/part-r-00000
19/12/11 02:03:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cla
ABE-ALB AA      {arrDelay=46, depDelay=20, totalFlight=2, rms=25.0799}
ABE-ATL DL      {arrDelay=47586, depDelay=48501, totalFlight=7755, rms=8.7617}
ABE-ATL EA      {arrDelay=9681, depDelay=16589, totalFlight=2606, rms=7.3704}
ABE-ATL EV      {arrDelay=51355, depDelay=70683, totalFlight=4384, rms=19.9292}
ABE-ATL OH      {arrDelay=14783, depDelay=23259, totalFlight=1796, rms=15.3449}
ABE-AVP EA      {arrDelay=4946, depDelay=5692, totalFlight=847, rms=8.9028}
ABE-AVP EV      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
ABE-AVP OH      {arrDelay=41, depDelay=24, totalFlight=2, rms=23.7539}
ABE-AVP US      {arrDelay=-1299, depDelay=-535, totalFlight=777, rms=1.8081}
ABE-AZO OO      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
ABE-BDL AA      {arrDelay=1, depDelay=0, totalFlight=1, rms=1.0000}
ABE-BHM OO      {arrDelay=-3, depDelay=11, totalFlight=1, rms=11.4018}
ABE-BWI OH      {arrDelay=61, depDelay=80, totalFlight=1, rms=100.6032}
ABE-BWI PI      {arrDelay=1308, depDelay=607, totalFlight=123, rms=11.7234}
ABE-BWI TW      {arrDelay=673, depDelay=1189, totalFlight=597, rms=2.2885}
ABE-BWI US      {arrDelay=8725, depDelay=6069, totalFlight=1838, rms=5.7825}
ABE-CLE XE      {arrDelay=-12266, depDelay=-21640, totalFlight=5860, rms=4.2448}
ABE-CLT US      {arrDelay=5979, depDelay=7334, totalFlight=7158, rms=1.3219}
ABE-CLT YV      {arrDelay=800, depDelay=393, totalFlight=103, rms=8.6536}
ABE-CVG DL      {arrDelay=4492, depDelay=2870, totalFlight=2239, rms=2.3808}
ABE-CVG EV      {arrDelay=182, depDelay=7872, totalFlight=918, rms=8.5775}
ABE-CVG OH      {arrDelay=-1496, depDelay=10530, totalFlight=3724, rms=2.8560}
ABE-DCA TW      {arrDelay=206, depDelay=145, totalFlight=395, rms=0.6378}
ABE-DTW 9E      {arrDelay=14361, depDelay=13040, totalFlight=2006, rms=9.6700}
ABE-DTW NW      {arrDelay=33415, depDelay=31861, totalFlight=15732, rms=2.9348}
ABE-FWA OO      {arrDelay=0, depDelay=0, totalFlight=2, rms=0.0000}
ABE-GRR OO      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
ABE-HPN EV      {arrDelay=0, depDelay=-4, totalFlight=1, rms=4.0000}
ABE-HPN UA      {arrDelay=450, depDelay=346, totalFlight=98, rms=5.7923}
ABE-IAD DH      {arrDelay=-2514, depDelay=4394, totalFlight=2075, rms=2.4397}
ABE-IND NW      {arrDelay=-15, depDelay=-1, totalFlight=1, rms=15.0333}
ABE-JFK OH      {arrDelay=320, depDelay=280, totalFlight=10, rms=42.5206}
ABE-LGA OH      {arrDelay=1243, depDelay=1002, totalFlight=14, rms=114.0411}
ABE-LGA TW      {arrDelay=893, depDelay=662, totalFlight=116, rms=9.5829}
ABE-LGA UA      {arrDelay=40, depDelay=419, totalFlight=86, rms=4.8942}
ABE-MCO DL      {arrDelay=207, depDelay=4242, totalFlight=884, rms=4.8044}
ABE-MCO US      {arrDelay=1721, depDelay=2903, totalFlight=984, rms=3.4297}
ABE-MDT AA      {arrDelay=12, depDelay=4, totalFlight=1, rms=12.6491}
ABE-MDT DL      {arrDelay=23347, depDelay=377830, totalFlight=8695, rms=43.5366}
ABE-MDT NW      {arrDelay=1625, depDelay=402, totalFlight=3167, rms=0.5286}
ABE-MDT UA      {arrDelay=3176, depDelay=2441, totalFlight=1008, rms=3.9739}
ABE-ORD AA      {arrDelay=6100, depDelay=7003, totalFlight=1514, rms=6.1342}
ABE-ORD DH      {arrDelay=8107, depDelay=10491, totalFlight=805, rms=16.4700}
ABE-ORD OO      {arrDelay=17196, depDelay=27435, totalFlight=2739, rms=11.8214}
ABE-ORD UA      {arrDelay=133825, depDelay=147202, totalFlight=17051, rms=11.6674}
ABE-ORD YV      {arrDelay=40284, depDelay=50940, totalFlight=2463, rms=26.3677}
ABE-PHL OH      {arrDelay=34, depDelay=46, totalFlight=2, rms=28.6007}
ABE-PHL UA      {arrDelay=87, depDelay=224, totalFlight=27, rms=8.9001}
ABE-PHL US      {arrDelay=6263, depDelay=5365, totalFlight=524, rms=15.7380}
ABE-PIT US      {arrDelay=77198, depDelay=61076, totalFlight=21753, rms=4.5252}
ABE-RDU AA      {arrDelay=1145, depDelay=252, totalFlight=86, rms=13.6326}
ABE-ROC EA      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
ABE-SBN OO      {arrDelay=0, depDelay=0, totalFlight=1, rms=0.0000}
ABI-CLL OO      {arrDelay=0, depDelay=0, totalFlight=3, rms=0.0000}
ABI-DFW AA      {arrDelay=0, depDelay=0, totalFlight=2, rms=0.0000}
ABI-DFW EV      {arrDelay=-9, depDelay=7, totalFlight=4, rms=2.8504}
ABI-DFW MQ      {arrDelay=53180, depDelay=71089, totalFlight=20067, rms=4.4241}
```

After I got this data I used secondary sorting to come with the final recommendation system which sorted each carrier per source destination pair with respect to least RMS values.

This gives us a Recommendation System in which we can tell a user which airlines is best suited for a particular source destination pair.

The output of this is given below:

Final Recommendation System

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /FinalProjectOutput/RecommendationSystem/2-Recommendation_System/part-r-00000
19/12/11 02:07:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
ABE-ALB : AA    25.079872407968907
ABE-ATL : EA    7.37037755523807
ABE-ATL : DL    8.761682758804241
ABE-ATL : OH    15.344853682134556
ABE-ATL : EV    19.929165111142492
ABE-AVP : EV    0.0
ABE-AVP : US    1.8080540539789332
ABE-AVP : EA    8.902804056783832
ABE-AVP : OH    23.753947040439407
ABE-AZO : OO    0.0
ABE-BDL : AA    1.0
ABE-BHM : OO    11.40175425099138
ABE-BWI : TW    2.2885326679043145
ABE-BWI : US    5.782474578119571
ABE-BWI : PI    11.723433464382353
ABE-BWI : OH    100.60318086422517
ABE-CLE : XE    4.244807590228362
ABE-CLT : US    1.321925995183028
ABE-CLT : YV    8.653579475658624
ABE-CVG : DL    2.380781074221435
ABE-CVG : OH    2.8559983029282363
ABE-CVG : EV    8.577454936302075
ABE-DCA : TW    0.6377586533979581
ABE-DTW : NW    2.934794067392623
ABE-DTW : 9E    9.669958124850933
ABE-FWA : OO    0.0
ABE-GRR : OO    0.0
ABE-HPN : EV    4.0
ABE-HPN : UA    5.792252361724888
ABE-IAD : DH    2.439688864080849
ABE-IND : NW    15.033296378372908
ABE-JFK : OH    42.5205832509386
ABE-LGA : UA    4.894243923254509
ABE-LGA : TW    9.582907674635843
ABE-LGA : OH    114.04109982361445
ABE-MCO : US    3.429671026573667
ABE-MCO : DL    4.804352446515052
ABE-MDT : NW    0.5285715066969867
ABE-MDT : UA    3.9738869989134735
ABE-MDT : AA    12.649110640673518
ABE-MDT : DL    43.536589491982426
ABE-ORD : AA    6.134211341345868
ABE-ORD : UA    11.667415129610452
ABE-ORD : OO    11.821366206962841
ABE-ORD : DH    16.47003210337515
ABE-ORD : YV    26.367722559682758
ABE-PHL : UA    8.900070128112606
ABE-PHL : US    15.73801564224951
ABE-PHL : OH    28.600699292150182
ABE-PIT : US    4.525206976344964
ABE-RDU : AA    13.632594042782774
ABE-ROC : EA    0.0
ABE-SBN : OO    0.0
ABI-CLL : OO    0.0
ABI-DFW : AA    0.0
ABI-DFW : EV    2.850438562747845
ABI-DFW : MQ    4.424142492907631
ABI-IAH : OO    6.460818889257863
```

4. Analysis using PIG

RUN COMMAND-> pig -x MapReduce Q1.pig

1. Top 20 cities grouped by count of flights

For each airport I calculated leaving, arriving and all flights. It helped in analysis of busiest airports

