ENGINEERING OF BIG-DATA SYSTEMS

Submitted by: Venkata Nitin Reddy Byreddy(002191804)

Overview of the dataset:

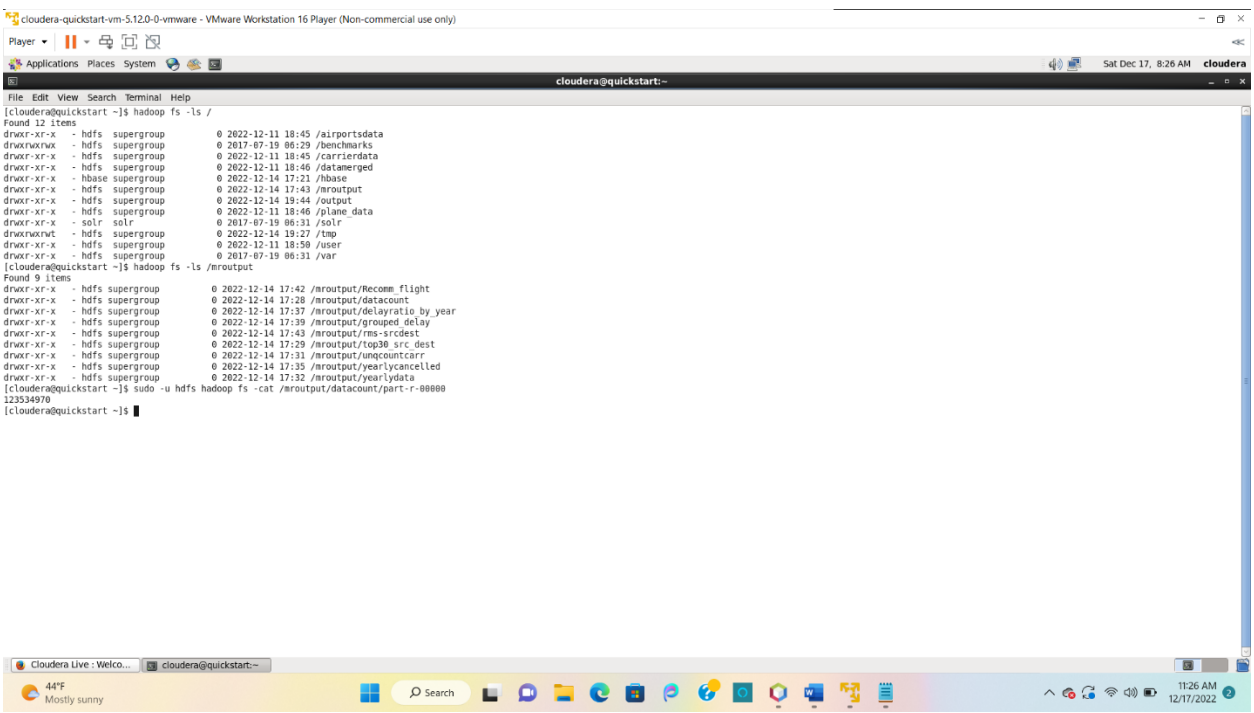
I used the Airline On time Statistics and Delay Causes dataset.

Dataset link: : https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009

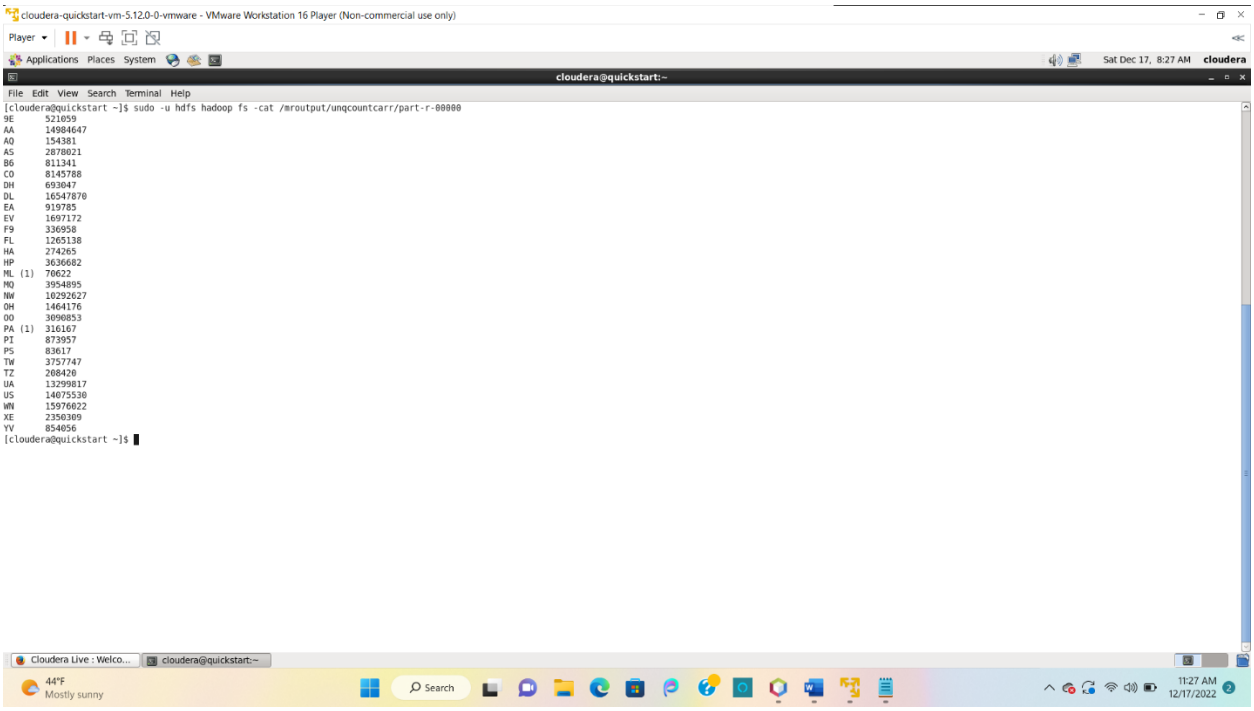
I decided to take this dataset as it has many columns which helped to explore different map reduce algorithms.

MapReduce Analysis on Flight Data:

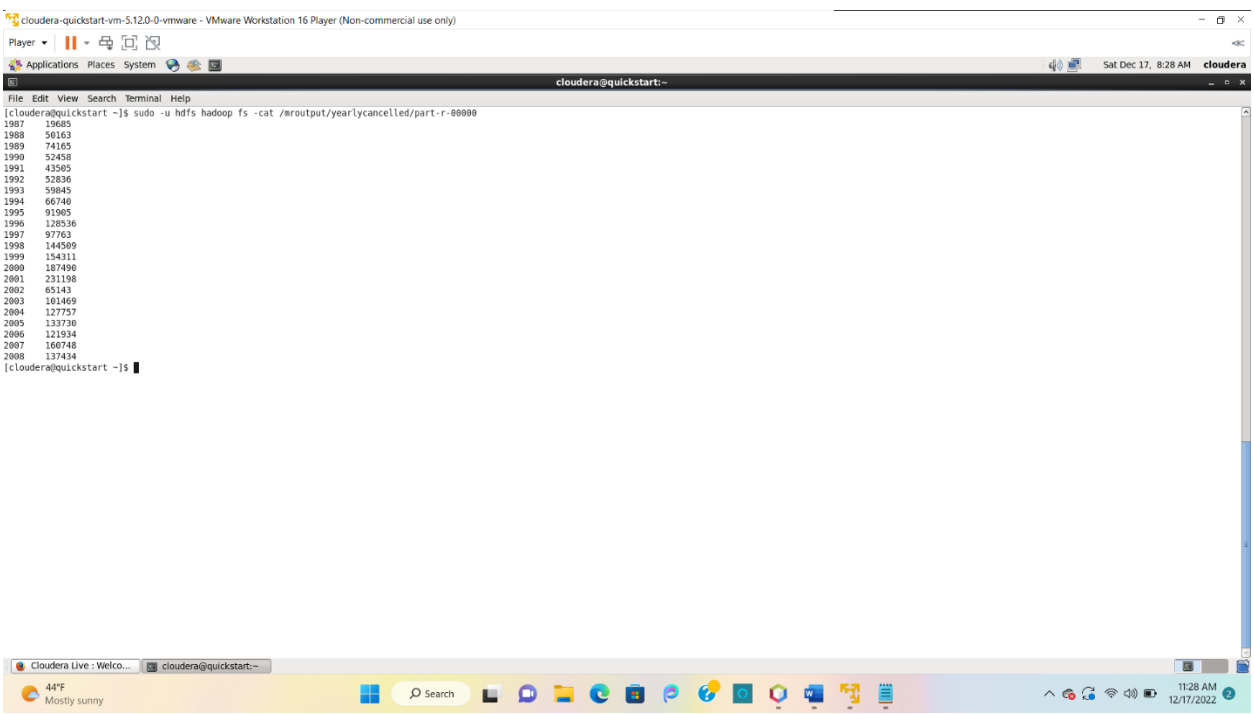
1)Counting the total number of records:



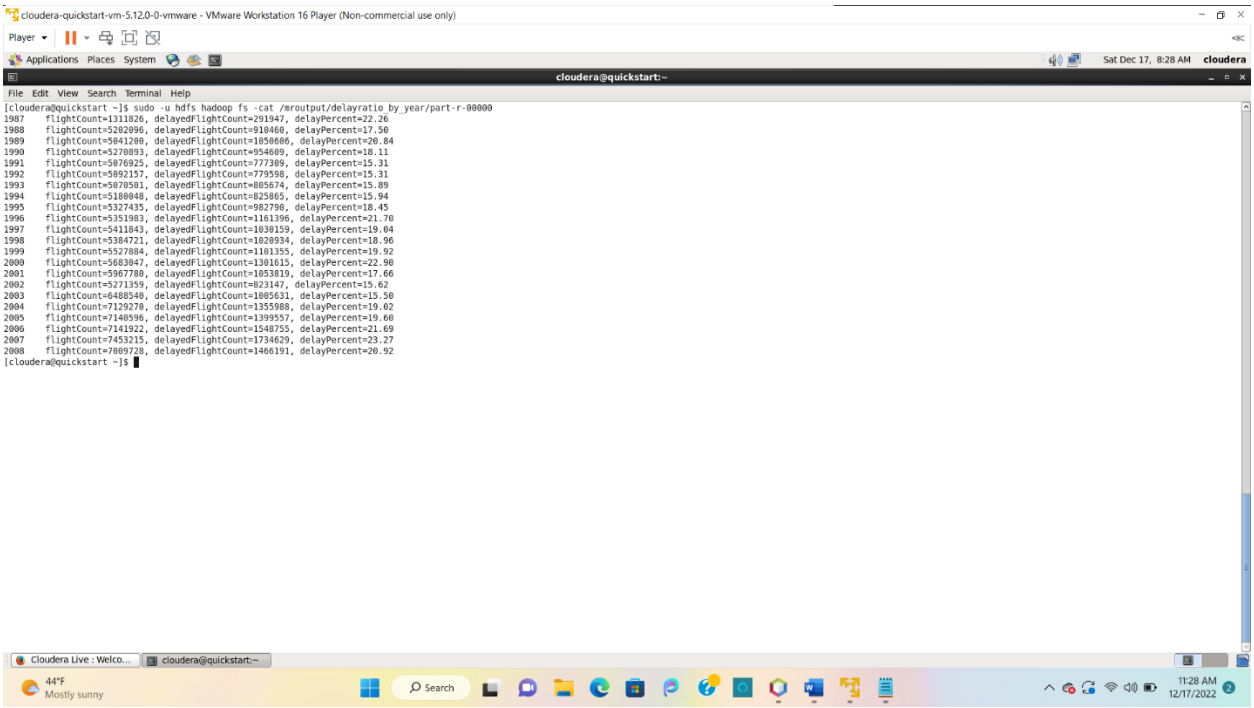
2)Calculating the number of unique carrier flights



