

ENGINEERING OF BIG-DATA SYSTEMS

Submitted by: Venkata Nitin Reddy Byreddy(002191804)

Overview of the dataset:

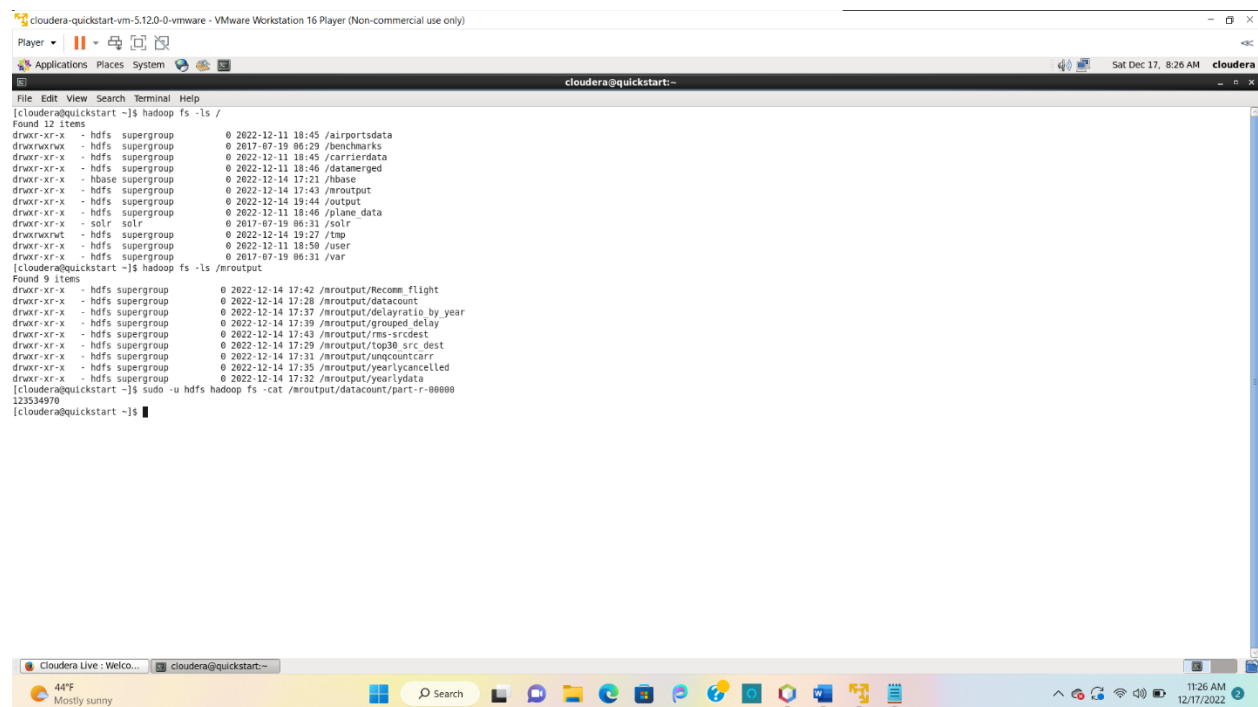
I used the Airline On time Statistics and Delay Causes dataset.

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.

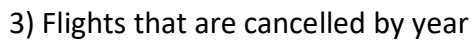
MapReduce Analysis on Flight Data:

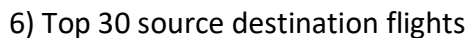
1)Counting the total number of records:



```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$ hadoop fs -ls /
Found 12 items
drwxr-xr-x - hdfs supergroup 0 2022-12-11 18:45 /airportsdata
drwxr-xr-x - hdfs supergroup 0 2017-07-19 08:29 /benchmarks
drwxr-xr-x - hdfs supergroup 0 2022-12-11 18:45 /carrierdata
drwxr-xr-x - hdfs supergroup 0 2022-12-11 18:46 /datamerged
drwxr-xr-x - hbase supergroup 0 2022-12-14 17:21 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:45 /mrooutput
drwxr-xr-x - hdfs supergroup 0 2022-12-14 19:44 /output
drwxr-xr-x - hdfs supergroup 0 2022-12-11 18:46 /plane_data
drwxr-xr-x - solr solr 0 2017-07-19 08:31 /solr
drwxr-xrwt - hdfs supergroup 0 2022-12-14 19:27 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-12-11 18:59 /user
drwxr-xr-x - hdfs supergroup 0 2017-07-19 08:31 /var
cloudera@quickstart:~$ hadoop fs -ls /mrooutput
Found 9 items
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:42 /mrooutput/Recomm_flight
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:28 /mrooutput/datacount
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:37 /mrooutput/delayratio_by_year
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:39 /mrooutput/grouped_delay
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:43 /mrooutput/ms-srcdest
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:29 /mrooutput/top30_src_dest
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:31 /mrooutput/unqcountcarr
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:35 /mrooutput/yearlycancelled
drwxr-xr-x - hdfs supergroup 0 2022-12-14 17:32 /mrooutput/yearlydata
cloudera@quickstart:~$ sudo -u hdfs hadoop fs -cat /mrooutput/datacount/part-r-00000
123534970
cloudera@quickstart:~$
```

2)Calculating the number of unique carrier flights





```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$ sudo -u hdfs hadoop fs -cat /mroutput/top30_src_dest/part-r-00000
SFO-LAX 338472
LAX-SFO 336038
LAX-LAS 292125
LAS-LAX 286328
PHX-LAX 278716
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 180032
DFW-IAH 180799
IAH-DFW 179036
cloudera@quickstart:~$
```

7)yearly data of flights

```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$ sudo -u hdfs hadoop fs -cat /mroutput/yearlydata/part-r-00000
1987 1311826
1988 5202096
1989 5841200
1990 5270893
1991 5076925
1992 5092157
1993 5070501
1994 5180048
1995 5327435
1996 5351983
1997 5411043
1998 5384721
1999 5527804
2000 5683047
2001 5967786
2002 5271359
2003 6488540
2004 7129270
2005 7140596
2006 7141922
2007 7453215
2008 7609728
cloudera@quickstart:~$
```

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;

The screenshot shows a terminal window titled 'cloudera@quickstart:~' with the following output:

```

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 323.02 sec
2022-12-11 19:35:23,794 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_1670812001682_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
CXB 17.375
CVG 6.946180184674737
ICT 5.681180721415886
FAY 4.349728079381738
PUB 0.6122783489792078
OMA 5.696606018798635
GUC 9.3447306098283
ITO -1.26349796487626
HLN 2.7478376364352735
DBQ 6.786110796428075
MBS 3.90246454477327
ONE 10.275010187448902
ROA 5.765441138689247
TYR 0.428095213292359
JFK 10.055813373921049
TYS 4.576944388212843
LEX 6.458628430373666
ROC 6.926218255708312
DSM 5.982650214643971
MSN 5.8220314502218995
SWF 5.0876044464339
HSD 3.0815529945218058
SFO 7.928491251135663
HSP 7.163652788332721
SLC 6.2403108080474505
YUM 4.198990948211901
ALB 5.7031606855957175
IDA 3.0949430220959404
GUM 3.4928421852631577
SLE -3.254457858243112
DOR -0.794117647958235
RDO 9.300890703284468
DAN 9.62928617648716
HRL 4.956449016663219
POX 5.642963301503295
FNN 164.0
JAC 7.935535159355352
FSD 4.89364108667551
DCA 4.89583588738624

```

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;

```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$
File Edit View Search Terminal Help
121603 NW 5 4
12296 OO 12 1
12037 PT 7 5
454 PS 4 3
172058 US 4 1
101098 WI 6 3
2873 9E 12 2
8481 DH 7 3
3069 F9 1 2
3184 F9 12 7
11521 FL 9 6
935 ML (1) 6 1
45929 MO 8 4
120307 NW 5 5
29648 OO 12 2
12455 PT 7 6
461 PS 4 4
109507 US 4 2
183186 WI 6 4
3001 9E 12 3
9279 DH 7 4
3424 F9 1 3
11674 FL 9 7
929 ML (1) 6 2
45632 MO 8 5
104718 NW 5 6
31345 OO 12 3
12552 PT 7 7
909 PS 4 5
157774 UA 4 1
107905 US 4 3
105228 WI 6 5
8370 YV 7 1
3003 9E 12 4
8903 DH 7 5
3238 F9 1 4
910 ML (1) 6 3
35944 MO 8 6
110046 NW 5 7
31992 OO 12 4
320 PS 4 6
44907 TW 4 1
157433 UA 4 2
109708 US 4 4
147228 WI 6 6
26553 XE 5 1
7439 YV 7 2
Time taken: 1312.385 seconds, Fetched: 2389 row(s)
hive>
```