SCRIPT—

```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/AirlinesDataFull' 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 '/PIG-OUTPUT-FULL/Q1/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 '/PIG-OUTPUT-FULL/Q1/OUTBOUND-TOP' USING PigStorage(',');



----------------------------------------------------------
-- 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 '/PIG-OUTPUT-FULL/Q1/MONTHLY-TRAFFIC-TOP/' USING PigStorage(',');

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

OUTPUT—

```
[Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q1/IN*/part-r-00000|head
19/12/11 19:32:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
1,CVG,164511
1,SLC,165998
1,PHL,180128
1,PIT,175495
1,LGA,193053
1,IAH,237414
1,LAS,215393
1,EWR,224267
1,BOS,190556
1,DEN,270033
Reemas-MacBook-Pro:bin reemadutta$
```

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q1/MO*/part-r-00000|head
19/12/11 19:35:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pla
1,CVG,329347
1,SLC,331980
1,PHL,360343
1,LGA,386296
1,PIT,350440
1,EWR,448267
1,BOS,381318
1,SFO,445337
1,LAS,431083
1,STL,451251
Reemas-MacBook-Pro:bin reemadutta$
```

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q1/OU*/part-r-00000|head
19/12/11 19:36:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
1,CVG,164836
1,SLC,165982
1,PHL,180215
1,PIT,174945
1,LAS,215690
1,EWR,224000
1,LGA,193243
1,CLT,211650
1,BOS,190762
1,MSP,224603
Reemas-MacBook-Pro:bin reemadutta$
```

2. Most used carriers by users

To get the popularity of carriers I computed total flights and took the median over a decade.

Script: -

```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/AirlinesDataFull' 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
```

Output-

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q2/COUNT_CARRIERS/part-r-00000|head
19/12/11 19:44:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
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
Reemas-MacBook-Pro:bin reemadutta$
```

3. Ratio of Flights Delayed

A flight is delayed if the delay is greater than 15 minutes.

Calculated delay over all time- day, month and year.

Script: -

```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/AirlinesDataFull' 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;
STORE COUNT_TOTAL INTO '/PIG-OUTPUT-FULL/Q3/COUNT_TOTAL' USING PigStorage(',');
```

Output: -

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q3/CO*/part-r-00000|head
19/12/11 20:09:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platf
(2,7),1312153,282274,0.21512277
(3,2),1505892,313956,0.20848507
(5,5),1507668,370952,0.24604356
(8,3),1572143,340427,0.21653692
(10,6),1363240,213903,0.15690781
(11,1),1518318,329369,0.21693018
Reemas-MacBook-Pro:bin reemadutta$
```

4. Carrier Delays

Calculating the proportion of delayed flights by carrier, ranked by carrier, at different time (hour, day, week, month year).

Script: -

```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/AirlinesDataFull' 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;
STORE COUNT_TOTAL INTO '/PIG-OUTPUT-FULL/Q4/COUNT_TOTAL' USING PigStorage(',');
```

Output: -

```
[Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q4/CO*/part-r-00000
19/12/11 20:13:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
US,14075530,3151816,0.22392166
Reemas-MacBook-Pro:bin reemadutta$
```

5. Busiest source destination pair (routes)

For this I created a symbol table with source and destination and then grouped and sorted them on the basis of number of flights.

Script: -

```
-- First, we load the raw data from a test dataset
RAW_DATA = LOAD '/AirlinesDataFull' 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;
STORE COUNT INTO '/PIG-OUTPUT-FULL/Q5/COUNT' USING PigStorage(',');
```

Output: -

```
(YUM,GJT),1
[Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /PIG-OUTPUT-FULL/Q5/CO*/part-r-00000|head
19/12/11 20:15:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
(ABY,ATL),8033
(ALB,DTW),28191
(ALO,STL),238
(ANC,LAS),1077
(ANC,SCC),8977
(ASE,ATL),16
(ATL,BGR),955
(ATL,BHM),74221
(ATL,BTR),30092
(ATL,BZN),165
```

5. Analysis using Apache HIVE

A. Load data into HIVE

Creating schema for flight data

create schema AirlineSchema;

use AirlineSchema;

```
hive> create schema AirlineSchema;
OK
Time taken: 0.173 seconds
hive> use AirlineSchema;
OK
```

Creating table to store flight data

```
hive> create external table flightData(Year INT, Month INT, DayofMonth INT, DayOfWeek INT, DepTime INT, CRSDepTime INT, ArrTime INT, CRSArrTime INT, UniqueCarrier String, Flight
tualElapsedTime INT, CRSElapsedTime INT, AirTime INT, ArrDelay INT, DepDelay INT, Origin String, Dest String, Distance INT, TaxiIn INT, TaxiOut INT, Cancelled INT, Cancellation
, CarrierDelay INT, WeatherDelay INT, NASDelay INT, SecurityDelay INT, LateAircraftDelay INT ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.518 seconds
```

Load flight data from HDFS into table

```
hive> SET hive.exec.dynamic.partition = true;
hive> SET hive.exec.dynamic.partition.mode = nonstrict;
hive> LOAD DATA INPATH '/AirlinesDataFull' OVERWRITE INTO TABLE flightData;
Loading data to table airlineschema.flightdata
OK
Time taken: 0.628 seconds
```

Count the total number of flights using

```
hive> select count(*) from flightData;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be availa
Query ID = reemadutta_20191211203417_4ab7325b-d277-4b79-b6ee-814
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576107470183_0001, Tracking URL = http://Reem
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill
Hadoop job information for Stage-1: number of mappers: 45; numbe
2019-12-11 20:34:27,943 Stage-1 map = 0%,  reduce = 0%
2019-12-11 20:34:47,733 Stage-1 map = 7%,  reduce = 0%
2019-12-11 20:34:48,778 Stage-1 map = 13%,  reduce = 0%
2019-12-11 20:35:05,344 Stage-1 map = 16%,  reduce = 0%
2019-12-11 20:35:06,383 Stage-1 map = 20%,  reduce = 0%
2019-12-11 20:35:07,426 Stage-1 map = 24%,  reduce = 0%
2019-12-11 20:35:08,460 Stage-1 map = 27%,  reduce = 0%
2019-12-11 20:35:22,923 Stage-1 map = 29%,  reduce = 0%
2019-12-11 20:35:23,956 Stage-1 map = 36%,  reduce = 0%
2019-12-11 20:35:24,997 Stage-1 map = 38%,  reduce = 0%
2019-12-11 20:35:29,123 Stage-1 map = 38%,  reduce = 13%
2019-12-11 20:35:38,410 Stage-1 map = 40%,  reduce = 13%
2019-12-11 20:35:39,445 Stage-1 map = 47%,  reduce = 13%
2019-12-11 20:35:40,520 Stage-1 map = 49%,  reduce = 16%
2019-12-11 20:35:53,942 Stage-1 map = 53%,  reduce = 16%
2019-12-11 20:35:54,979 Stage-1 map = 60%,  reduce = 16%
2019-12-11 20:35:59,113 Stage-1 map = 60%,  reduce = 20%
2019-12-11 20:36:09,430 Stage-1 map = 64%,  reduce = 20%
2019-12-11 20:36:10,462 Stage-1 map = 71%,  reduce = 21%
2019-12-11 20:36:16,651 Stage-1 map = 71%,  reduce = 24%
2019-12-11 20:36:23,875 Stage-1 map = 73%,  reduce = 24%
2019-12-11 20:36:24,902 Stage-1 map = 76%,  reduce = 24%
2019-12-11 20:36:25,939 Stage-1 map = 82%,  reduce = 24%
2019-12-11 20:36:29,044 Stage-1 map = 82%,  reduce = 27%
2019-12-11 20:36:38,343 Stage-1 map = 84%,  reduce = 27%
2019-12-11 20:36:40,411 Stage-1 map = 87%,  reduce = 29%
2019-12-11 20:36:41,440 Stage-1 map = 93%,  reduce = 29%
2019-12-11 20:36:46,587 Stage-1 map = 93%,  reduce = 31%
2019-12-11 20:36:48,656 Stage-1 map = 96%,  reduce = 31%
2019-12-11 20:36:49,684 Stage-1 map = 100%,  reduce = 31%
2019-12-11 20:36:50,711 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1576107470183_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 45  Reduce: 1   HDFS Read: 12029868272 HDFS
Total MapReduce CPU Time Spent: 0 msec
OK
123534991
Time taken: 155.576 seconds, Fetched: 1 row(s)
```

Create table and load airports data from HDFS

```
hive> create external table airports (Iata String, aiport String, city String, state String, country String, lat String, longi String) ROW FORMAT DELIMITED FIELDS TERMI
OK
Time taken: 0.279 seconds
hive> LOAD DATA INPATH '/AirportsList.csv' OVERWRITE INTO TABLE airports;
FAILED: SemanticException Line 1:17 Invalid path ''/AirportsList.csv'': No files matching path hdfs://localhost:9000/AirportsList.csv
hive> LOAD DATA INPATH '/AirportsList' OVERWRITE INTO TABLE airports;
Loading data to table airlineschema.airports
OK
Time taken: 0.539 seconds
```

Display only the cities from airlineschema.airports table in ascending order