3) Flights that are cancelled by year

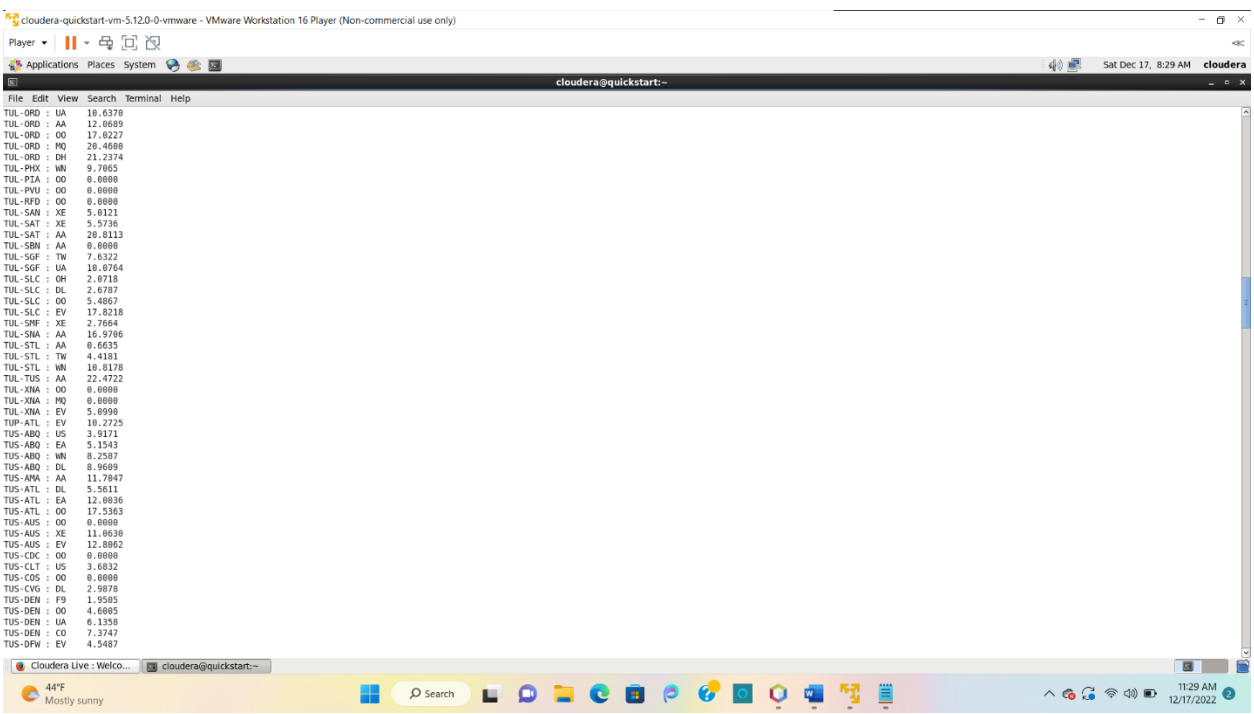


4) Ratio of delayed flights



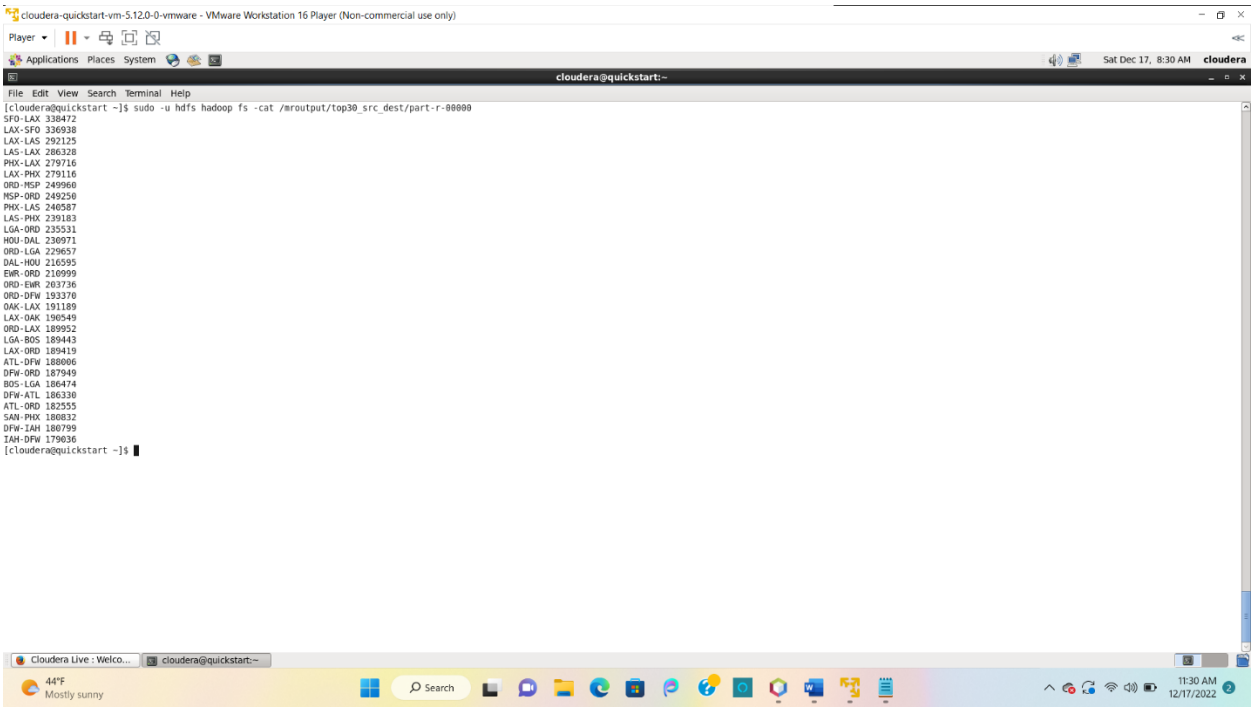
```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -cat /mroutput/delayratio_by_year/part-r-00000
1987    flightCount=1311826, delayedFlightCount=291947, delayPercent=22.26
1988    flightCount=5202096, delayedFlightCount=910460, delayPercent=17.50
1989    flightCount=5041200, delayedFlightCount=1050606, 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
[cloudera@quickstart ~]$
```

5)Recommendation system

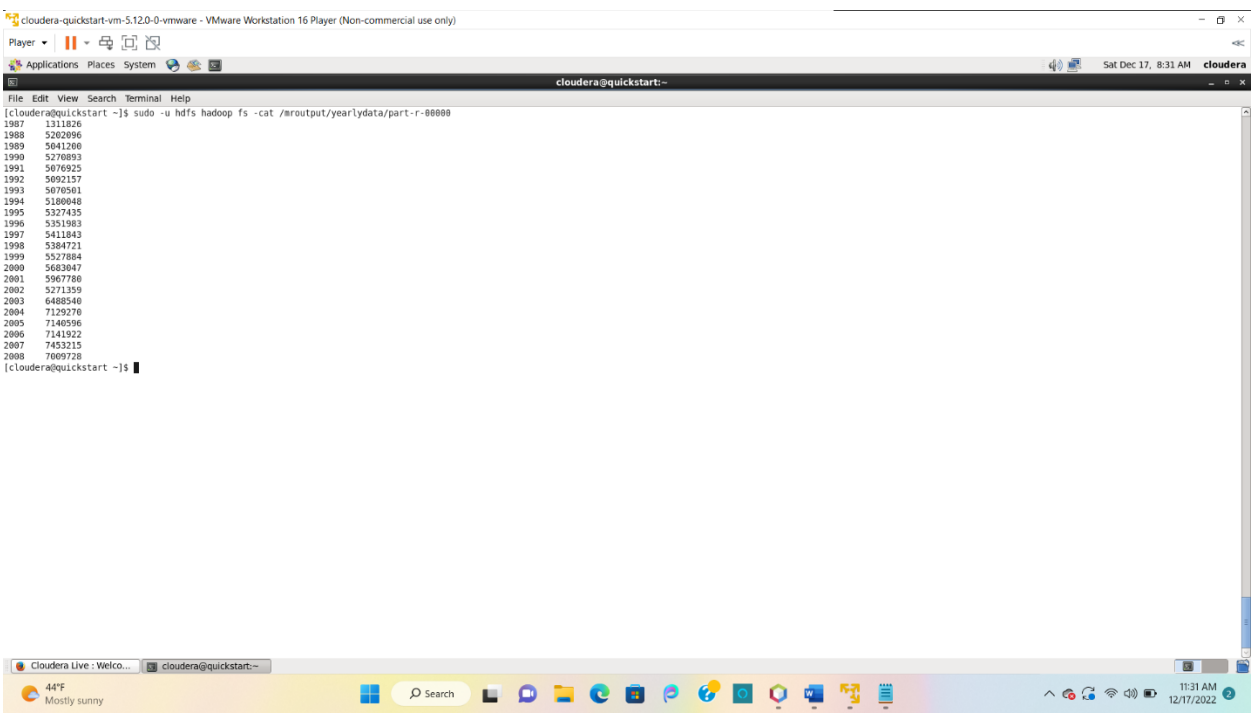


```
TUL-ORD : UA    10.6370
TUL-ORD : AA    12.0689
TUL-ORD : OO    17.0227
TUL-ORD : MQ    20.4600
TUL-ORD : DH    21.2374
TUL-PHX : WN    9.7065
TUL-PIA : OO    0.0000
TUL-PVU : OO    0.0000
TUL-RFD : OO    0.0000
TUL-SAN : XE    5.0121
TUL-SAT : XE    5.5736
TUL-SAT : AA    20.0113
TUL-SBN : AA    0.0000
TUL-SGF : TW    7.6322
TUL-SGF : UA    10.0764
TUL-SLC : OH    2.0718
TUL-SLC : DL    2.6787
TUL-SLC : OO    5.4867
TUL-SLC : EV    17.8218
TUL-SMF : XE    2.7664
TUL-SNA : AA    16.9706
TUL-STL : AA    0.6635
TUL-STL : TW    4.4181
TUL-STL : WN    10.8178
TUL-TUS : AA    22.4722
TUL-XNA : OO    0.0000
TUL-XNA : MQ    0.0000
TUL-XNA : EV    5.0990
TUP-ATL : EV    10.2725
TUS-ABQ : US    3.9171
TUS-ABQ : EA    5.1543
TUS-ABQ : WN    8.2587
TUS-ABQ : DL    8.9609
TUS-AMA : AA    11.7047
TUS-ATL : DL    5.5611
TUS-ATL : EA    12.0036
TUS-ATL : OO    17.5363
TUS-AUS : OO    0.0000
TUS-AUS : XE    11.0630
TUS-AUS : EV    12.8062
TUS-CDC : OO    0.0000
TUS-CLT : US    3.6032
TUS-COS : OO    0.0000
TUS-CVG : DL    2.9878
TUS-DEN : F9    1.9505
TUS-DEN : OO    4.6005
TUS-DEN : UA    6.1358
TUS-DEN : CO    7.3747
TUS-DFW : EV    4.5487
```

