INFO 7250: Final Project
Summer 2020
MIT KATWALA: 001440383

DataSet: Airline On-Time Statistics and Delay Causes (JUNE 2003 – April 2020)

1) About the dataset:

Everyday there are millions of flights that fly between different airports in the United States. Many flights are delayed, canceled or diverted. It will be of interest to any airline business/airports to study and analyze the reasons why flights are delayed and how to improve travel services

This dataset have the following variables:

Column	Description
year	Year (yyyy)
month	Month (mm)
carrier	Airline carrier abbreviation
carrier_name	Carrier name
airport	Airport code
airport_name	Airport full name
arr_flights	Total Number of arriving flights / record
arr_del15	Total Number of delayed flights / record
carrier_ct	Total number of delayed flights because of air carrier (subset of arr_del15)
weather_ct	Total number of delayed flights because of weather (subset of arr_del15)
nas_ct	Total number of delayed flights because of national aviation system carrier (subset of arr_del15)
security_ct	Total number of delayed flights because of airport security (subset of arr_del15)
late_aircraft_ct	Total number of delayed flights because of previous delayed flight using same aircraft (subset of arr_del15)

arr_cancelled	Number of canceled flights
arr_diverted	Number of flights diverted
arr_delay	Arrival delay in minutes
carrier_delay	Carrier delay in minutes (subset of arr_delay)
weather_delay	Weather delay in minutes (subset of arr_delay)
nas_delay	National aviation system delay in minutes (subset of arr_delay)
security_delay	Security delay in minutes (subset of arr_delay)
late_aircraft_delay	Previous Aircraft delay in minutes (subset of arr_delay)

2) Analysis to be performed:

a) What is the biggest contribution of delays for every airline? (counter)
b) Top 10 most punctual flights (MR chaining)?
c) Busiest airport in USA? (Secondary sort by year)
d) Number of canceled flights / carrier? (Hive)
e) Best month for travel with least weather related delays? (Hive)
f) Which airport had the most NAS delays? (Pig)
g) Which carrier had the least number of diverted flights? (mongoDb)

3) Analysis:

a) What is the biggest contribution of delays for every airline? (counter)

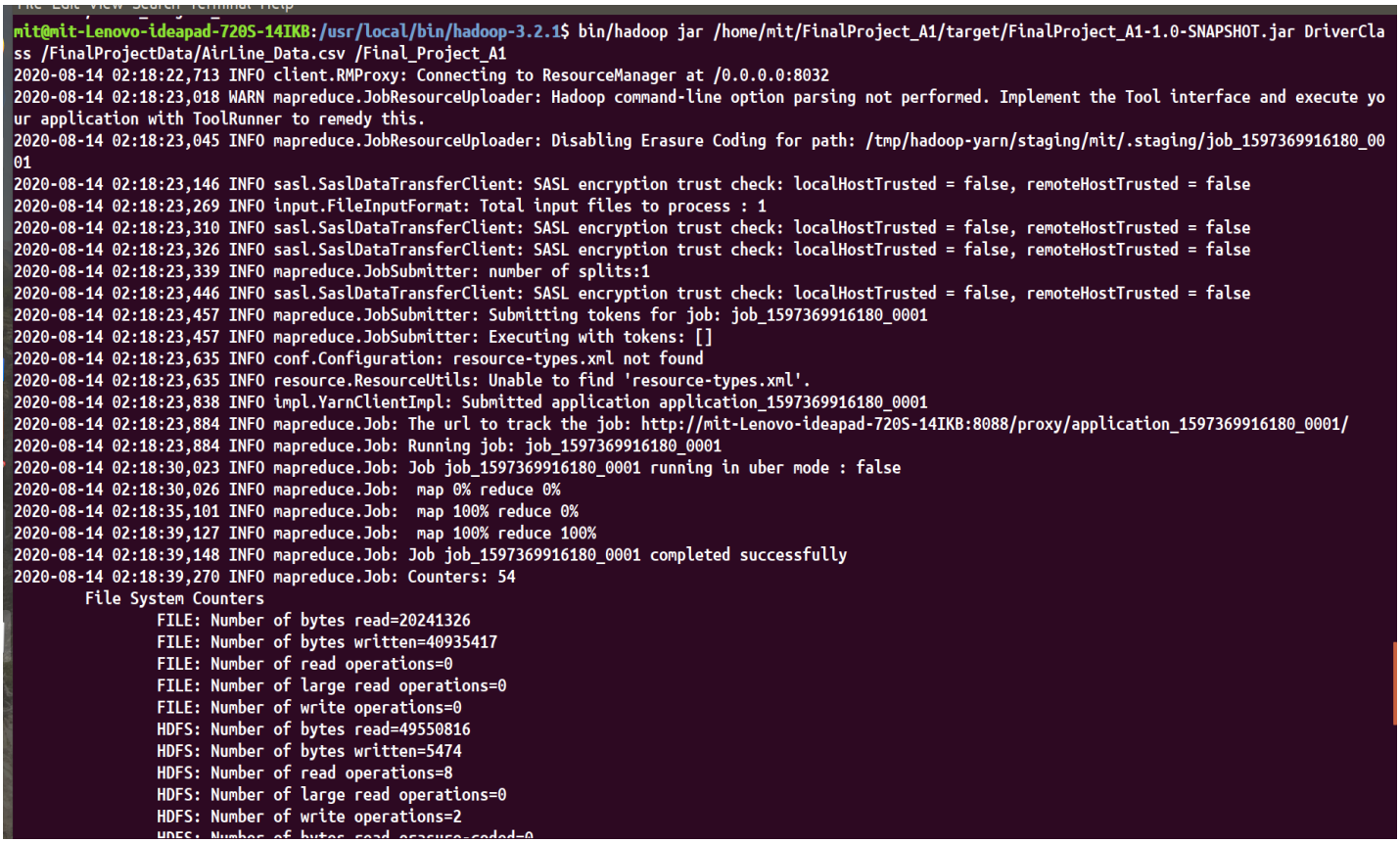
This analysis helps the airline know what are the causes which result in maximum delay and see how it can be minimized in the future