```
hive> select city from airlineschema.airports sort by city asc limit 50;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the
Query ID = reemadutta_20191211213650_5dda2319-bb2b-43cd-9a3c-e699cf2a9a15
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576107470183_0003, Tracking URL = http://Reemas-MacBook-
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_157610
Hadoop job information for Stage-1: number of mappers: 1; number of reducers
2019-12-11 21:36:56,141 Stage-1 map = 0%,  reduce = 0%
2019-12-11 21:37:01,352 Stage-1 map = 100%,  reduce = 0%
2019-12-11 21:37:05,492 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1576107470183_0003
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576107470183_0004, Tracking URL = http://Reemas-MacBook-
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_157610
Hadoop job information for Stage-2: number of mappers: 1; number of reducers
2019-12-11 21:37:16,986 Stage-2 map = 0%,  reduce = 0%
2019-12-11 21:37:21,120 Stage-2 map = 100%,  reduce = 0%
2019-12-11 21:37:26,282 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_1576107470183_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   HDFS Read: 217311 HDFS Write: 1378 SUCCES
Stage-Stage-2: Map: 1  Reduce: 1   HDFS Read: 6346 HDFS Write: 1119 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
 Jr."
 R G Le Tourneau "
 Ryan "
 Sr."
 Troy Shelton "
"Pullman/Moscow
"Westport
Abbeville
Abbeville
Aberdeen
Aberdeen
Aberdeen-Amory
Abilene
Abilene.
Abingdon
Ackerman
Ada
Adak
Adams/Friendship
Adel
Adrian
Afton
```

Create table and load carrier's data from HDFS

```
Time taken: 0.0745 seconds, Fetched: 50 row(s)
hive> create external table carriers (Code String, Description String) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;
OK
Time taken: 0.477 seconds
hive> LOAD DATA INPATH '/AirlineCarriers' OVERWRITE INTO TABLE carriers;
Loading data to table default.carriers
OK
Time taken: 0.344 seconds
```

B. Analysis using HIVE

1: FLIGHTS THAT TRAVELLED LESS THAN OR MORE THAN 500 AIRTIME

SCRIPT:

```
hive> INSERT OVERWRITE DIRECTORY '/HiveMROutput/1.1' select count(*) from airlineschema.flightData where AirTime > 500;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different
Query ID = reemadutta_20191211221103_060b26b7-8cc3-4acb-9fe2-208447404f30
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576107470183_0005, Tracking URL = http://Reemas-MacBook-Pro.local:8088/proxy/application_157610747018
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_1576107470183_0005
Hadoop job information for Stage-1: number of mappers: 45; number of reducers: 1
2019-12-11 22:11:11,252 Stage-1 map = 0%,  reduce = 0%
2019-12-11 22:11:33,431 Stage-1 map = 2%,  reduce = 0%
2019-12-11 22:11:34,467 Stage-1 map = 9%,  reduce = 0%
2019-12-11 22:11:35,572 Stage-1 map = 13%,  reduce = 0%
2019-12-11 22:11:55,248 Stage-1 map = 21%,  reduce = 0%
2019-12-11 22:11:56,358 Stage-1 map = 27%,  reduce = 0%
2019-12-11 22:12:17,053 Stage-1 map = 37%,  reduce = 0%
2019-12-11 22:12:18,085 Stage-1 map = 40%,  reduce = 0%
2019-12-11 22:12:35,613 Stage-1 map = 47%,  reduce = 0%
2019-12-11 22:12:36,672 Stage-1 map = 51%,  reduce = 0%
2019-12-11 22:12:37,724 Stage-1 map = 51%,  reduce = 17%
2019-12-11 22:12:54,203 Stage-1 map = 62%,  reduce = 17%
2019-12-11 22:12:55,239 Stage-1 map = 62%,  reduce = 21%
2019-12-11 22:13:11,753 Stage-1 map = 73%,  reduce = 21%
2019-12-11 22:13:13,832 Stage-1 map = 73%,  reduce = 24%
2019-12-11 22:13:30,310 Stage-1 map = 84%,  reduce = 24%
2019-12-11 22:13:31,339 Stage-1 map = 84%,  reduce = 28%
2019-12-11 22:13:47,830 Stage-1 map = 96%,  reduce = 28%
2019-12-11 22:13:49,907 Stage-1 map = 96%,  reduce = 32%
2019-12-11 22:13:55,051 Stage-1 map = 100%,  reduce = 32%
2019-12-11 22:13:56,081 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1576107470183_0005
Moving data to directory /HiveMROutput/1.1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 45  Reduce: 1   HDFS Read: 12029887064 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
```

OUTPUT:

```
1988422PHXMKE\N1460\N\N
[Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /HiveMROutput/1.1/000000_0
19/12/12 00:28:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library
30711
Reemas-MacBook-Pro:bin reemadutta$
```

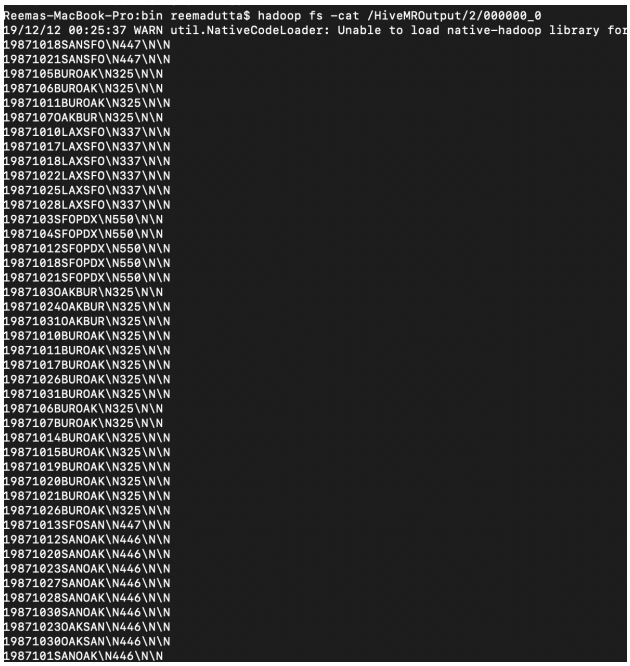
2. COUNT OF ALL THE FLIGHTS THAT WERE ON TIME WHILE ARRIVING AND DEPARTURE

SCRIPT:

```
FAILED: SemanticException [Error 10001]: Line 1:123 Table not found 'flightData'
hive> INSERT OVERWRITE DIRECTORY '/HiveMROutput/2' select Year,Month,DayofMonth,Origin,Dest,AirTime,Distance,TaxiIn,TaxiOut from airlineschema.flightData where DepTime<=CRSDepTime and ArrTime<=CRSArrTime

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = reemadutta_20191211233839_4db1c753-1009-437f-8f95-d65b5a2e2174
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1576107470183_0006, Tracking URL = http://Reemas-MacBook-Pro.local:8088/proxy/application_1576107470183_0006/
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_1576107470183_0006
Hadoop job information for Stage-1: number of mappers: 45; number of reducers: 0
2019-12-11 23:38:45,136 Stage-1 map = 0%,  reduce = 0%
2019-12-11 23:39:14,093 Stage-1 map = 13%,  reduce = 0%
2019-12-11 23:39:35,787 Stage-1 map = 18%,  reduce = 0%
2019-12-11 23:39:38,886 Stage-1 map = 19%,  reduce = 0%
2019-12-11 23:39:39,919 Stage-1 map = 20%,  reduce = 0%
2019-12-11 23:39:40,942 Stage-1 map = 27%,  reduce = 0%
2019-12-11 23:40:01,558 Stage-1 map = 30%,  reduce = 0%
2019-12-11 23:40:02,588 Stage-1 map = 32%,  reduce = 0%
2019-12-11 23:40:06,712 Stage-1 map = 39%,  reduce = 0%
2019-12-11 23:40:07,745 Stage-1 map = 40%,  reduce = 0%
2019-12-11 23:40:33,577 Stage-1 map = 47%,  reduce = 0%
2019-12-11 23:40:34,607 Stage-1 map = 53%,  reduce = 0%
2019-12-11 23:41:00,404 Stage-1 map = 54%,  reduce = 0%
2019-12-11 23:41:02,459 Stage-1 map = 59%,  reduce = 0%
2019-12-11 23:41:03,493 Stage-1 map = 62%,  reduce = 0%
2019-12-11 23:41:04,536 Stage-1 map = 66%,  reduce = 0%
2019-12-11 23:41:05,566 Stage-1 map = 67%,  reduce = 0%
2019-12-11 23:41:29,341 Stage-1 map = 68%,  reduce = 0%
2019-12-11 23:41:31,416 Stage-1 map = 69%,  reduce = 0%
2019-12-11 23:41:33,500 Stage-1 map = 70%,  reduce = 0%
2019-12-11 23:41:34,544 Stage-1 map = 72%,  reduce = 0%
2019-12-11 23:41:35,588 Stage-1 map = 73%,  reduce = 0%
2019-12-11 23:41:36,634 Stage-1 map = 74%,  reduce = 0%
2019-12-11 23:41:39,751 Stage-1 map = 79%,  reduce = 0%
2019-12-11 23:41:40,820 Stage-1 map = 80%,  reduce = 0%
2019-12-11 23:42:07,717 Stage-1 map = 81%,  reduce = 0%
2019-12-11 23:42:09,774 Stage-1 map = 84%,  reduce = 0%
2019-12-11 23:42:10,803 Stage-1 map = 87%,  reduce = 0%
2019-12-11 23:42:11,838 Stage-1 map = 89%,  reduce = 0%
2019-12-11 23:42:12,872 Stage-1 map = 93%,  reduce = 0%
2019-12-11 23:42:27,286 Stage-1 map = 98%,  reduce = 0%
2019-12-11 23:42:28,308 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_1576107470183_0006
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:9000/HiveMROutput/2/.hive-staging_hive_2019-12-11_23-38-39_300_5565093627066015218-1/-ext-10000
Moving data to directory /HiveMROutput/2
MapReduce Jobs Launched:
Stage-Stage-1: Map: 45   HDFS Read: 12029873117 HDFS Write: 1548690839 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 230.173 seconds
```

OUTPUT:


Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /HiveMROutput/2/000000_0
19/12/12 00:25:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
1987101SANSFO\N447\N\N
1987102SANSFO\N447\N\N
1987105BUROAK\N325\N\N
1987106BUROAK\N325\N\N
1987111BUROAK\N325\N\N
1987107OAKBUR\N325\N\N
1987110LAXSFO\N337\N\N
1987117LAXSFO\N337\N\N
1987118LAXSFO\N337\N\N
1987122LAXSFO\N337\N\N
1987125LAXSFO\N337\N\N
1987128LAXSFO\N337\N\N
1987103SFOPDX\N550\N\N
1987104SFOPDX\N550\N\N
1987112SFOPDX\N550\N\N
1987118SFOPDX\N550\N\N
1987121SFOPDX\N550\N\N
1987103OAKBUR\N325\N\N
1987102OAKBUR\N325\N\N
1987103OAKBUR\N325\N\N
1987110BUROAK\N325\N\N
1987111BUROAK\N325\N\N
1987117BUROAK\N325\N\N
1987126BUROAK\N325\N\N
1987131BUROAK\N325\N\N
1987106BUROAK\N325\N\N
1987107BUROAK\N325\N\N
1987114BUROAK\N325\N\N
1987115BUROAK\N325\N\N
1987119BUROAK\N325\N\N
1987120BUROAK\N325\N\N
1987121BUROAK\N325\N\N
1987126BUROAK\N325\N\N
1987113SFOSAN\N447\N\N
1987112SANOAK\N446\N\N
1987120SANOAK\N446\N\N
1987123SANOAK\N446\N\N
1987127SANOAK\N446\N\N
1987128SANOAK\N446\N\N
1987130SANOAK\N446\N\N
1987123OAKSAN\N446\N\N
1987130OAKSAN\N446\N\N
1987101SANOAK\N446\N\N

3: COUNT OF ALL THE FLIGHTS THAT TOOK MORE THAN 45 MINS TO DEP AND ARRIVAL DELAY

SCRIPT:

```
hive> select count(*) from airlineschema.flightData where ArrDelay + DepDelay >45;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions
Query ID = reemadutta_20191211234452_f4e22358-4070-4c78-9fbe-16fe1f3fb411
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576107470183_0007, Tracking URL = http://Reemas-MacBook-Pro.local:8088/
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_1576107470183_0007
Hadoop job information for Stage-1: number of mappers: 45; number of reducers: 1
2019-12-11 23:44:58,698 Stage-1 map = 0%,  reduce = 0%
2019-12-11 23:45:21,409 Stage-1 map = 4%,  reduce = 0%
2019-12-11 23:45:22,473 Stage-1 map = 13%,  reduce = 0%
2019-12-11 23:45:43,131 Stage-1 map = 16%,  reduce = 0%
2019-12-11 23:45:44,164 Stage-1 map = 18%,  reduce = 0%
2019-12-11 23:45:46,264 Stage-1 map = 21%,  reduce = 0%
2019-12-11 23:45:47,308 Stage-1 map = 25%,  reduce = 0%
2019-12-11 23:45:48,338 Stage-1 map = 27%,  reduce = 0%
2019-12-11 23:46:07,096 Stage-1 map = 28%,  reduce = 0%
2019-12-11 23:46:08,124 Stage-1 map = 31%,  reduce = 0%
2019-12-11 23:46:09,160 Stage-1 map = 38%,  reduce = 10%
2019-12-11 23:46:15,356 Stage-1 map = 38%,  reduce = 13%
2019-12-11 23:46:27,757 Stage-1 map = 42%,  reduce = 13%
2019-12-11 23:46:28,790 Stage-1 map = 49%,  reduce = 13%
2019-12-11 23:46:33,969 Stage-1 map = 49%,  reduce = 16%
2019-12-11 23:46:47,428 Stage-1 map = 50%,  reduce = 16%
2019-12-11 23:46:48,458 Stage-1 map = 52%,  reduce = 16%
2019-12-11 23:46:49,498 Stage-1 map = 59%,  reduce = 16%
2019-12-11 23:46:50,532 Stage-1 map = 60%,  reduce = 16%
2019-12-11 23:46:51,561 Stage-1 map = 60%,  reduce = 20%
2019-12-11 23:47:08,093 Stage-1 map = 62%,  reduce = 20%
2019-12-11 23:47:09,129 Stage-1 map = 71%,  reduce = 20%
2019-12-11 23:47:10,169 Stage-1 map = 71%,  reduce = 23%
2019-12-11 23:47:15,344 Stage-1 map = 71%,  reduce = 24%
2019-12-11 23:47:27,754 Stage-1 map = 78%,  reduce = 24%
2019-12-11 23:47:28,782 Stage-1 map = 82%,  reduce = 24%
2019-12-11 23:47:33,968 Stage-1 map = 82%,  reduce = 27%
2019-12-11 23:47:47,394 Stage-1 map = 89%,  reduce = 27%
2019-12-11 23:47:48,500 Stage-1 map = 93%,  reduce = 27%
2019-12-11 23:47:51,614 Stage-1 map = 93%,  reduce = 31%
2019-12-11 23:47:58,842 Stage-1 map = 96%,  reduce = 31%
2019-12-11 23:47:59,874 Stage-1 map = 100%,  reduce = 31%
2019-12-11 23:48:00,900 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1576107470183_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 45  Reduce: 1   HDFS Read: 12029893142 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
14570465
Time taken: 189.137 seconds, Fetched: 1 row(s)
```

4: COUNT OF FLIGHTS FOR EACH CARRIER

SCRIPT:

```
hive> INSERT OVERWRITE DIRECTORY '/HiveMROutput/3' Select carriers.description, uniqCount.countCancelled, uniqCount.countCarrier from (Select UniqueCarrier, sum(cancelled) as countCancelled, count(*) as c
ountCarrier from airlineschema.flightData group by UniqueCarrier) AS uniqCount, carriers where carriers.code = uniqCount.UniqueCarrier;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = reemadutta_20191211234955_a898e507-ff78-4d3b-8f83-5eabac67e476
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 47
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576107470183_0008, Tracking URL = http://Reemas-MacBook-Pro.local:8088/proxy/application_1576107470183_0008/
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_1576107470183_0008
Hadoop job information for Stage-1: number of mappers: 45; number of reducers: 47
2019-12-11 23:50:03,875 Stage-1 map = 0%,  reduce = 0%
2019-12-11 23:50:27,699 Stage-1 map = 4%,  reduce = 0%
2019-12-11 23:50:29,775 Stage-1 map = 6%,  reduce = 0%
2019-12-11 23:50:30,917 Stage-1 map = 13%,  reduce = 0%
2019-12-11 23:50:51,578 Stage-1 map = 18%,  reduce = 0%
2019-12-11 23:50:53,630 Stage-1 map = 27%,  reduce = 0%
2019-12-11 23:51:14,280 Stage-1 map = 30%,  reduce = 0%
2019-12-11 23:51:15,307 Stage-1 map = 31%,  reduce = 0%
2019-12-11 23:51:16,353 Stage-1 map = 34%,  reduce = 0%
2019-12-11 23:51:17,383 Stage-1 map = 40%,  reduce = 0%
2019-12-11 23:51:38,066 Stage-1 map = 43%,  reduce = 0%
2019-12-11 23:51:39,113 Stage-1 map = 44%,  reduce = 0%
2019-12-11 23:51:40,144 Stage-1 map = 50%,  reduce = 0%
2019-12-11 23:51:41,196 Stage-1 map = 51%,  reduce = 0%
2019-12-11 23:51:59,840 Stage-1 map = 58%,  reduce = 0%
2019-12-11 23:52:00,874 Stage-1 map = 60%,  reduce = 0%
2019-12-11 23:52:01,919 Stage-1 map = 60%,  reduce = 1%
2019-12-11 23:52:14,296 Stage-1 map = 67%,  reduce = 1%
2019-12-11 23:52:25,643 Stage-1 map = 71%,  reduce = 1%
2019-12-11 23:52:26,672 Stage-1 map = 73%,  reduce = 1%
2019-12-11 23:52:31,832 Stage-1 map = 73%,  reduce = 2%
```