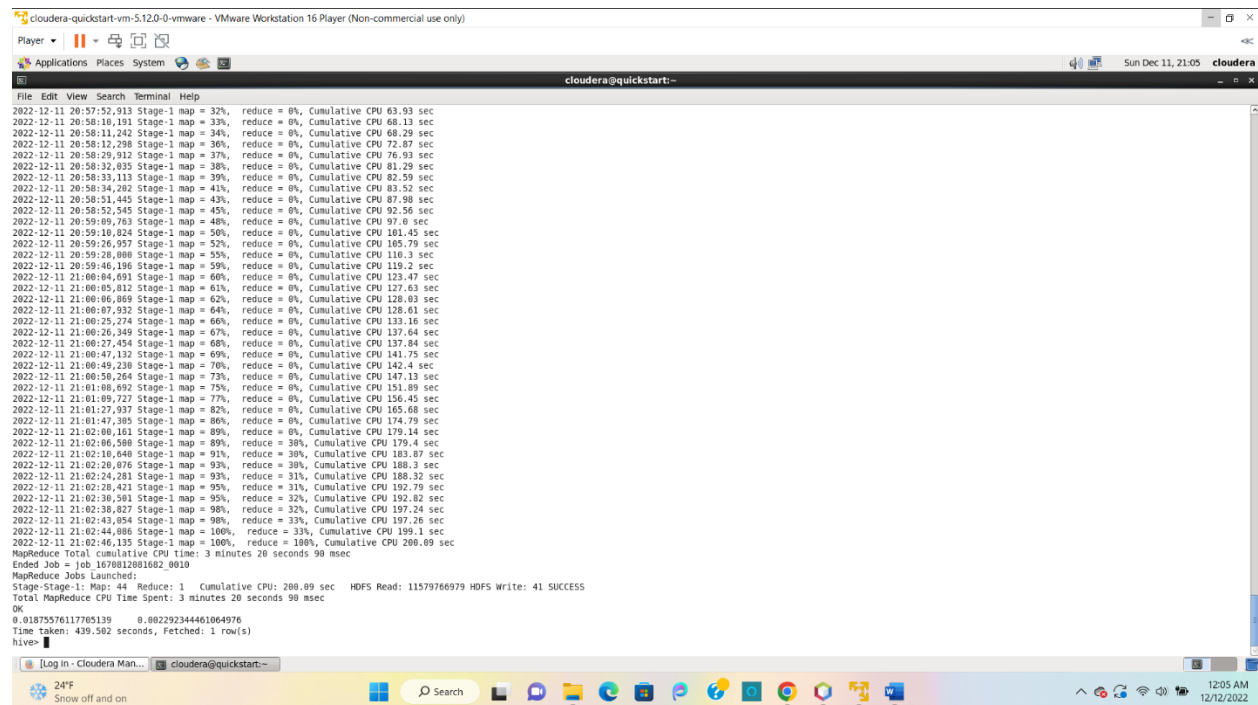
4 Identifying the most common destinations and origins for flights

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

```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$
File Edit View Search Terminal Help
AVL PIT 1
DAK PIT 1
LHM MCO 1
PLN DCA 1
ABY VLD 1
DEN FWA 1
LGB TUS 1
CRW FWA 1
CRW GSP 1
CAE OPM 1
BOI FSD 1
LFT MCO 1
MKE MGN 1
AEX AUS 1
ROA MGN 1
HTJ PUB 1
BOS SWF 1
PSC HTJ 1
TUL SNA 1
AGS JFK 1
COS CLE 1
SLC LBF 1
DFW AGS 1
GRR FLL 1
HRY SBP 1
MKE COS 1
MSY RSW 1
HFA SBA 1
PTA LHR 1
SAV MLU 1
RSW CAE 1
EVV MGR 1
SAV SAT 1
LBB STL 1
AVP BUF 1
BIL CPR 1
BNA SDF 1
SLC OGD 1
SHV TYR 1
FSD CYS 1
TUS PIH 1
COS FSD 1
PFN CSG 1
PSP PIH 1
RAP ORB 1
SJC BFL 1
ABE SBN 1
Time taken: 1221.499 seconds, Fetched: 8496 row(s)
hive>
```