Possible delay causes are

- carrier delay
- weather delay
- nas delay
- security delay
- late_aircraft_delay

We also have total delay for every observation so we will be calculating the percentage for every delay cause per airline

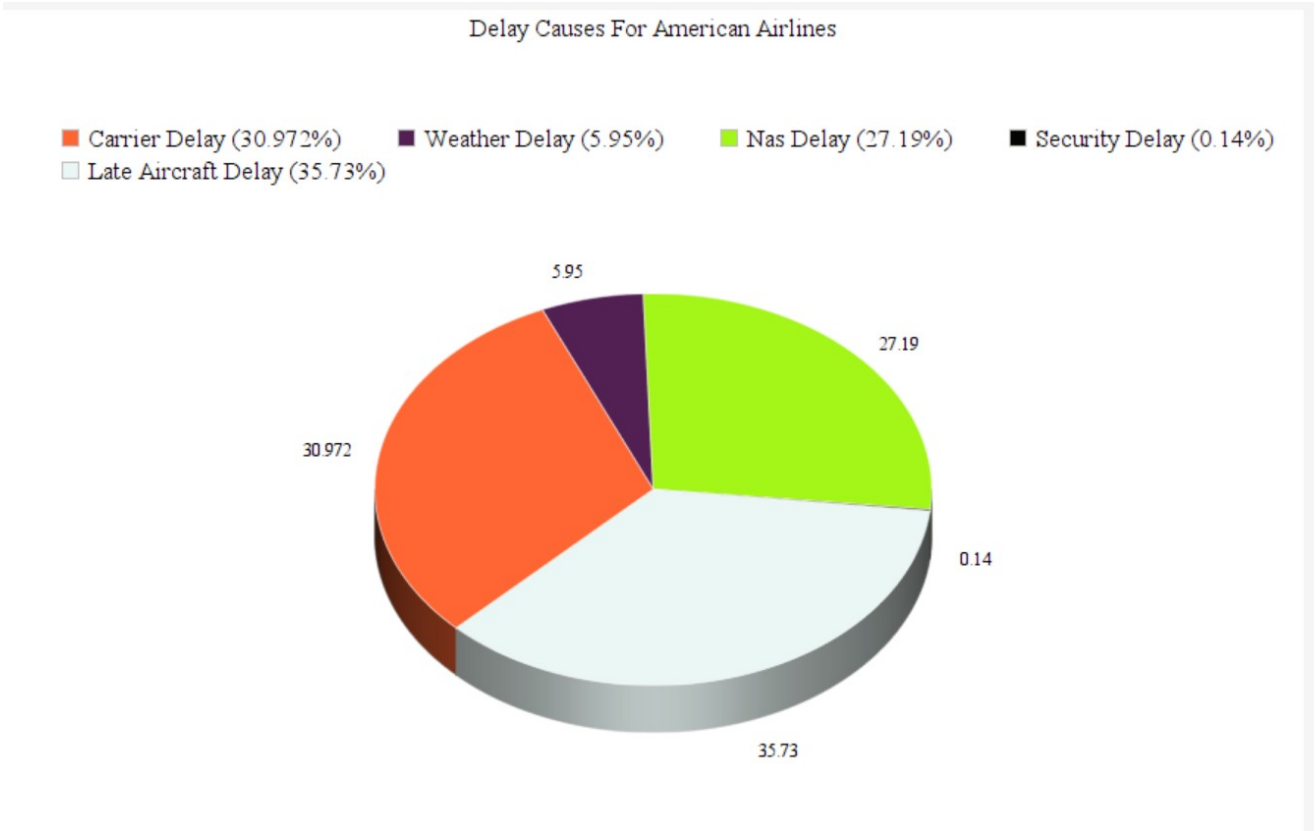
Output:



```
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$ bin/hadoop fs -cat /Final_Project_A1/part-r-00000
2020-08-14 02:14:10,152 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
"ATA Airlines d/b/a ATA"        -> Carrier: 19.84699124049079% Weather: 1.1323685816662596% Nas: 42.677609241803474% Security: 0.6318671502987298% Late A
ircraft: 35.71116378574075%
"AirTran Airways Corporation"   -> Carrier: 15.888351738935603% Weather: 1.2295797192396538% Nas: 32.27395755881494% Security: 0.0% Late Aircraft: 50.608
1109830098%
"Alaska Airlines Inc."  -> Carrier: 32.423597360082695% Weather: 2.7101080580034 29% Nas: 26.51918162118687% Security: 0.5112096948525172% Late Aircraft:
37.835903265874485%
"Allegiant Air" -> Carrier: 36.29782122968193% Weather: 6.783315373328914% Nas: 15.424353302133081% Security: 0.24410530621735055% Late Aircraft: 41.2504
04788638726%
"Aloha Airlines Inc."    -> Carrier: 55.22500104793443% Weather: 0.8753110459909839% Nas: 6.071541530593969 5% Security: 0.7129764767301398% Late Aircraft:
 37.11516989875048%
"America West Airlines Inc."    -> Carrier: 39.66832113953219% Weather: 1.6739166563962369% Nas: 30.906780911192527% Security: 0.5664608846014514% Late A
ircraft: 27.184520408277592%
"American Airlines Inc."        -> Carrier: 30.97274993389567% Weather: 5.954840146216426% Nas: 27.194107662048822% Security: 0.144025400820566 47% Late A
ircraft: 35.734276857018514%
"American Eagle Airlines Inc."  -> Carrier: 26.417465115863585% Weather: 6.999093413771422% Nas: 27.511344512530123% Security: 0.04611549446310614% Late
Aircraft: 39.025981463371764%
"Atlantic Coast Airlines"       -> Carrier: 21.700267378043183% Weather: 6.470854569594772% Nas: 30.954260203093103% Security: 0.06436652206633668% Late
Aircraft: 40.81025132720261%
"Atlantic Southeast Airlines"   -> Carrier: 38.03589851523679% Weather: 11.871024958329759% Nas: 25.422652750958836% Security: 0.10930097594908494% Late
Aircraft: 24.561122799525535%
"Comair Inc."   -> Carrier: 39.753005690925676% Weather: 22.911187650227856% Nas: 32.1179429902154% Security: 0.14336557925093613% Late Aircraft: 5.07449
8089380134%
"Continental Air Lines Inc."    -> Carrier: 21.602832346100197% Weather: 4.714825723701522% Nas: 46.82684730129721% Security: 0.4099299180140164% Late Ai
rcraft: 26.44556471088706%
"Delta Air Lines Inc."  -> Carrier: 34.0512831502242% Weather: 5.3229793664743 06% Nas: 30.8207261402701 92% Security: 0.08125143910591578% Late Aircraft:
29.72375990392539%
"Endeavor Air Inc."     -> Carrier: 27.591133288847952% Weather: 6.4487340750883% Nas: 26.420349727621613% Security: 0.03505580147564636% Late Aircraft: 3
9.504727107334 92%
"Envoy Air"     -> Carrier: 24.89119291538764% Weather: 8.735130677068549% Nas: 27.00683678597845% Security: 0.11473239699024101% Late Aircraft: 39.25210
722457512%
"ExpressJet Airlines Inc."      -> Carrier: 27.74888269777671% Weather: 3.51506423609551% Nas: 30.26817593624897% Security: 0.11729291288149336% Late Air
craft: 38.35058421699732%
"ExpressJet Airlines LLC"       -> Carrier: 30.685152006533905% Weather: 3.829712584626673% Nas: 37.052709594165826% Security: 0.0% Late Aircraft: 28.432
425814673596%
```