6) Top 30 source destination flights



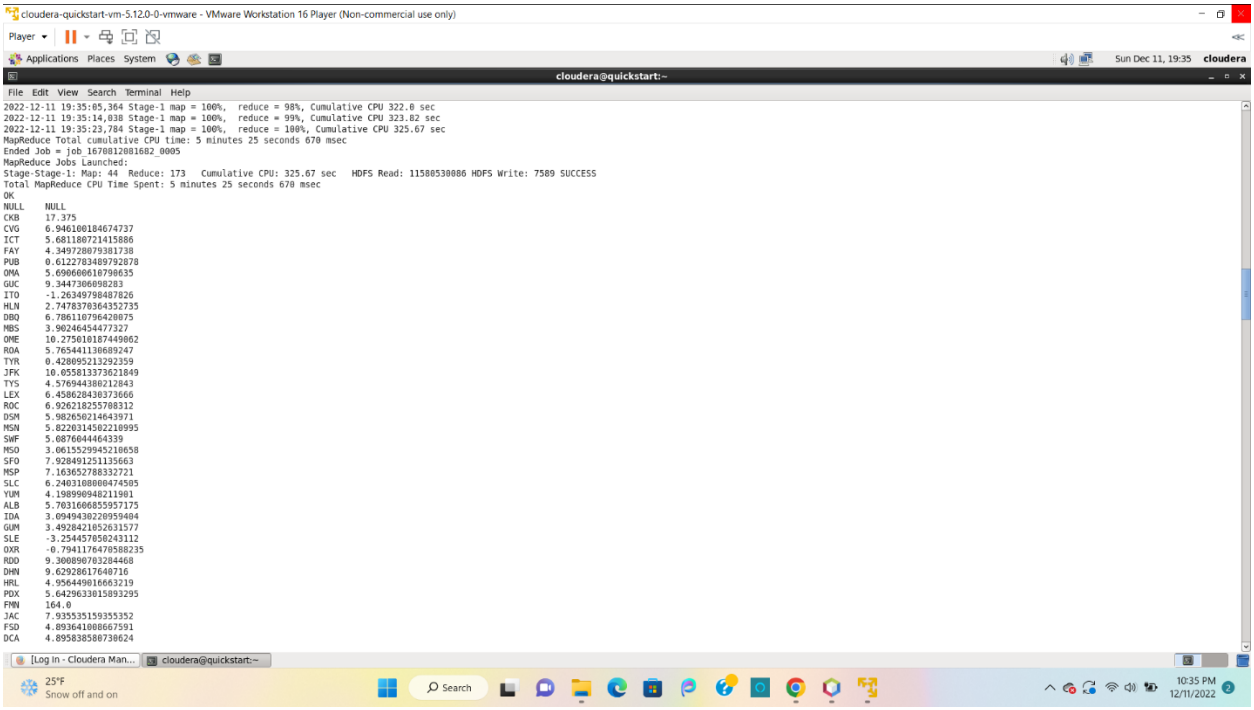
7)yearly data of flights



Hive Analysis on flight data:

1 Calculating the average delay for flights by origin.

Select origin,AVG(ArrDelay) AS avg_arr_delay FROM flightData GROUP BY Origin;



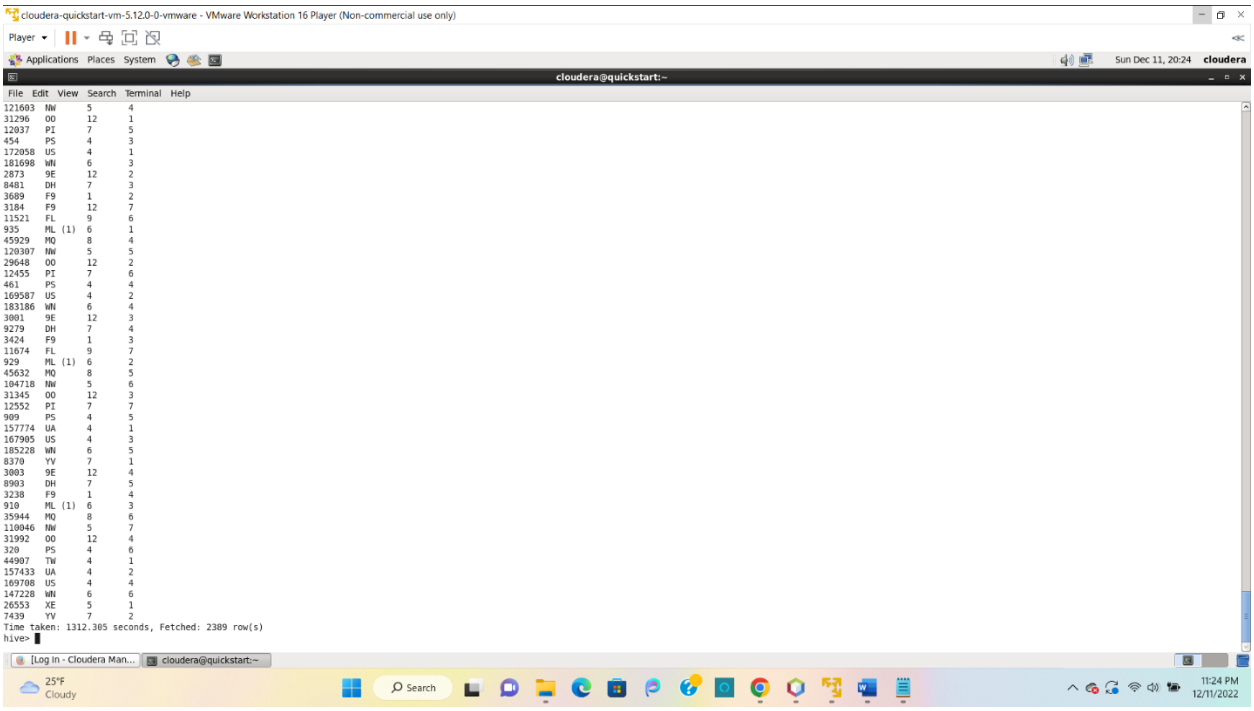
```
2022-12-11 19:35:05,364 Stage-1 map = 100%,  reduce = 98%, Cumulative CPU 322.0 sec
2022-12-11 19:35:14,038 Stage-1 map = 100%,  reduce = 99%, Cumulative CPU 323.82 sec
2022-12-11 19:35:23,784 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 325.67 sec
MapReduce Total cumulative CPU time: 5 minutes 25 seconds 670 msec
Ended Job = job_1670812081682_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 44  Reduce: 173   Cumulative CPU: 325.67 sec   HDFS Read: 11580530086 HDFS Write: 7589 SUCCESS
Total MapReduce CPU Time Spent: 5 minutes 25 seconds 670 msec
OK
NULL    NULL
CKB     17.375
CVG     6.946100184674737
ICT     5.681180721415886
FAY     4.349720079381738
PUB     0.6122783489792878
OMA     5.690600610790635
GUC     9.344730609828 3
ITO     -1.26349798487826
HLN     2.747837836432735
DBQ     6.786110796420075
MBS     3.90246454477327
OME     10.275010187449062
ROA     5.765441130689247
TYR     0.4280952132923 59
JFK     10.055813373621849
TYS     4.576944380212843
LEX     6.458628430373666
ROC     6.926218255708312
DSM     5.982650214643971
MSN     5.822031450221 0995
SWF     5.0876044464339
MSO     3.0615529945210658
SFO     7.928491251135663
MSP     7.163652788332721
SLC     6.240310800047450 5
YUM     4.198990948211981
ALB     5.703160685595717 5
IDA     3.094943022095940 4
GUM     3.492842105263157 7
SLE     -3.254457050243112
OXR     -0.7941176470588235
RDD     9.300890703284468
DHN     9.629280617640716
HRL     4.956449016663219
PDX     5.642963301589329 5
FMN     164.0
JAC     7.935535159355352
FSD     4.893641008667591
DCA     4.895838580730624
```