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;
```



The screenshot shows a terminal window titled "cloudera@quickstart:~" with a list of Hadoop MapReduce job progress logs. Each line represents a task completion, showing the timestamp, stage, map/reduce status, and cumulative CPU time. The logs show a sequence of tasks from 2022-12-11 20:57:52 to 21:02:46. The final summary indicates a successful job completion with a total CPU time of 3 minutes 20 seconds 90 msec.

```
2022-12-11 20:57:52,913 Stage-1 map = 32%, reduce = 0%, Cumulative CPU 63.93 sec  
2022-12-11 20:58:10,191 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 68.13 sec  
2022-12-11 20:58:11,242 Stage-1 map = 34%, reduce = 0%, Cumulative CPU 68.29 sec  
2022-12-11 20:58:12,298 Stage-1 map = 36%, reduce = 0%, Cumulative CPU 72.87 sec  
2022-12-11 20:58:29,912 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 76.93 sec  
2022-12-11 20:58:32,035 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 81.29 sec  
2022-12-11 20:58:33,113 Stage-1 map = 39%, reduce = 0%, Cumulative CPU 82.59 sec  
2022-12-11 20:58:34,202 Stage-1 map = 41%, reduce = 0%, Cumulative CPU 83.52 sec  
2022-12-11 20:58:51,445 Stage-1 map = 43%, reduce = 0%, Cumulative CPU 87.98 sec  
2022-12-11 20:58:52,545 Stage-1 map = 45%, reduce = 0%, Cumulative CPU 92.56 sec  
2022-12-11 20:58:59,763 Stage-1 map = 46%, reduce = 0%, Cumulative CPU 97.6 sec  
2022-12-11 20:59:10,824 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 101.45 sec  
2022-12-11 20:59:26,957 Stage-1 map = 52%, reduce = 0%, Cumulative CPU 105.79 sec  
2022-12-11 20:59:29,088 Stage-1 map = 55%, reduce = 0%, Cumulative CPU 110.3 sec  
2022-12-11 20:59:46,196 Stage-1 map = 59%, reduce = 0%, Cumulative CPU 119.2 sec  
2022-12-11 21:00:04,691 Stage-1 map = 60%, reduce = 0%, Cumulative CPU 123.47 sec  
2022-12-11 21:00:05,812 Stage-1 map = 61%, reduce = 0%, Cumulative CPU 127.63 sec  
2022-12-11 21:00:06,869 Stage-1 map = 62%, reduce = 0%, Cumulative CPU 128.03 sec  
2022-12-11 21:00:07,932 Stage-1 map = 64%, reduce = 0%, Cumulative CPU 128.61 sec  
2022-12-11 21:00:25,274 Stage-1 map = 66%, reduce = 0%, Cumulative CPU 133.16 sec  
2022-12-11 21:00:26,349 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 137.64 sec  
2022-12-11 21:00:27,454 Stage-1 map = 68%, reduce = 0%, Cumulative CPU 137.84 sec  
2022-12-11 21:00:47,132 Stage-1 map = 69%, reduce = 0%, Cumulative CPU 141.75 sec  
2022-12-11 21:00:49,238 Stage-1 map = 70%, reduce = 0%, Cumulative CPU 142.4 sec  
2022-12-11 21:00:50,264 Stage-1 map = 73%, reduce = 0%, Cumulative CPU 147.13 sec  
2022-12-11 21:01:08,692 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 151.89 sec  
2022-12-11 21:01:09,727 Stage-1 map = 77%, reduce = 0%, Cumulative CPU 156.45 sec  
2022-12-11 21:01:27,937 Stage-1 map = 82%, reduce = 0%, Cumulative CPU 165.68 sec  
2022-12-11 21:01:47,385 Stage-1 map = 86%, reduce = 0%, Cumulative CPU 174.79 sec  
2022-12-11 21:02:00,161 Stage-1 map = 89%, reduce = 0%, Cumulative CPU 179.14 sec  
2022-12-11 21:02:06,580 Stage-1 map = 89%, reduce = 30%, Cumulative CPU 179.4 sec  
2022-12-11 21:02:10,648 Stage-1 map = 91%, reduce = 30%, Cumulative CPU 183.87 sec  
2022-12-11 21:02:20,076 Stage-1 map = 93%, reduce = 30%, Cumulative CPU 188.3 sec  
2022-12-11 21:02:24,281 Stage-1 map = 93%, reduce = 31%, Cumulative CPU 188.32 sec  
2022-12-11 21:02:28,421 Stage-1 map = 95%, reduce = 31%, Cumulative CPU 192.79 sec  
2022-12-11 21:02:30,591 Stage-1 map = 95%, reduce = 32%, Cumulative CPU 192.82 sec  
2022-12-11 21:02:38,827 Stage-1 map = 98%, reduce = 32%, Cumulative CPU 197.24 sec  
2022-12-11 21:02:43,054 Stage-1 map = 98%, reduce = 33%, Cumulative CPU 197.26 sec  
2022-12-11 21:02:44,886 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 199.1 sec  
2022-12-11 21:02:46,135 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 200.89 sec  
MapReduce Total cumulative CPU time: 3 minutes 20 seconds 90 msec  
Ended Job = job_1670812081062_0010  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 44 Reduce: 1 Cumulative CPU: 200.89 sec HDFS Read: 11579766979 HDFS Write: 41 SUCCESS  
Total MapReduce CPU Time Spent: 3 minutes 20 seconds 90 msec  
OK  
0.01875576117705139 0.002292344461064976  
Time taken: 439.562 seconds, Fetched: 1 row(s)  
hive>
```

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

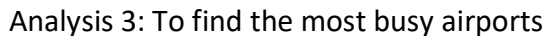
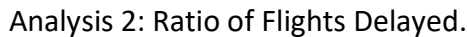
```
SELECT (SUM(WeatherDelay) / SUM(ArrDelay)) AS weather_delay_pct, (SUM(SecurityDelay) /  
SUM(ArrDelay)) AS security_delay_pct, (SUM(CarrierDelay) / SUM(ArrDelay)) AS  
carrier_delay_pct  
FROM flightData;
```