Example: American Airlines

(same can be visualized for other airlines)



b) Top 10 most punctual flights (MR chaining)?

This analysis will help the customer know which airlines are the best when it comes to punctuality.

We will do MR chaining to obtain the final result.

Map 1: Emit carrier name and number of flights delayed for every observation
Reduce 1: Calculate total number of delayed flights / carrier

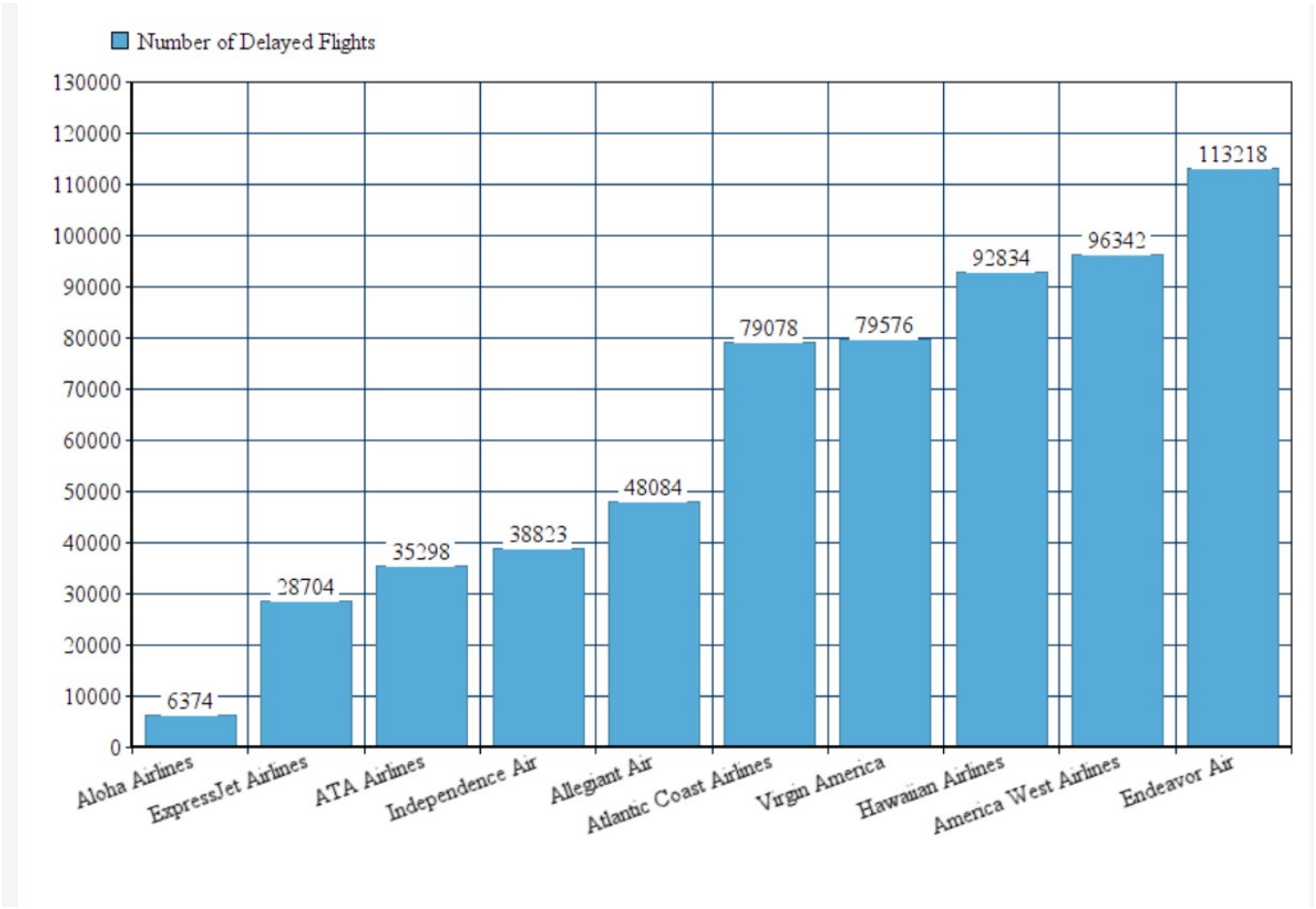
Map 2: Emit total delayed flights and carrier name. (key = total delayed flights)
Reduce 2: Set a counter and just write the first 10 entries. Sorting done by framework

Output:

```
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$ bin/hadoop jar /home/mit/FinalProject_A2/target/FinalProject_A2-1.0-SNAPSHOT.jar DriverCla
ss /FinalProjectData/AirLine_Data.csv /Final_Project_A2
2020-08-14 02:47:26,598 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-08-14 02:47:26,892 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute yo
ur application with ToolRunner to remedy this.
2020-08-14 02:47:26,915 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/mit/.staging/job_1597369916180_00
02
2020-08-14 02:47:26,999 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-08-14 02:47:27,110 INFO input.FileInputFormat: Total input files to process : 1
2020-08-14 02:47:27,148 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-08-14 02:47:27,169 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-08-14 02:47:27,176 INFO mapreduce.JobSubmitter: number of splits:1
2020-08-14 02:47:27,315 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-08-14 02:47:27,734 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1597369916180_0002
2020-08-14 02:47:27,734 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-08-14 02:47:27,914 INFO conf.Configuration: resource-types.xml not found
2020-08-14 02:47:27,914 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-08-14 02:47:27,970 INFO impl.YarnClientImpl: Submitted application application_1597369916180_0002
2020-08-14 02:47:28,023 INFO mapreduce.Job: The url to track the job: http://mit-Lenovo-ideapad-720S-14IKB:8088/proxy/application_1597369916180_0002/
2020-08-14 02:47:28,024 INFO mapreduce.Job: Running job: job_1597369916180_0002
2020-08-14 02:47:33,097 INFO mapreduce.Job: Job job_1597369916180_0002 running in uber mode : false
2020-08-14 02:47:33,099 INFO mapreduce.Job:  map 0% reduce 0%
2020-08-14 02:47:38,176 INFO mapreduce.Job:  map 100% reduce 0%
2020-08-14 02:47:43,217 INFO mapreduce.Job:  map 100% reduce 100%
2020-08-14 02:47:43,238 INFO mapreduce.Job: Job job_1597369916180_0002 completed successfully
2020-08-14 02:47:43,326 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=9288486
                FILE: Number of bytes written=19029777
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=49550816
                HDFS: Number of bytes written=1018
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
```

```
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$ bin/hadoop fs -ls /Final_Project_A2/out2/
Found 2 items
-rw-r--r--   1 mit supergroup          0 2020-08-10 01:01 /Final_Project_A2/out2/_SUCCESS
-rw-r--r--   1 mit supergroup        305 2020-08-10 01:01 /Final_Project_A2/out2/part-r-00000
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$ bin/hadoop fs -cat /Final_Project_A2/out2/part-r-00000
2020-08-14 02:45:34,281 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
6374.0  "Aloha Airlines Inc."
28704.0 "ExpressJet Airlines LLC"
35298.0 "ATA Airlines d/b/a ATA"
38823.0 "Independence Air"
48084.0 "Allegiant Air"
79078.0 "Atlantic Coast Airlines"
79576.0 "Virgin America"
92834.0 "Hawaiian Airlines Inc."
96342.0 "America West Airlines Inc."
113218.0        "Endeavor Air Inc."
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$
```