2 Calculating the average delay for flights by destination.

Select dest,AVG(ArrDelay) AS avg_arr_delay FROM flightData GROUP BY Origin;

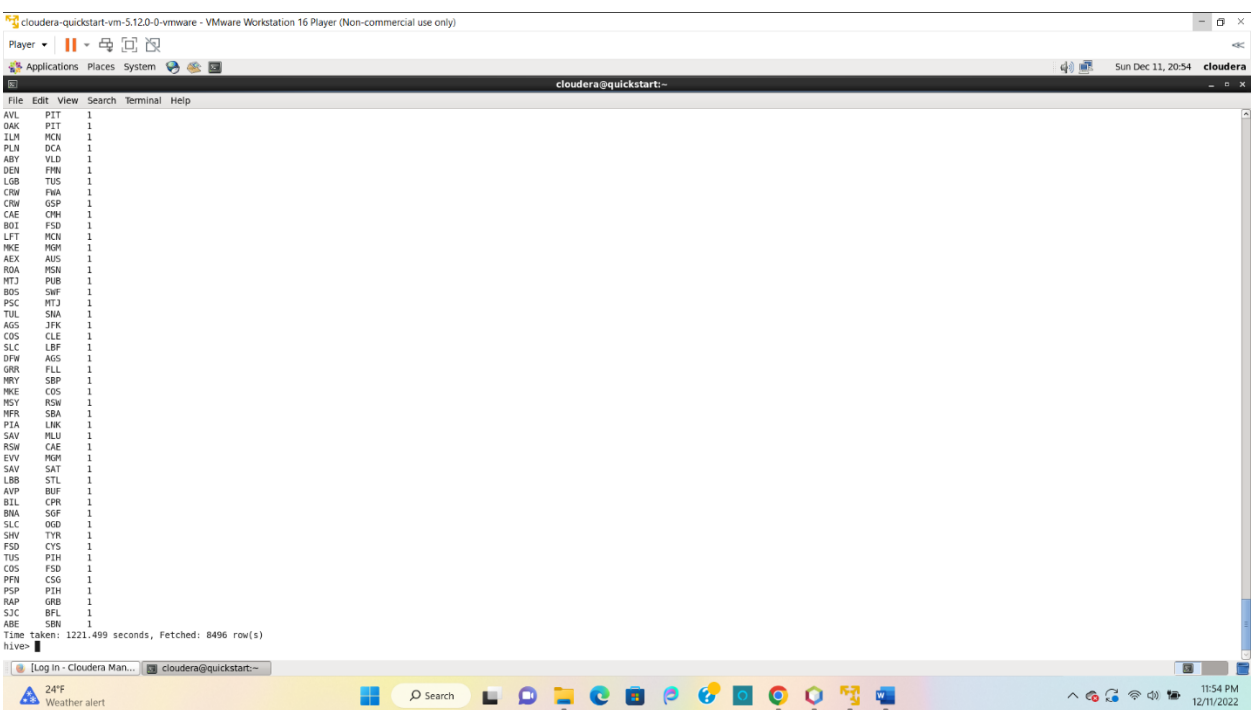
3 Counting the number offlights per carrier,per month, or per day of the week

SELECT COUNT(*) AS num_flights, UniqueCarrier, Month, DayOfWeek FROM flightData GROUP BY UniqueCarrier, Month, DayOfWeek;



4 Identifying the most common destinations and origins for flights

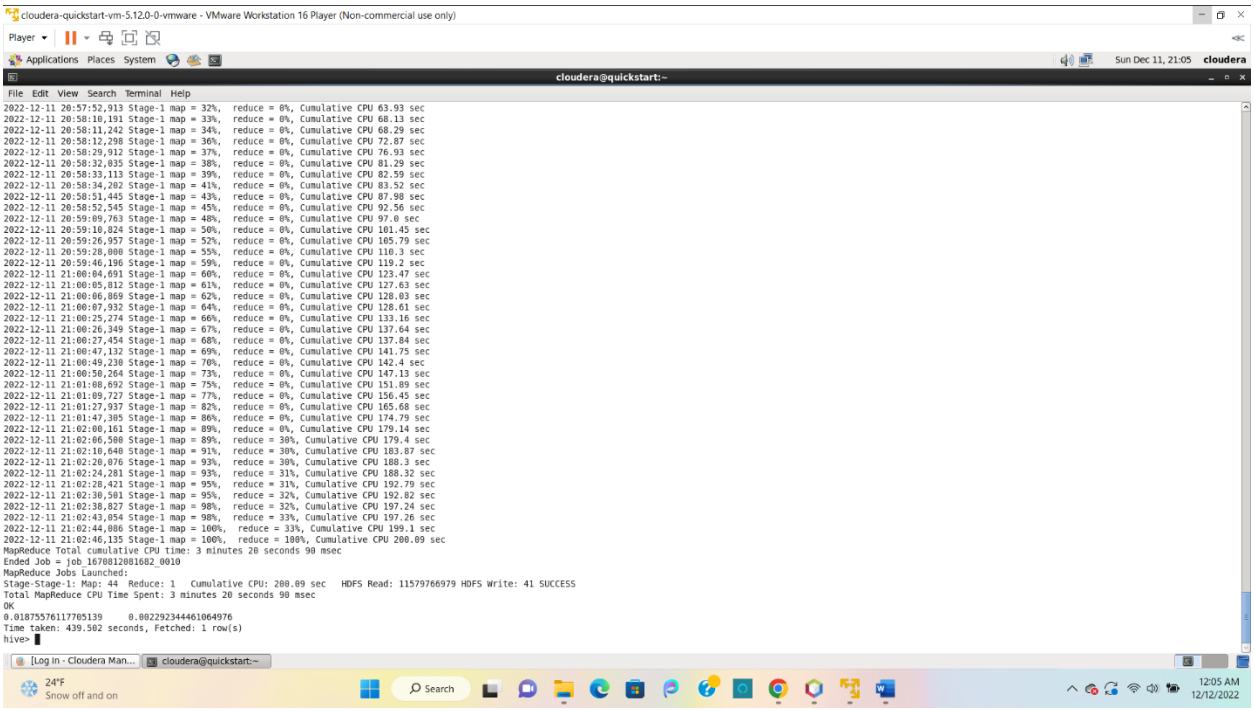
SELECT Origin, Dest, COUNT(*) AS num_flights FROM flightData GROUP BY Origin, Dest ORDER BY num_flights DESC;



5 Determining the percentage of cancelled or diverted flights

SELECT (COUNT(CASE WHEN Cancelled = 1 THEN 1 END) / COUNT(*)) AS cancelled_pct, (COUNT(CASE WHEN Diverted = 1 THEN 1 END) / COUNT(*)) AS diverted_pct