```
cloudera-quickstart-vm-5.120-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$
File Edit View Search Terminal Help
2022-12-11 21:11:05,679 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 197.48 sec
2022-12-11 21:11:19,438 Stage-1 map = 35%, reduce = 0%, Cumulative CPU 210.14 sec
2022-12-11 21:11:26,473 Stage-1 map = 36%, reduce = 0%, Cumulative CPU 210.87 sec
2022-12-11 21:11:44,963 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 226.2 sec
2022-12-11 21:11:57,563 Stage-1 map = 41%, reduce = 0%, Cumulative CPU 236.73 sec
2022-12-11 21:12:23,138 Stage-1 map = 42%, reduce = 0%, Cumulative CPU 251.43 sec
2022-12-11 21:12:35,755 Stage-1 map = 45%, reduce = 0%, Cumulative CPU 262.11 sec
2022-12-11 21:13:01,249 Stage-1 map = 47%, reduce = 0%, Cumulative CPU 277.33 sec
2022-12-11 21:13:13,868 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 287.3 sec
2022-12-11 21:13:39,377 Stage-1 map = 52%, reduce = 0%, Cumulative CPU 302.65 sec
2022-12-11 21:13:56,877 Stage-1 map = 55%, reduce = 0%, Cumulative CPU 312.4 sec
2022-12-11 21:14:16,583 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 326.77 sec
2022-12-11 21:14:38,230 Stage-1 map = 59%, reduce = 0%, Cumulative CPU 338.06 sec
2022-12-11 21:14:55,786 Stage-1 map = 61%, reduce = 0%, Cumulative CPU 353.45 sec
2022-12-11 21:15:07,242 Stage-1 map = 62%, reduce = 0%, Cumulative CPU 366.91 sec
2022-12-11 21:15:08,291 Stage-1 map = 64%, reduce = 0%, Cumulative CPU 363.88 sec
2022-12-11 21:15:34,931 Stage-1 map = 65%, reduce = 0%, Cumulative CPU 378.74 sec
2022-12-11 21:15:45,524 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 386.82 sec
2022-12-11 21:15:46,543 Stage-1 map = 68%, reduce = 0%, Cumulative CPU 389.31 sec
2022-12-11 21:16:08,677 Stage-1 map = 69%, reduce = 0%, Cumulative CPU 396.5 sec
2022-12-11 21:16:42,274 Stage-1 map = 70%, reduce = 0%, Cumulative CPU 410.26 sec
2022-12-11 21:17:06,586 Stage-1 map = 71%, reduce = 0%, Cumulative CPU 416.95 sec
2022-12-11 21:17:18,693 Stage-1 map = 73%, reduce = 0%, Cumulative CPU 426.53 sec
2022-12-11 21:17:17,036 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 424.01 sec
2022-12-11 21:17:27,711 Stage-1 map = 77%, reduce = 0%, Cumulative CPU 427.66 sec
2022-12-11 21:17:33,036 Stage-1 map = 80%, reduce = 0%, Cumulative CPU 431.46 sec
2022-12-11 21:17:43,622 Stage-1 map = 82%, reduce = 0%, Cumulative CPU 435.18 sec
2022-12-11 21:17:46,998 Stage-1 map = 84%, reduce = 0%, Cumulative CPU 438.91 sec
2022-12-11 21:17:57,464 Stage-1 map = 86%, reduce = 0%, Cumulative CPU 442.69 sec
2022-12-11 21:18:05,776 Stage-1 map = 89%, reduce = 0%, Cumulative CPU 446.44 sec
2022-12-11 21:18:06,885 Stage-1 map = 89%, reduce = 29%, Cumulative CPU 446.7 sec
2022-12-11 21:18:13,161 Stage-1 map = 89%, reduce = 30%, Cumulative CPU 446.74 sec
2022-12-11 21:18:16,257 Stage-1 map = 91%, reduce = 30%, Cumulative CPU 450.59 sec
2022-12-11 21:18:24,633 Stage-1 map = 93%, reduce = 30%, Cumulative CPU 454.36 sec
2022-12-11 21:18:30,507 Stage-1 map = 93%, reduce = 31%, Cumulative CPU 454.39 sec
2022-12-11 21:18:34,848 Stage-1 map = 95%, reduce = 31%, Cumulative CPU 458.25 sec
2022-12-11 21:18:37,192 Stage-1 map = 95%, reduce = 32%, Cumulative CPU 458.29 sec
2022-12-11 21:18:40,549 Stage-1 map = 98%, reduce = 32%, Cumulative CPU 467.55 sec
2022-12-11 21:18:54,892 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 469.8 sec
2022-12-11 21:18:55,851 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 470.77 sec
MapReduce Total cumulative CPU time: 7 minutes 56 seconds 770 msec
Ended Job = job_1670812081682_0010
MapReduce Jobs Launched:
Stage-1 Stage-1: Map: 44 Reduce: 1 Cumulative CPU: 470.77 sec HDFS Read: 11579752107 HDFS Write: 62 SUCCESS
Total MapReduce CPU Time Spent: 7 minutes 56 seconds 770 msec
OK
0.029310639576461246 0.0010313628534721901 0.1354236602330673
Time taken: 780.295 seconds, Fetched: 1 row(s)
hive>
```