Result: Aloha Airline is the most punctual airlines (2003 - 2020)

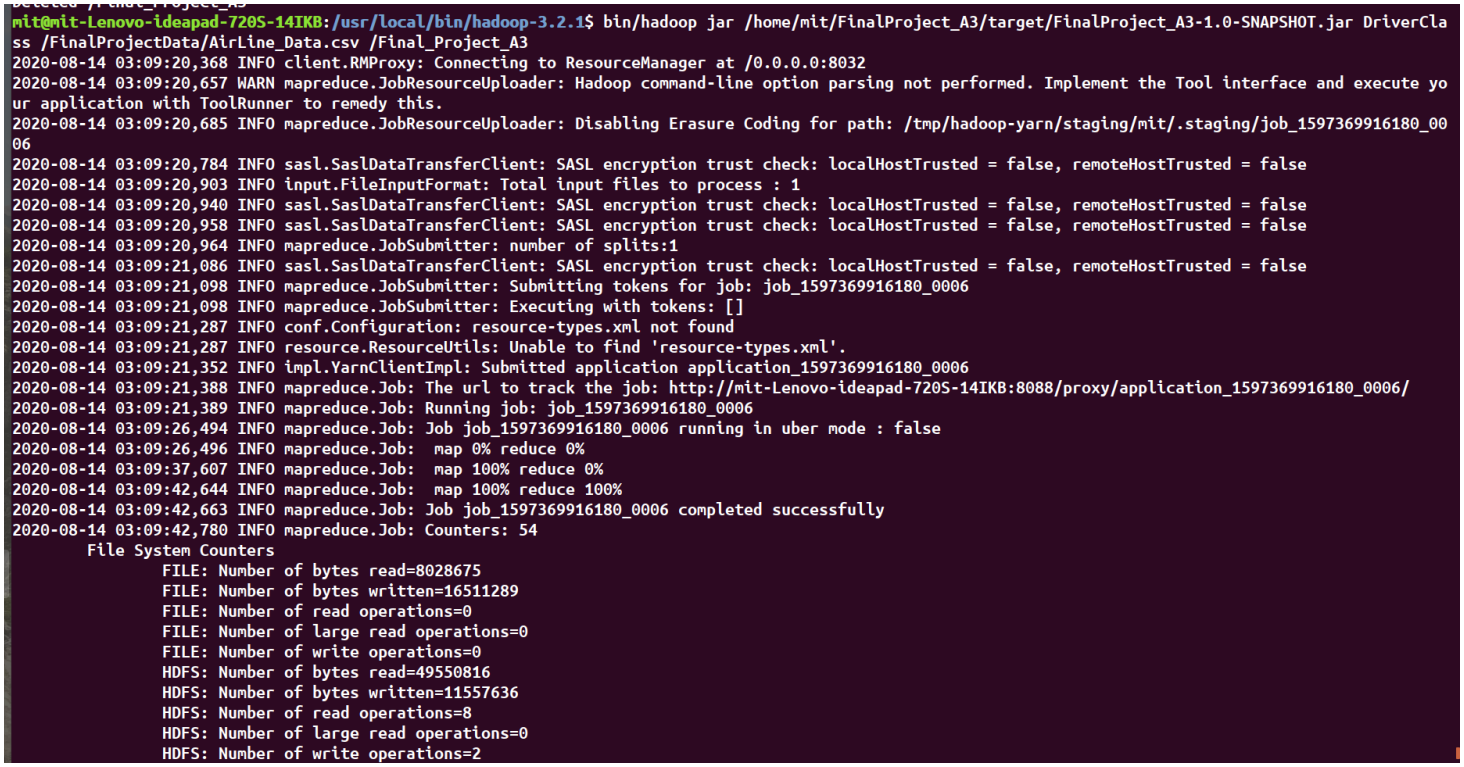


c) Busiest airport in USA? (Secondary sort by year)

This analysis will help us understand how many airlines does an airport have to deal with and how it has changed overtime.