```
2019-12-11 23:54:19,298 Stage-1 map = 100%,  reduce = 62%
2019-12-11 23:54:20,327 Stage-1 map = 100%,  reduce = 64%
2019-12-11 23:54:21,360 Stage-1 map = 100%,  reduce = 66%
2019-12-11 23:54:22,402 Stage-1 map = 100%,  reduce = 68%
2019-12-11 23:54:23,436 Stage-1 map = 100%,  reduce = 70%
2019-12-11 23:54:26,532 Stage-1 map = 100%,  reduce = 72%
2019-12-11 23:54:31,694 Stage-1 map = 100%,  reduce = 74%
2019-12-11 23:54:32,724 Stage-1 map = 100%,  reduce = 77%
2019-12-11 23:54:33,773 Stage-1 map = 100%,  reduce = 79%
2019-12-11 23:54:34,804 Stage-1 map = 100%,  reduce = 81%
2019-12-11 23:54:35,835 Stage-1 map = 100%,  reduce = 83%
2019-12-11 23:54:38,928 Stage-1 map = 100%,  reduce = 85%
2019-12-11 23:54:43,057 Stage-1 map = 100%,  reduce = 87%
2019-12-11 23:54:44,084 Stage-1 map = 100%,  reduce = 89%
2019-12-11 23:54:45,124 Stage-1 map = 100%,  reduce = 91%
2019-12-11 23:54:46,152 Stage-1 map = 100%,  reduce = 96%
2019-12-11 23:54:48,218 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_1576107470183_0008
2019-12-11 23:55:00     Starting to launch local task to process map join;      maximum
2019-12-11 23:55:01     Dump the side-table for tag: 1 with group count: 1492 into file
11_23-49-55_334_8831561791488485038-1/-local-10003/HashTable-Stage-4/MapJoin-mapfile01-
2019-12-11 23:55:01     Uploaded 1 File to: file:/var/folders/l6/fgcp8zh13mdbfl1_hhsc6c
10003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable (63478 bytes)
2019-12-11 23:55:01     End of local task; Time Taken: 1.042 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1576107470183_0009, Tracking URL = http://Reemas-MacBook-Pro.local:8
Kill Command = /usr/local/bin/hadoop-2.8.5//bin/hadoop job  -kill job_1576107470183_000
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2019-12-11 23:55:10,445 Stage-4 map = 0%,  reduce = 0%
2019-12-11 23:55:14,583 Stage-4 map = 100%,  reduce = 0%
Ended Job = job_1576107470183_0009
Moving data to directory /HiveMROutput/3
MapReduce Jobs Launched:
Stage-Stage-1: Map: 45  Reduce: 47   HDFS Read: 12030088857 HDFS Write: 5344 SUCCESS
Stage-Stage-4: Map: 1   HDFS Read: 22479 HDFS Write: 1185 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 321.33 seconds
```

Output:-

```
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /HiveMROutput/3/000000_0|head
19/12/12 00:23:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Southwest Airlines Co.15505315976022
Pinnacle Airlines Inc.15039521059
Midway Airlines Inc. (1)134270622
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)554313636682
American Airlines Inc.28688914984647
Trans World Airways LLC690883757747
Northwest Airlines Inc.21415410292627
Piedmont Aviation Inc.8910873957
ATA Airlines d/b/a ATA2307208420
Atlantic Southeast Airlines486761697172
Reemas-MacBook-Pro:bin reemadutta$ hadoop fs -cat /HiveMROutput/3/000000_0
19/12/12 00:23:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Southwest Airlines Co.15505315976022
Pinnacle Airlines Inc.15039521059
Midway Airlines Inc. (1)134270622
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)554313636682
American Airlines Inc.28688914984647
Trans World Airways LLC690883757747
Northwest Airlines Inc.21415410292627
Piedmont Aviation Inc.8910873957
ATA Airlines d/b/a ATA2307208420
Atlantic Southeast Airlines486761697172
Independence Air22176693047
Frontier Airlines Inc.1778336958
United Air Lines Inc.29050613299817
Delta Air Lines Inc.25838216547870
American Eagle Airlines Inc.1574783954895
Expressjet Airlines Inc.519912350309
Pan American World Airways (1)3521316167
Pacific Southwest Airlines115183617
Mesa Airlines Inc.30050854056
Aloha Airlines Inc.2837154381
Alaska Airlines Inc.571212878021
Comair Inc.471741464176
JetBlue Airways9281811341
AirTran Airways Corporation128541265138
Skywest Airlines Inc.653903090853
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)29165014075530
Continental Air Lines Inc.1130648145788
Hawaiian Airlines Inc.1329274265
Eastern Air Lines Inc.28702919785
Reemas-MacBook-Pro:bin reemadutta$
```

5. Most cancellation due to Bad weather grouped by months.

SCRIPT:

```
hive> SELECT month,COUNT(cancelled) as t FROM airlineschema.flightdata WHERE cancelled = 1 AND cancell
ationcode = 'B' GROUP BY month ORDER BY t DESC LIMIT 5;
```

OUTPUT:

```
2019-12-12 00:38:44,135 Stage-1 map = 0%,  reduce = 0%
2019-12-12 00:39:05,990 Stage-1 map = 4%,  reduce = 0%
2019-12-12 00:39:08,077 Stage-1 map = 7%,  reduce = 0%
2019-12-12 00:39:09,123 Stage-1 map = 13%,  reduce = 0%
2019-12-12 00:39:29,895 Stage-1 map = 18%,  reduce = 0%
2019-12-12 00:39:31,982 Stage-1 map = 21%,  reduce = 0%
2019-12-12 00:39:33,042 Stage-1 map = 27%,  reduce = 0%
2019-12-12 00:39:59,041 Stage-1 map = 40%,  reduce = 0%
2019-12-12 00:40:20,769 Stage-1 map = 43%,  reduce = 0%
2019-12-12 00:40:21,798 Stage-1 map = 51%,  reduce = 0%
2019-12-12 00:40:38,331 Stage-1 map = 58%,  reduce = 1%
2019-12-12 00:40:39,373 Stage-1 map = 60%,  reduce = 1%
2019-12-12 00:40:52,792 Stage-1 map = 67%,  reduce = 1%
2019-12-12 00:41:05,182 Stage-1 map = 73%,  reduce = 1%
```

```
Total MapReduce CPU Time Spent: 0 msec
OK
12      48868
1       42641
2       38234
9       27524
3       23179
Time taken: 308.562 seconds, Fetched: 5 row(s)
hive>
```

6. Top 10 routes that has maximum diversions

SCRIPT:

```
hive> SELECT origin,dest,COUNT(diverted) as t FROM airlineschema.flightdata
    > WHERE diverted = 1
    > GROUP BY origin,dest
    > ORDER BY t DESC
    > LIMIT 10;
```

```
Hadoop job information for Stage-1: number of mappers: 45; number of reducers: 47
2019-12-12 00:47:02,664 Stage-1 map = 0%,  reduce = 0%
2019-12-12 00:47:31,644 Stage-1 map = 4%,  reduce = 0%
2019-12-12 00:47:32,678 Stage-1 map = 6%,  reduce = 0%
2019-12-12 00:47:33,715 Stage-1 map = 12%,  reduce = 0%
2019-12-12 00:47:34,761 Stage-1 map = 13%,  reduce = 0%
```

OUTPUT:

```
ORD     LGA     1190
LGA     DFW     987
DFW     LGA     898
MIA     LGA     823
DAL     HOU     758
ATL     LGA     748
DEN     EWR     731
ORD     DCA     718
BOS     DCA     696
LAX     JFK     669
Time taken: 367.66 seconds, Fetched: 10 row(s)
hive>
```

7. Top 5 visited destination

SCRIPT:

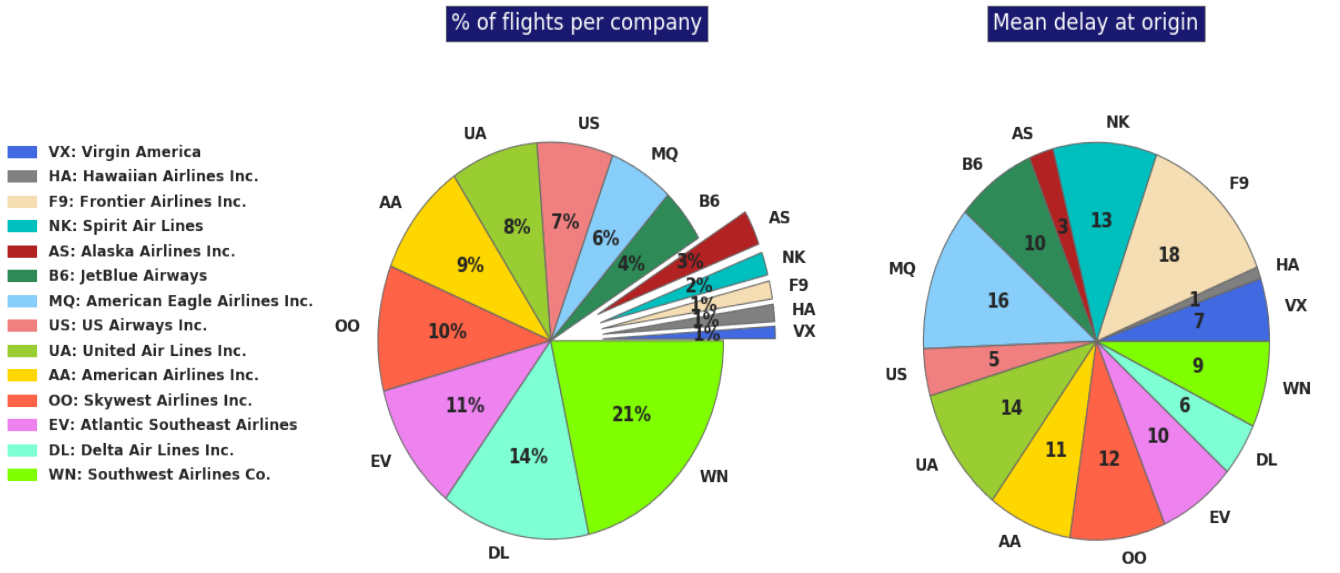
```
hive> SELECT dest,COUNT(dest) as x FROM airlineschema.flightdata
    > GROUP BY dest
    > ORDER BY x DESC
    > LIMIT 5;
```

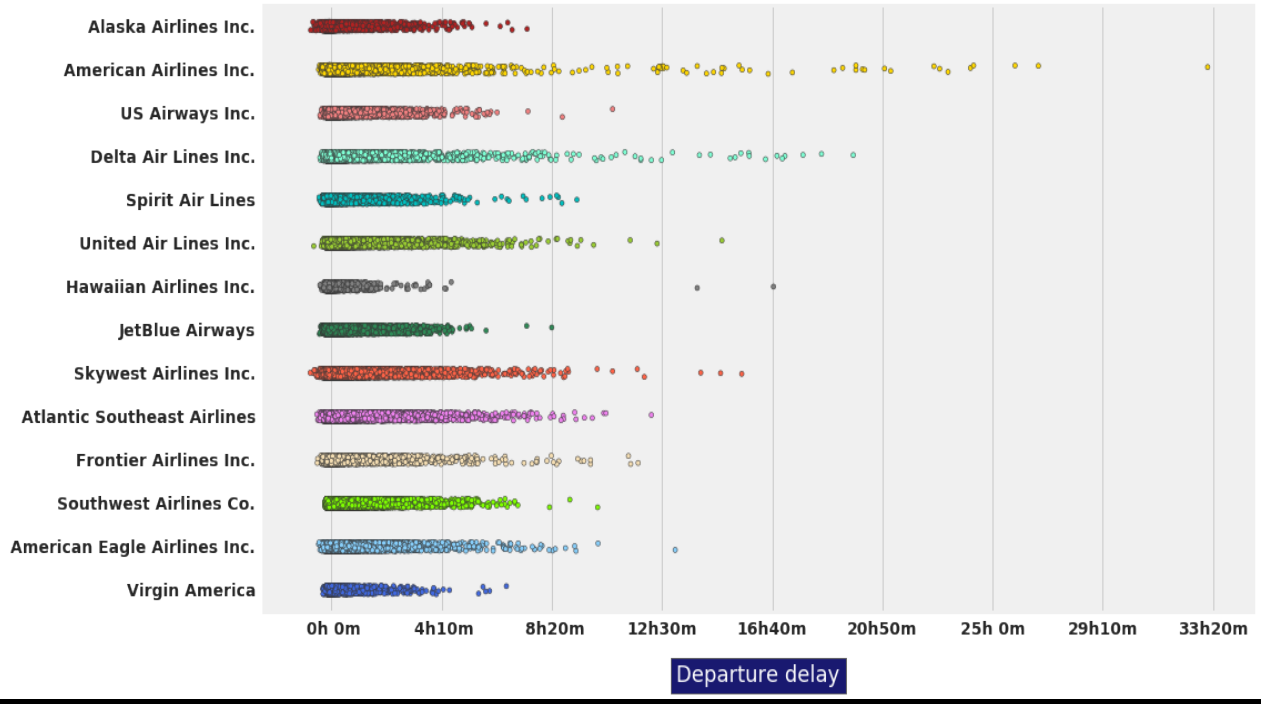
OUTPUT:

```
ORD     6638035
ATL     6094186
DFW     5745593
LAX     4086930
PHX     3497764
Time taken: 315.871 seconds, Fetched: 5 row(s)
```

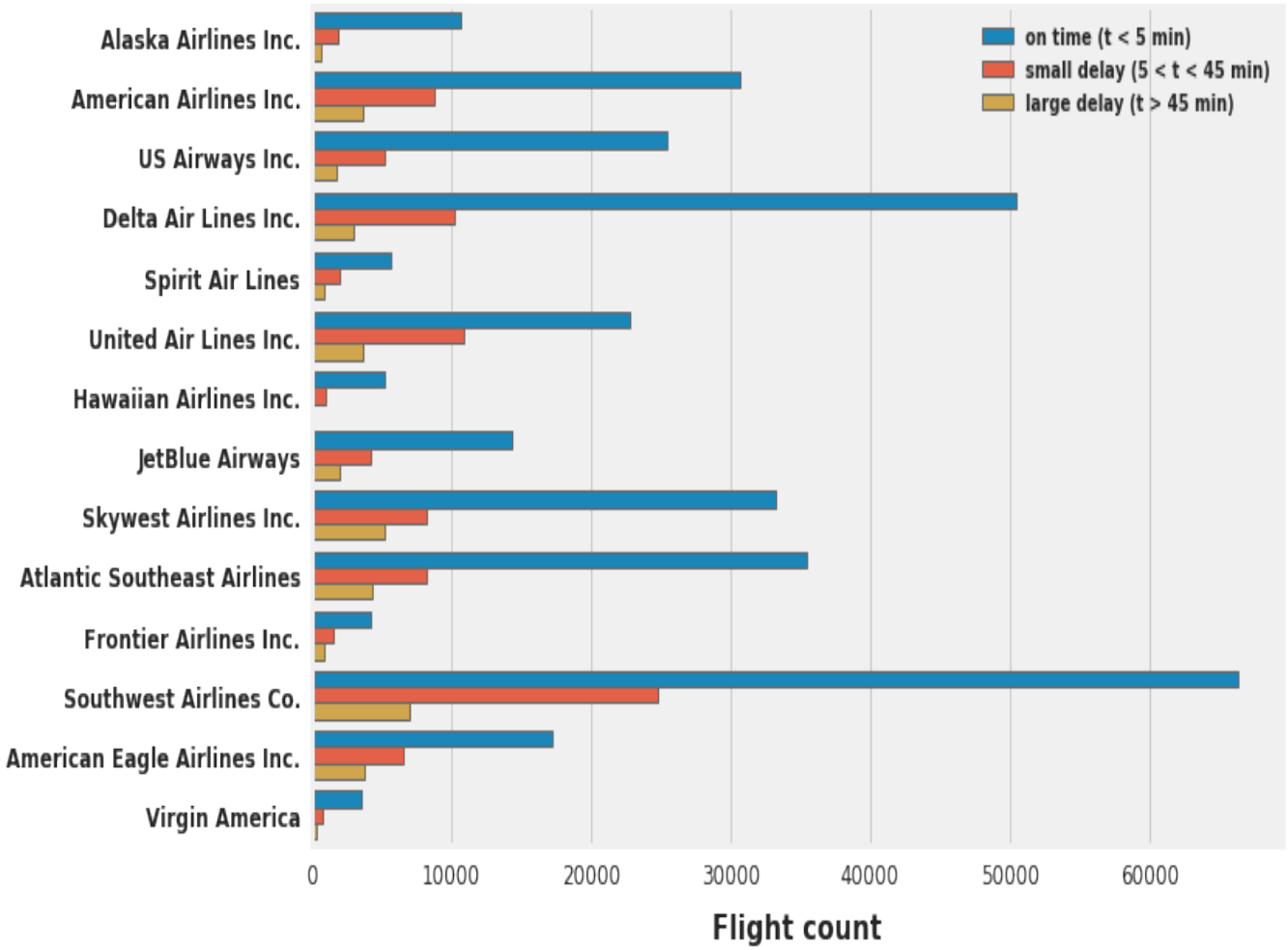
6. Visualization of analyzed results

A. Percentage of flights per carrier and their delay

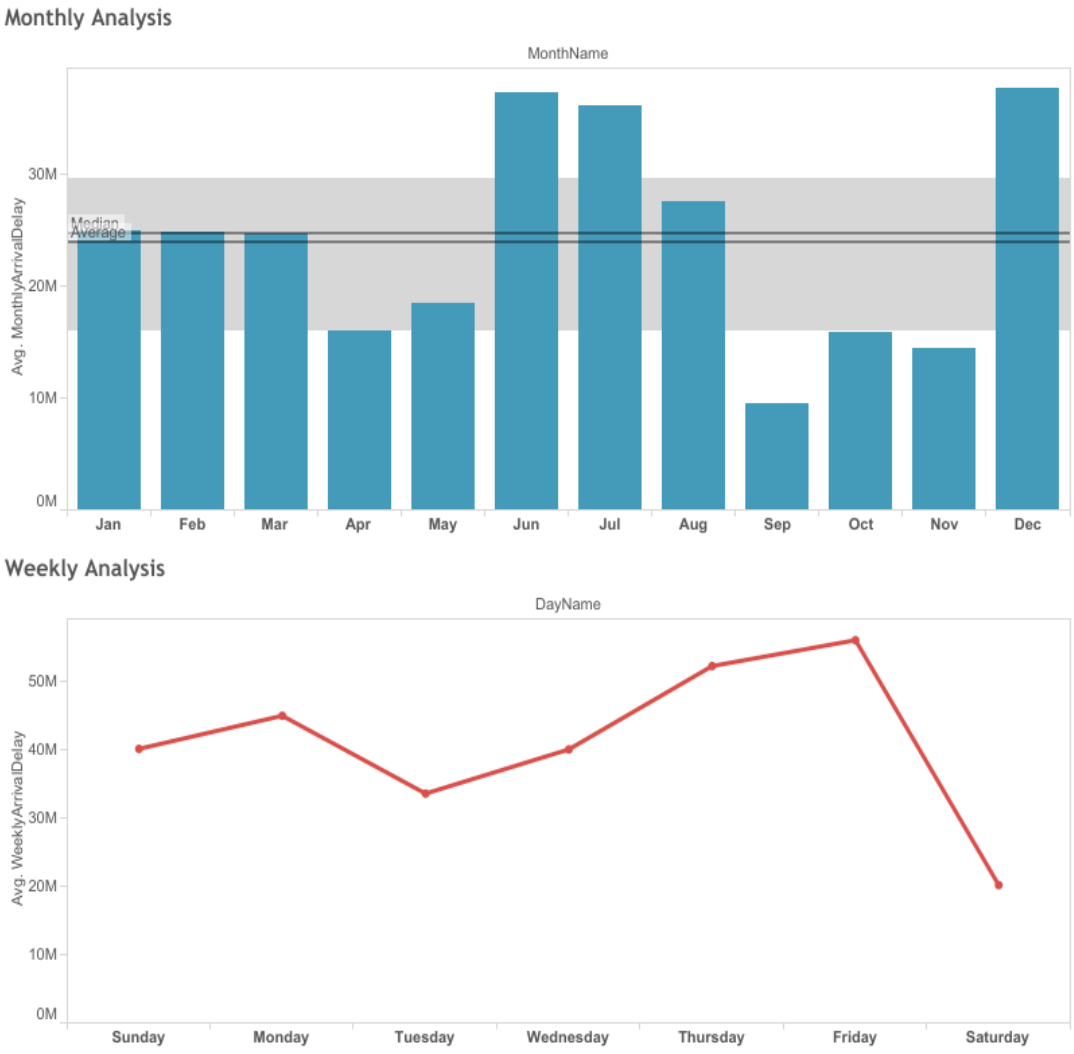




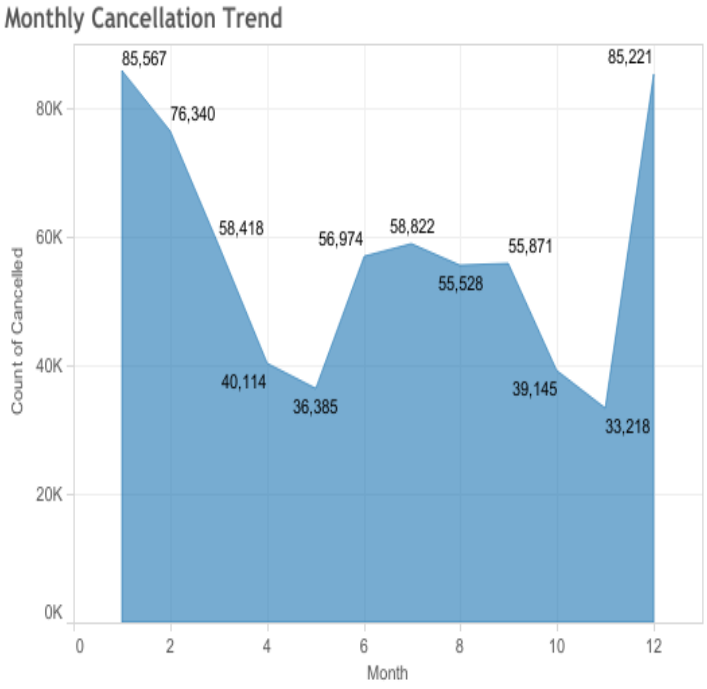
B. Various delay count of each carriers

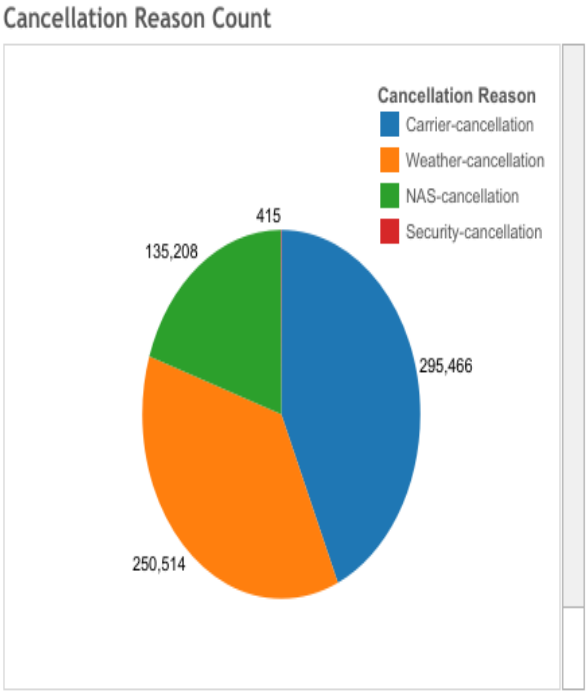


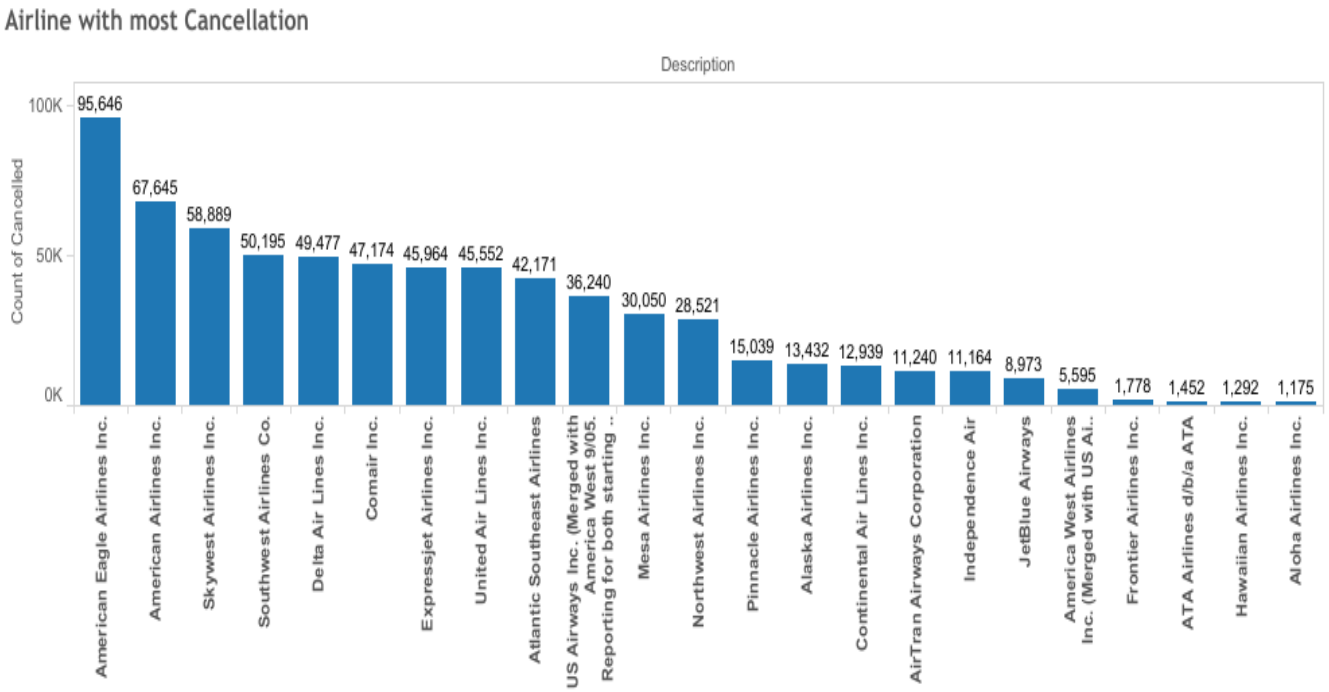
C. Daily, Monthly and weekly analysis of flight count over complete dataset