```
cloudera-quickstart-vm-5.120-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$
File Edit View Search Terminal Help
EVV MFM 1
SAV SAT 1
LBB STL 1
AVP BUF 1
BIL CPR 1
BNA SGF 1
SLC OGD 1
SHV TYR 1
FSD CYS 1
TUS PIH 1
COS FSD 1
PFM CSG 1
PSP PIH 1
MAP GMB 1
SJC BFL 1
ABE SBN 1
Time taken: 1221.499 seconds, Fetched: 8496 row(s)
hive> 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;
Query ID = cloudera_20221211205555_72fd89d3-aad1-4b3b-91e7-1bde4c93909d
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.reducers=<number>
Starting Job = job_1670812081682_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1670812081682_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1670812081682_0010
Hadoop job information for Stage-1: number of mappers: 44; number of reducers: 1
2022-12-11 20:55:36,661 Stage-1 map = 0%, reduce = 0%
2022-12-11 20:55:58,591 Stage-1 map = 5%, reduce = 0%, Cumulative CPU 9.18 sec
2022-12-11 20:56:19,164 Stage-1 map = 9%, reduce = 0%, Cumulative CPU 18.32 sec
2022-12-11 20:56:37,449 Stage-1 map = 11%, reduce = 0%, Cumulative CPU 23.01 sec
2022-12-11 20:56:38,512 Stage-1 map = 14%, reduce = 0%, Cumulative CPU 27.74 sec
2022-12-11 20:56:55,779 Stage-1 map = 16%, reduce = 0%, Cumulative CPU 32.27 sec
2022-12-11 20:56:56,018 Stage-1 map = 18%, reduce = 0%, Cumulative CPU 36.86 sec
2022-12-11 20:57:16,250 Stage-1 map = 23%, reduce = 0%, Cumulative CPU 45.86 sec
2022-12-11 20:57:33,457 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 50.33 sec
2022-12-11 20:57:34,584 Stage-1 map = 27%, reduce = 0%, Cumulative CPU 54.90 sec
2022-12-11 20:57:51,831 Stage-1 map = 30%, reduce = 0%, Cumulative CPU 59.37 sec
2022-12-11 20:57:52,913 Stage-1 map = 32%, reduce = 0%, Cumulative CPU 63.93 sec
2022-12-11 20:58:10,191 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 68.13 sec
2022-12-11 20:58:11,242 Stage-1 map = 34%, reduce = 0%, Cumulative CPU 68.29 sec
2022-12-11 20:58:12,298 Stage-1 map = 36%, reduce = 0%, Cumulative CPU 72.87 sec
```

Pig Analysis:

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