We will use a composite key to sort the data by year in descending order

Output:



Head of output file

```
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$ bin/hadoop fs -head /Final_Project_A3/part-r-00000
2020-08-14 03:10:36,849 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Year: 2020 airportName: "Appleton       89.0
Year: 2020 airportName: "Austin 4.0
Year: 2020 airportName: "Asheville      75.0
Year: 2020 airportName: "Scranton/Wilkes-Barre  49.0
Year: 2020 airportName: "Kalamazoo      88.0
Year: 2020 airportName: "Birmingham     127.0
Year: 2020 airportName: "Billings       31.0
Year: 2020 airportName: "Bloomington/Normal     84.0
Year: 2020 airportName: "Nashville      147.0
Year: 2020 airportName: "Boise  62.0
Year: 2020 airportName: "Beaumont/Port Arthur   80.0
Year: 2020 airportName: "Brownsville    89.0
Year: 2020 airportName: "Baton Rouge    151.0
Year: 2020 airportName: "Buffalo        147.0
Year: 2020 airportName: "Baltimore      121.0
Year: 2020 airportName: "Bozeman        31.0
Year: 2020 airportName: "Columbia       133.0
Year: 2020 airportName: "Chattanooga    18.0
Year: 2020 airportName: "Charlottesville        146.0
Year: 2020 airportName: "Charleston     84.0
Year: 2020 airportName: "Cedar Rapids/Iowa City 128.0
Year: 2020 airportName: "Cleveland      293.0
Year: 2020 airportName: "College Station/Bryan  49.0
Year: 2020 airportName: "Charlotte      544.mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$
```

Tail of output file

```
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$ bin/hadoop fs -tail /Final_Project_A3/part-r-00000
2020-08-14 03:10:18,813 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
rlington        339.0
Year: 2003 airportName: "Boston 875.0
Year: 2003 airportName: "Nashville      235.0
Year: 2003 airportName: "Bloomington/Normal     60.0
Year: 2003 airportName: "Birmingham     93.0
Year: 2003 airportName: "Bangor 124.0
Year: 2003 airportName: "Binghamton     93.0
Year: 2003 airportName: "Hartford       124.0
Year: 2003 airportName: "Kalamazoo      98.0
Year: 2003 airportName: "Scranton/Wilkes-Barre  62.0
Year: 2003 airportName: "Asheville      31.0
Year: 2003 airportName: "Austin 30.0
Year: 2003 airportName: "Appleton       62.0
Year: 2003 airportName: "Albany 184.0
Year: 2003 airportName: "Atlantic City  13.0
Year: 2003 airportName: "Allentown/Bethlehem/Easton     184.0
Year: 2003 airportName: "Tucson 30.0
Year: 2003 airportName: "Tulsa  102.0
Year: 2003 airportName: "Tampa  368.0
Year: 2003 airportName: "Charlotte Amalie       4.0
Year: 2003 airportName: "St. Louis      1.0
Year: 2003 airportName: "Sarasota/Bradenton     31.0
Year: 2003 airportName: "Santa Ana      219.0
Year: 2003 airportName: "Sacramento     62.0
Year: 2003 airportName: "Salt Lake City 123.0
mit@mit-Lenovo-ideapad-720S-14IKB:/usr/local/bin/hadoop-3.2.1$
```

d) Number of canceled flights / carrier? (Hive)

This analysis will help up determine which airlines have the worst record in canceled flights

```
hive> select carrier_name ,count(arr_canceled) from airlineData GROUP BY carrier_name;
Query ID = mit_20200810223108_27b65c4b-cd13-4fdf-b429-b13cf4538c95
Total jobs = 1
Launching Job 1 out of 1
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_1597107532326_0002, Tracking URL = http://mit-Lenovo-ideapad-720S-14IKB:8088/proxy/application_1597107532326_0002/
Kill Command = /usr/local/bin/hadoop-3.2.1//bin/mapred job  -kill job_1597107532326_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-10 22:31:17,528 Stage-1 map = 0%,  reduce = 0%
2020-08-10 22:31:21,683 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.91 sec
2020-08-10 22:31:26,819 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.43 sec
MapReduce Total cumulative CPU time: 5 seconds 430 msec
Ended Job = job_1597107532326_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.43 sec   HDFS Read: 49568205 HDFS Write: 1373 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 430 msec
OK
```

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.43 sec   HDFS Read: 49568205 HDFS Write: 1373 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 430 msec
OK
"ATA Airlines d/b/a ATA"        918
"AirTran Airways Corporation"   6955
"Alaska Airlines Inc."  11128
"Allegiant Air" 3318
"Aloha Airlines Inc."   253
"America West Airlines Inc."    1579
"American Airlines Inc."        17298
"American Eagle Airlines Inc."  15404
"Atlantic Coast Airlines"       1253
"Atlantic Southeast Airlines"   12159
"Comair Inc."   7687
"Continental Air Lines Inc."    6909
"Delta Air Lines Inc."  24607
"Endeavor Air Inc."     3464
"Envoy Air"     6392
"ExpressJet Airlines Inc."      24632
"ExpressJet Airlines LLC"       1217
"Frontier Airlines Inc."        9903
"Hawaiian Airlines Inc."        3127
"Independence Air"      670
"JetBlue Airways"       10114
"Mesa Airlines Inc."    12386
"Northwest Airlines Inc."       8069
"PSA Airlines Inc."     2590
"Pinnacle Airlines Inc."        6437
"Republic Airline"      2501
"SkyWest Airlines Inc." 32873
"Southwest Airlines Co."        15098
"Spirit Air Lines"      2527
"US Airways Inc."       10693
"United Air Lines Inc." 16841
"Virgin America"        1426
Time taken: 19.527 seconds, Fetched: 32 row(s)
hive>
```

e) Best month for travel with least weather related delays? (Hive)

This analysis gives users an overview on which month in the year is the best to travel as less chances of weather delays