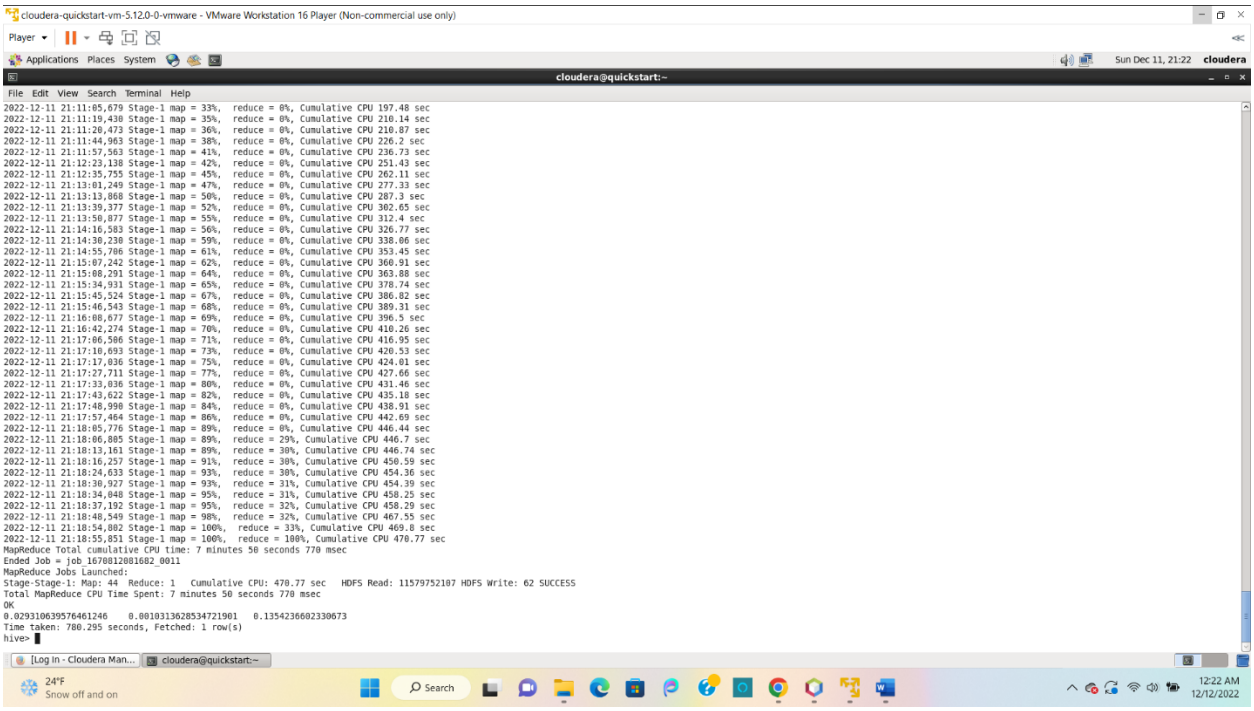
FROM flightData;

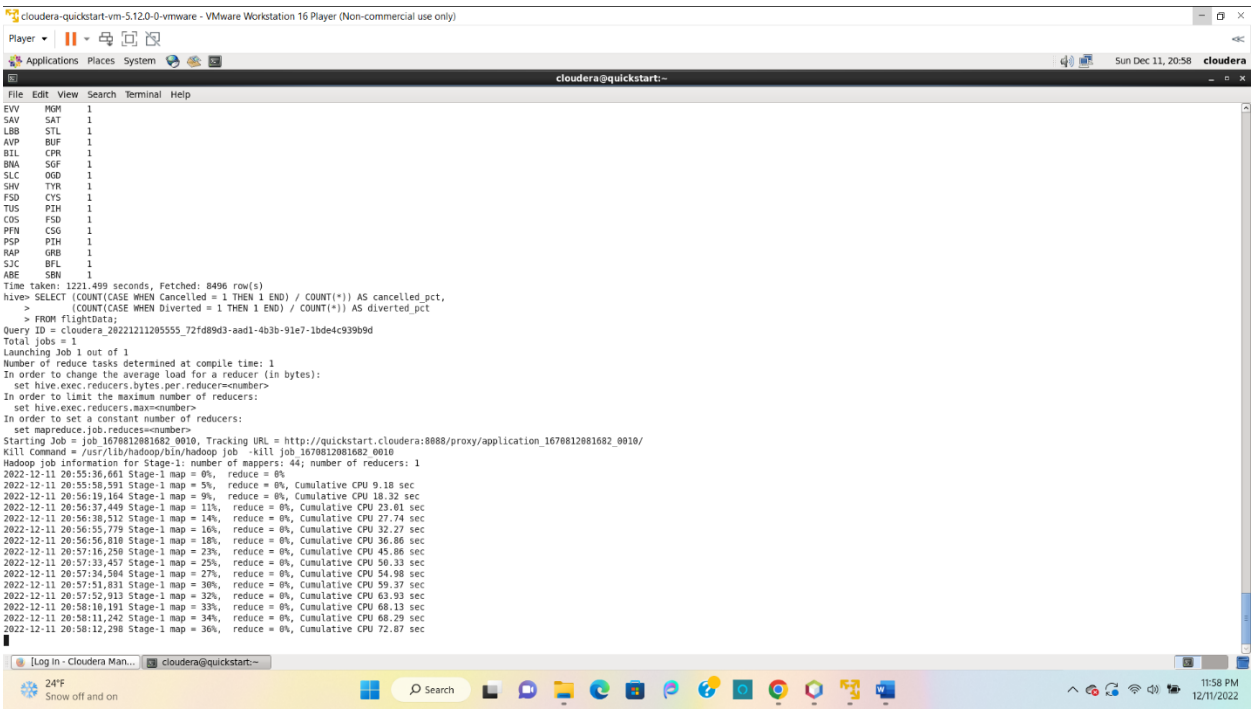


6 Analyzing the causes of flight delays, such as weather, security, or carrier delays

SELECT (SUM(WeatherDelay) / SUM(ArrDelay)) AS weather_delay_pct, (SUM(SecurityDelay) / SUM(ArrDelay)) AS security_delay_pct, (SUM(CarrierDelay) / SUM(ArrDelay)) AS carrier_delay_pct

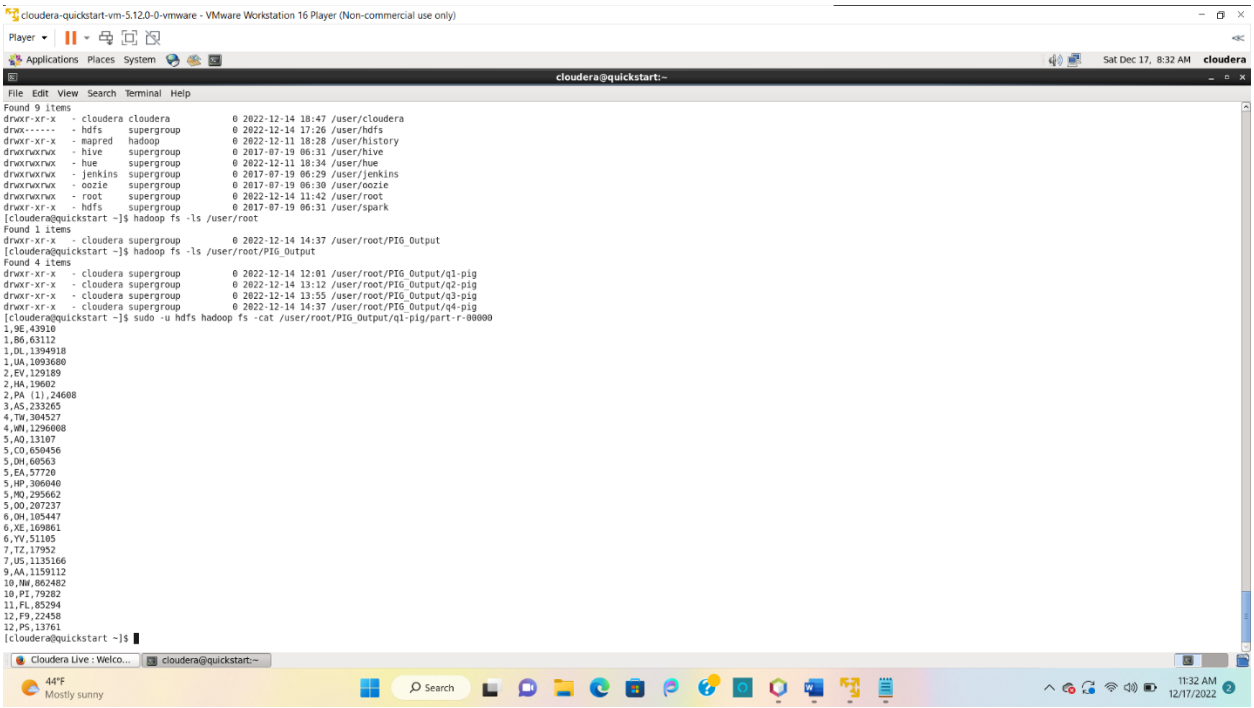
FROM flightData;