```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$
File Edit View Search Terminal Help
(SFO, MEM), 1125
(SFO, ORD), 1637
(SFO, TPA), 2717
(STL, CID), 26855
(STL, HLL), 5636
(STL, MLI), 24184
(STL, SAT), 24976
(STL, SLC), 46272
(STL, SNA), 15998
(STL, SUX), 2004
(STL, TYS), 5788
(STT, LGA), 94
(STT, SJU), 11942
(SUR, LAX), 14
(SUR, TWF), 62
(SWF, PCO), 1431
(SWF, ORD), 21574
(SWF, TPA), 454
(SYR, BUF), 4645
(SYR, CLT), 16240
(SYR, DTW), 27691
(TLH, JFK), 595
(TLH, MLB), 1
(TLH, PBI), 94
(TOL, CVG), 8896
(TPA, DAY), 4425
(TPA, GSP), 579
(TPA, IAH), 37348
(TPA, PHL), 51074
(TPA, PIK), 7526
(TPA, PTE), 1
(TPA, RJC), 204
(TUL, HOU), 32738
(TUL, TCT), 2289
(TUL, LAS), 2882
(TUL, SFO), 111
(TUS, GEG), 168
(TUS, GJT), 6
(TUS, MEM), 183
(TUS, ONT), 950
(TUS, ORD), 25073
(TVC, MSN), 1
(TXK, ACT), 1
(TYS, ATL), 54928
(TYS, DTW), 15965
(XNA, SAT), 1
(XNA, SLC), 1247
(YAK, ANC), 3
cloudera@quickstart ~$
```

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

```
cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places System
cloudera@quickstart:~$
File Edit View Search Terminal Help
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, 203762
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, 321097
11, PHX, 274589
11, ATL, 480785
11, EWR, 215496
12, SLC, 159809
12, MCO, 162745
12, CLT, 209352
12, PIT, 173987
12, PHL, 177853
12, MSP, 224668
12, STL, 223515
12, EWR, 222515
12, LAS, 212562
12, LGA, 186480
12, BOS, 184361
12, ORD, 537712
12, DTW, 242534
12, LAX, 332736
12, DFW, 469710
12, PHX, 286003
12, IAH, 234978
12, DEN, 268629
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.