```
hive> select month ,cast(sum(weather_delay) as decimal(8,0)) from airlineData GROUP BY month order by sum(weather_delay);
Query ID = mit_20200810224826_d0afd930-ef1f-4936-962c-2460716c4c39
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_1597107532326_0005, Tracking URL = http://mit-Lenovo-ideapad-720S-14IKB:8088/proxy/application_1597107532326_0005/
Kill Command = /usr/local/bin/hadoop-3.2.1//bin/mapred job  -kill job_1597107532326_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-10 22:48:33,467 Stage-1 map = 0%,  reduce = 0%
2020-08-10 22:48:38,669 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.61 sec
2020-08-10 22:48:44,820 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.99 sec
MapReduce Total cumulative CPU time: 7 seconds 990 msec
Ended Job = job_1597107532326_0005
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_1597107532326_0006, Tracking URL = http://mit-Lenovo-ideapad-720S-14IKB:8088/proxy/application_1597107532326_0006/
Kill Command = /usr/local/bin/hadoop-3.2.1//bin/mapred job  -kill job_1597107532326_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-08-10 22:48:56,824 Stage-2 map = 0%,  reduce = 0%
2020-08-10 22:49:01,989 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.14 sec
2020-08-10 22:49:07,123 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.77 sec
MapReduce Total cumulative CPU time: 4 seconds 770 msec
Ended Job = job_1597107532326_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.99 sec   HDFS Read: 49567815 HDFS Write: 480 SUCCESS
```

{1: Jan, 2: Feb, 3: Mar, 4: Apr, 5: May, 6: June, 7: July, 8: Aug, 9: Sept, 10: Oct, 11: Nov, 12: Dec}