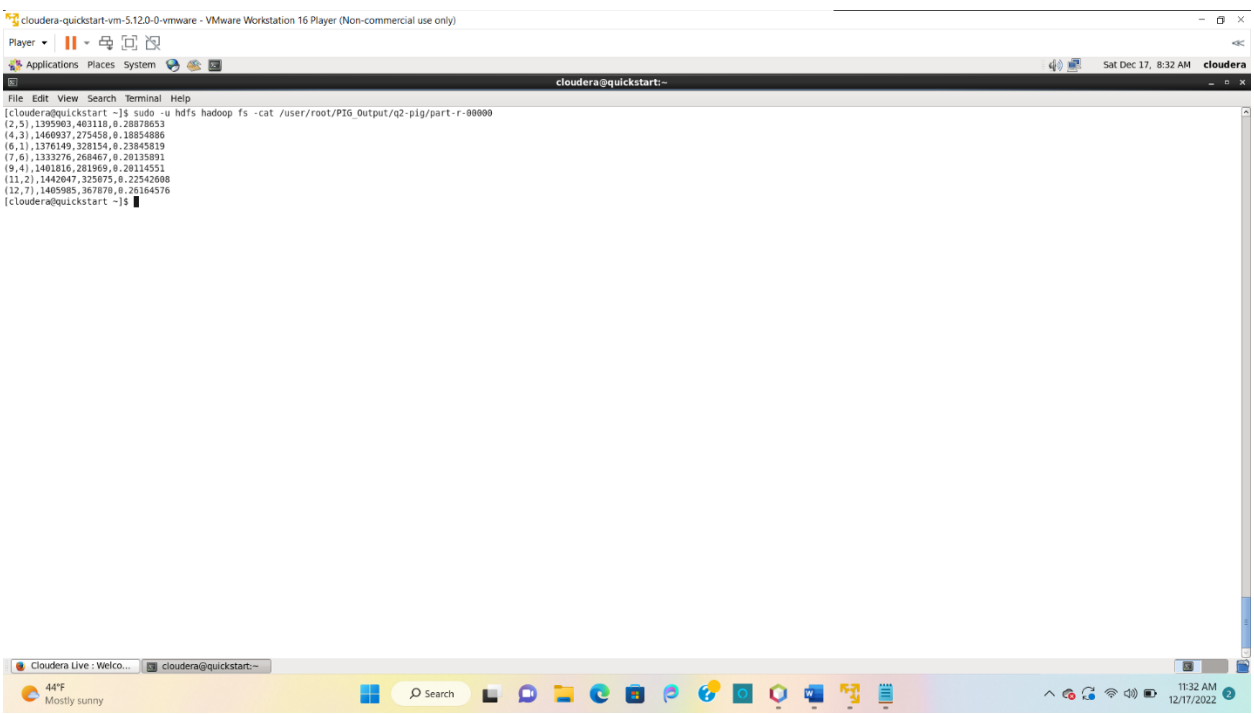
Pig Analysis:

Analysis 1: Carrier Popularity Computing the volume of total flights over each year, by carrier.



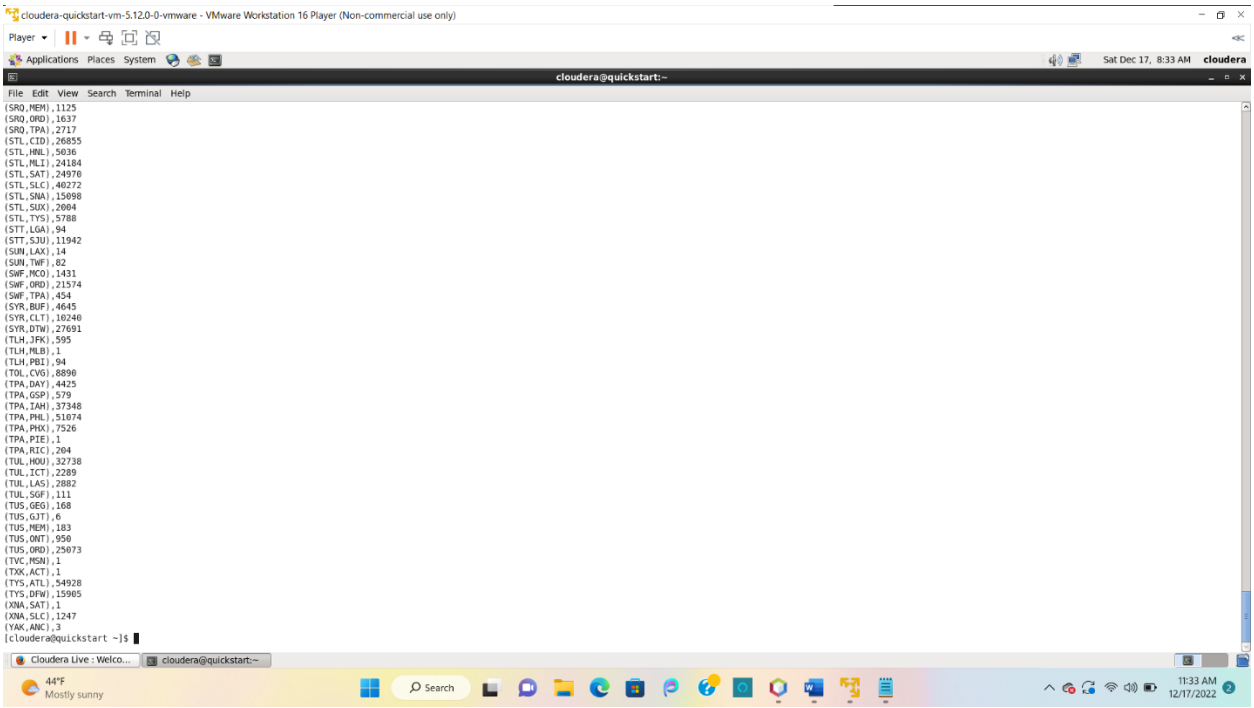
```
Found 9 items
drwxr-xr-x   - cloudera cloudera          0 2022-12-14 18:47 /user/cloudera
drwx------   - hdfs     supergroup        0 2022-12-14 17:26 /user/hdfs
drwxr-xr-x   - mapred   hadoop            0 2022-12-11 18:28 /user/history
drwxrwxrwx   - hive     supergroup        0 2017-07-19 06:31 /user/hive
drwxrwxrwx   - hue      supergroup        0 2022-12-11 18:34 /user/hue
drwxrwxrwx   - jenkins  supergroup        0 2017-07-19 06:29 /user/jenkins
drwxrwxrwx   - oozie    supergroup        0 2017-07-19 06:30 /user/oozie
drwxrwxrwx   - root     supergroup        0 2022-12-14 11:42 /user/root
drwxr-xr-x   - hdfs     supergroup        0 2017-07-19 06:31 /user/spark
[cloudera@quickstart ~]$ hadoop fs -ls /user/root
Found 1 items
drwxr-xr-x   - cloudera supergroup        0 2022-12-14 14:37 /user/root/PIG_Output
[cloudera@quickstart ~]$ hadoop fs -ls /user/root/PIG_Output
Found 4 items
drwxr-xr-x   - cloudera supergroup        0 2022-12-14 12:01 /user/root/PIG_Output/q1-pig
drwxr-xr-x   - cloudera supergroup        0 2022-12-14 13:12 /user/root/PIG_Output/q2-pig
drwxr-xr-x   - cloudera supergroup        0 2022-12-14 13:55 /user/root/PIG_Output/q3-pig
drwxr-xr-x   - cloudera supergroup        0 2022-12-14 14:37 /user/root/PIG_Output/q4-pig
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -cat /user/root/PIG_Output/q1-pig/part-r-00000
1,9E,43910
1,B6,63112
1,DL,1394918
1,UA,1093680
2,EV,129189
2,HA,19602
2,PA (1),24608
3,AS,233265
4,TW,304527
4,WN,1296008
5,AQ,13107
5,CO,650456
5,DH,60563
5,EA,57720
5,HP,306040
5,MQ,295662
5,OO,207237
6,OH,105447
6,XE,169861
6,YV,51105
7,TZ,17952
7,US,1135166
9,AA,1159112
10,NW,862482
10,PI,79282
11,FL,85294
12,F9,22458
12,PS,13761
[cloudera@quickstart ~]$
```