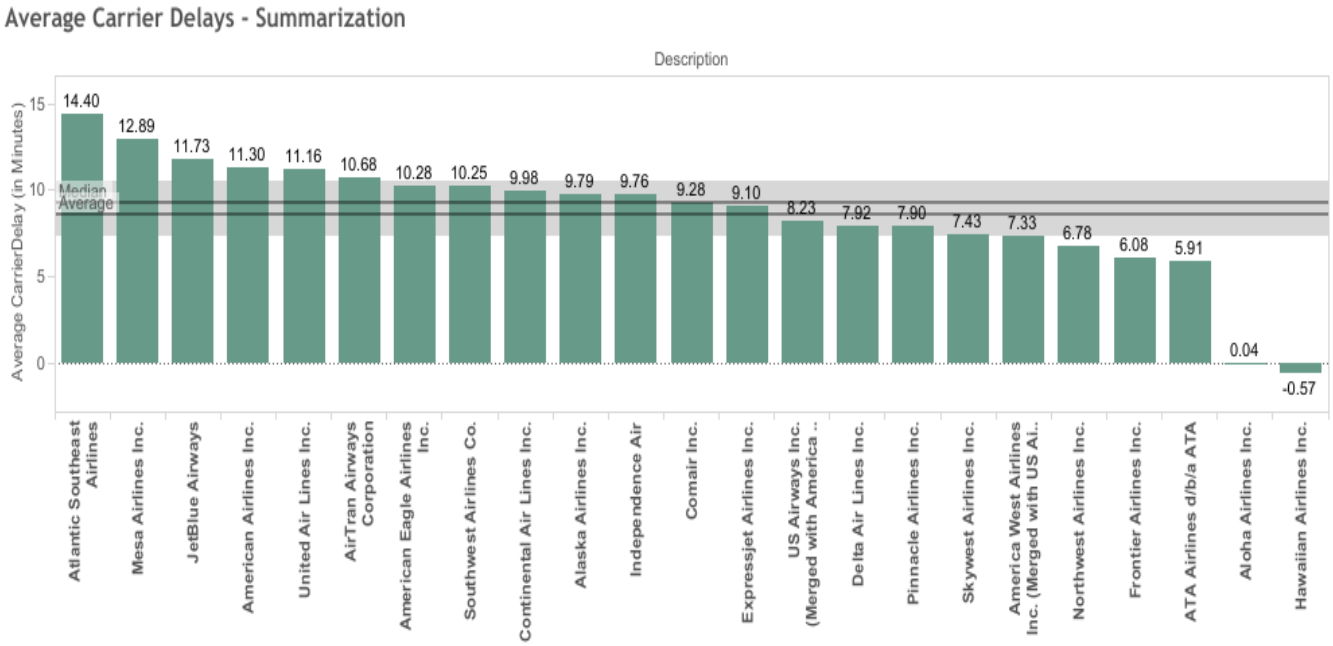
D. Cancellation trend of airlines and airlines with most cancellation

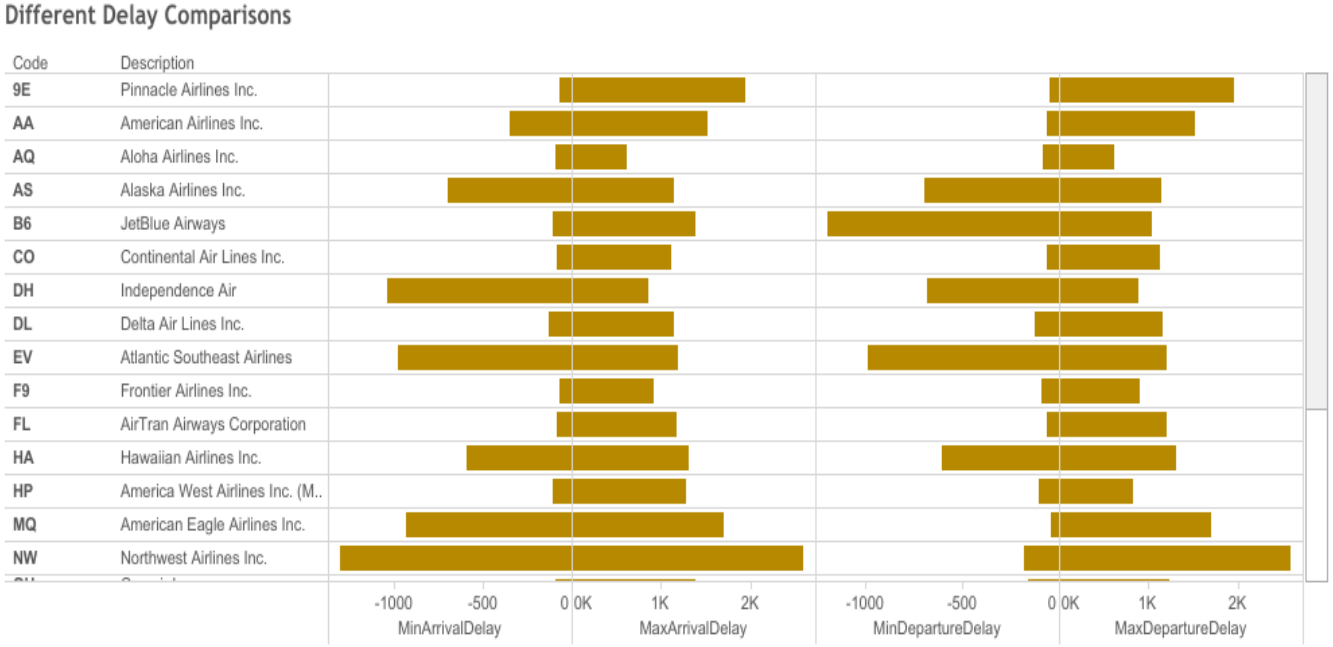




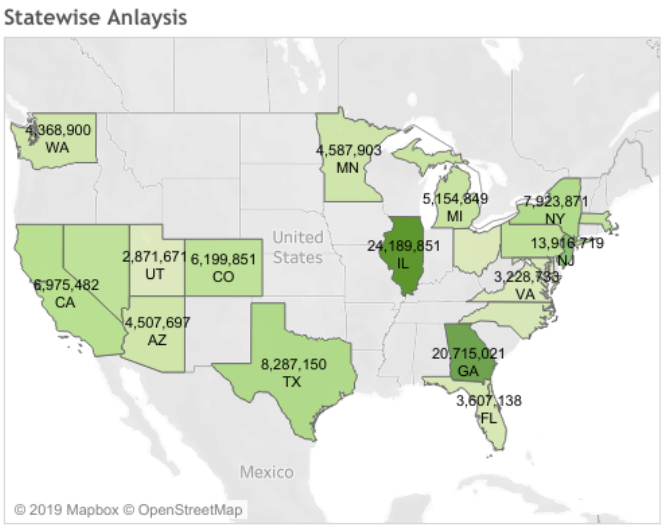


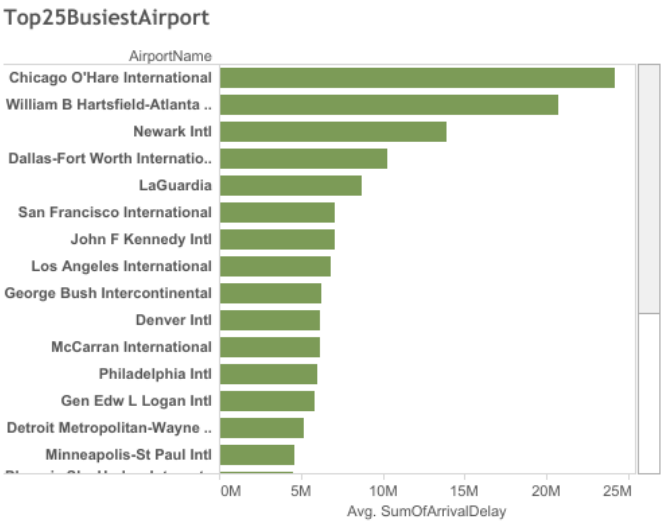
E. Average carrier delay for arrival and departure

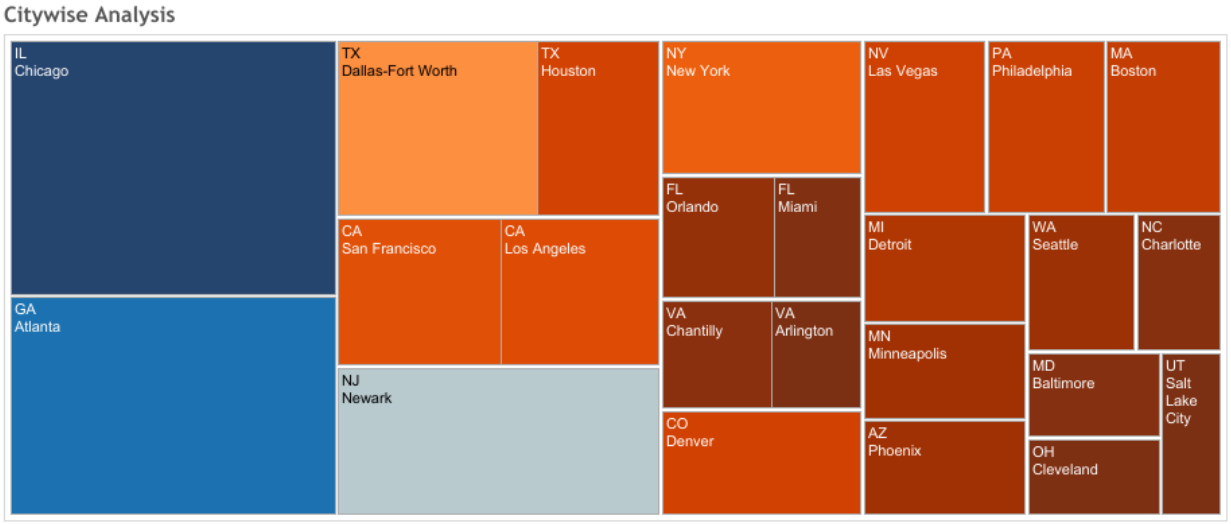




F. State and city wise analysis of flight count with top 25 busiest airport







7. 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
5. https://en.wikipedia.org/wiki/Root-mean-square_deviation
6. https://en.wikipedia.org/wiki/Mean_absolute_error