```
MapReduce Total cumulative CPU time: 7 seconds 990 msec
Ended Job = job_1597107532326_0005
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_1597107532326_0006, Tracking URL = http://mit-Lenovo-ideapad-720S-14IKB:8088/proxy/application_1597107532326_0006/
Kill Command = /usr/local/bin/hadoop-3.2.1//bin/mapred job  -kill job_1597107532326_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-08-10 22:48:56,824 Stage-2 map = 0%,  reduce = 0%
2020-08-10 22:49:01,989 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.14 sec
2020-08-10 22:49:07,123 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.77 sec
MapReduce Total cumulative CPU time: 4 seconds 770 msec
Ended Job = job_1597107532326_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.99 sec   HDFS Read: 49567815 HDFS Write: 480 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.77 sec   HDFS Read: 8479 HDFS Write: 366 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 760 msec
OK
9       22785121
11      23407582
10      24811244
4       26653286
5       27476861
2       28333013
3       30546262
1       31296797
8       35247213
12      36549342
6       37942083
7       39356828
Time taken: 42.817 seconds, Fetched: 12 row(s)
hive>
```

f) Which airport had the most NAS delays? (Pig)

NAS: National Aviation System

Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control

These are the delays which could have been avoided if corrective action were taken by the airport or federal aviation administration.

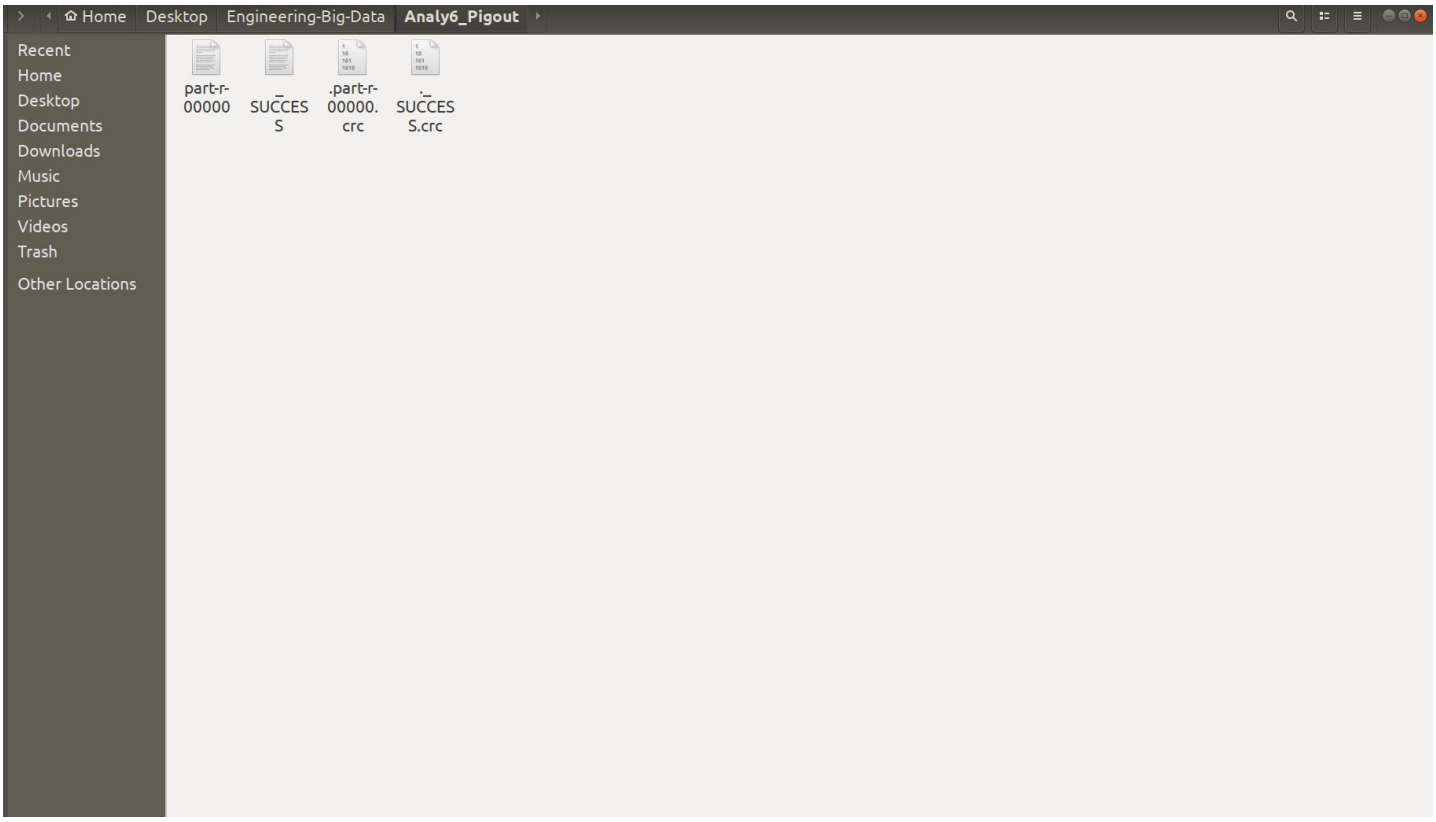
```
mit@mit-Lenovo-ideapad-720S-14IKB:~$ pig -x local
2020-08-14 01:45:11,722 INFO  [main] pig.ExecTypeProvider (ExecTypeProvider.java:selectExecType(41)) - Trying ExecType : LOCAL
2020-08-14 01:45:11,723 INFO  [main] pig.ExecTypeProvider (ExecTypeProvider.java:selectExecType(43)) - Picked LOCAL as the ExecType
2020-08-14 01:45:11,749 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2020-08-14 01:45:11,750 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/mit/pig_1597383911745.log
2020-08-14 01:45:11,762 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/mit/.pigbootup not found
2020-08-14 01:45:11,898 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrack
er.address