Analysis 2: Ratio of Flights Delayed.



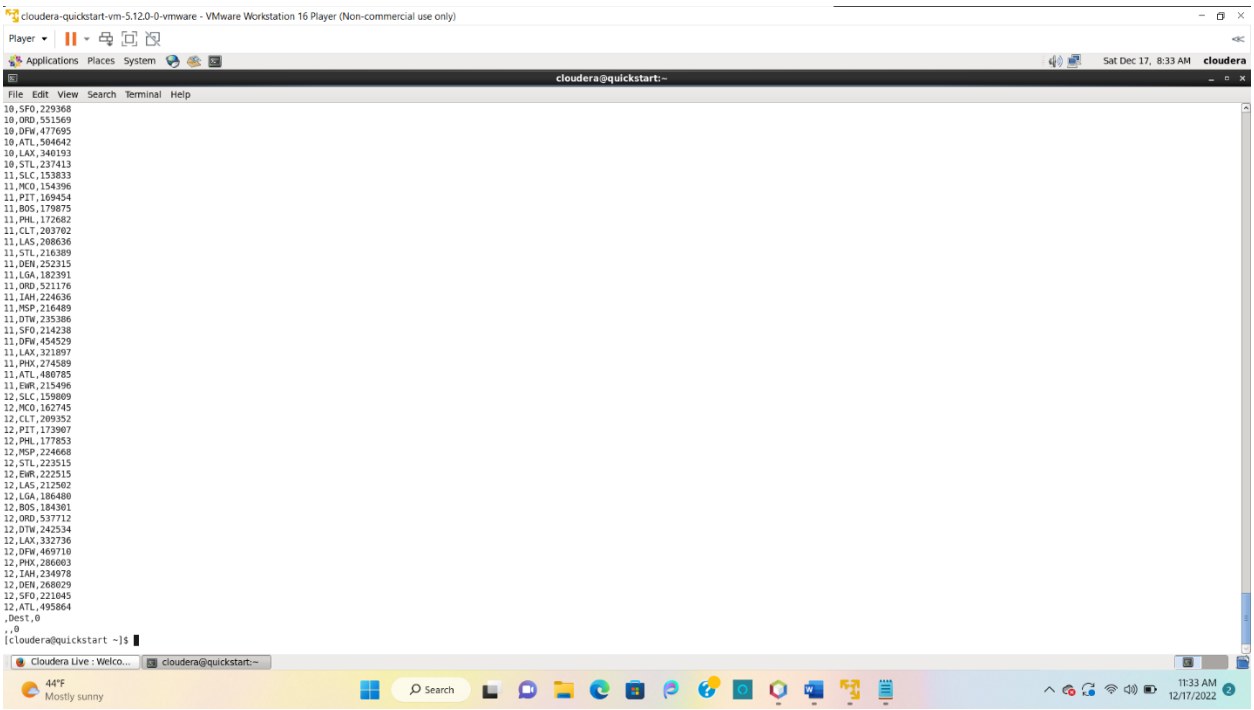
```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -cat /user/root/PIG_Output/q2-pig/part-r-00000
(2,5),1395903,403118,0.28878653
(4,3),1460937,275458,0.18854886
(6,1),1376149,328154,0.23845819
(7,6),1333276,268467,0.20135891
(9,4),1401816,281969,0.20114551
(11,2),1442047,325075,0.22542608
(12,7),1405985,367870,0.26164576
[cloudera@quickstart ~]$
```

Analysis 3: To find the most busy airports



```
(SRQ,MEM),1125
(SRQ,ORD),1637
(SRQ,TPA),2717
(STL,CID),26855
(STL,HNL),5036
(STL,MLI),24184
(STL,SAT),24970
(STL,SLC),40272
(STL,SNA),15098
(STL,SUX),2004
(STL,TYS),5788
(STT,LGA),94
(STT,SJU),11942
(SUN,LAX),14
(SUN,TWF),82
(SWF,MCO),1431
(SWF,ORD),21574
(SWF,TPA),454
(SYR,BUF),4645
(SYR,CLT),10240
(SYR,DTW),27691
(TLH,JFK),595
(TLH,MLB),1
(TLH,PBI),94
(TOL,CVG),8898
(TPA,DAY),4425
(TPA,GSP),579
(TPA,IAH),37348
(TPA,PHL),51074
(TPA,PHX),7526
(TPA,PIE),1
(TPA,RIC),204
(TUL,HOU),32738
(TUL,ICT),2289
(TUL,LAS),2882
(TUL,SGF),111
(TUS,GEG),168
(TUS,GJT),6
(TUS,MEM),183
(TUS,OMT),950
(TUS,ORD),25073
(TVC,MSN),1
(TXK,ACT),1
(TYS,ATL),54928
(TYS,DFW),15905
(XNA,SAT),1
(XNA,SLC),1247
(YAK,ANC),3
[cloudera@quickstart ~]$
```

Analysis 4: What are the busiest cities by total flight traffic?



```
10,SFO,229368
10,ORD,551569
10,DFW,477695
10,ATL,504642
10,LAX,340193
10,STL,237413
11,SLC,153833
11,MCO,154396
11,PIT,169454
11,BOS,179875
11,PHL,172682
11,CLT,203702
11,LAS,208636
11,STL,216389
11,DEN,252315
11,LGA,182391
11,ORD,521176
11,IAH,224636
11,MSP,216489
11,DTW,235386
11,SFO,214238
11,DFW,454529
11,LAX,321897
11,PHX,274589
11,ATL,480785
11,EWR,215496
12,SLC,159809
12,MCO,162745
12,CLT,209352
12,PIT,173907
12,PHL,177853
12,MSP,224668
12,STL,223515
12,EWR,222515
12,LAS,212502
12,LGA,186480
12,BOS,184301
12,ORD,537712
12,DTW,242534
12,LAX,332736
12,DFW,469710
12,PHX,286003
12,IAH,234978
12,DEN,268029
12,SFO,221045
12,ATL,495864
,Dest,0
,,0
[cloudera@quickstart ~]$
```

Summary:

To summarize this project can give the detailed analysis on the reason of flight delays. Performed different computations using mapreduce, hive and pig. Flight delays can be of any reason which be weather or security or carrier delay.