2020-08-14 01:45:11,899 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2020-08-14 01:45:12,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-c
hecksum
2020-08-14 01:45:12,057 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-67a1a662-3a49-46f0-b440-6fb8ac88fbe1
2020-08-14 01:45:12,057 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> data = LOAD '/home/mit/Desktop/Engineering-Big-Data/datasets-projects/AirlineData_Full/AirLine_Data_NoHeader.csv' using PigStorage(',') AS (year,m
onth,carrier,carrier_name,airport,airport_name,arr_flights,arr_del15,carrier_ct,weather_ct,nas_ct,security_ct,late_aircraft_ct,arr_canceled,arr_diverted,
arr_delay,carrier_delay,weather_delay,nas_delay,security_delay,late_aircraft_delay);
2020-08-14 01:45:20,119 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-c
hecksum
grunt> airportgroup = GROUP data by airport;
grunt> describe airportgroup;
airportgroup: {group: bytearray,data: {(year: bytearray,month: bytearray,carrier: bytearray,carrier_name: bytearray,airport: bytearray,airport_name: byte
array,arr_flights: bytearray,arr_del15: bytearray,carrier_ct: bytearray,weather_ct: bytearray,nas_ct: bytearray,security_ct: bytearray,late_aircraft_ct:
bytearray,arr_canceled: bytearray,arr_diverted: bytearray,arr_delay: bytearray,carrier_delay: bytearray,weather_delay: bytearray,nas_delay: bytearray,sec
urity_delay: bytearray,late_aircraft_delay: bytearray)}}
grunt> nasDelayCount = FOREACH airportgroup GENERATE group,SUM(data.nas_delay) as totalNasDelay;
grunt> describe nasDelayCount;
nasDelayCount: {group: bytearray,totalNasDelay: double}
grunt> sortedNasDelayCount = ORDER nasDelayCount BY totalNasDelay DESC;
grunt> STORE sortedNasDelayCount into '/home/mit/Desktop/Engineering-Big-Data/Analy6_Pigout';
2020-08-14 01:49:35,500 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-c
hecksum
2020-08-14 01:49:35,519 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use map
reduce.output.textoutputformat.separator
2020-08-14 01:49:35,539 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY
2020-08-14 01:49:35,579 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-c
hecksum
2020-08-14 01:49:35,606 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, Const
antCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimize
```

```
Output(s):
Successfully stored 409 records in: "/home/mit/Desktop/Engineering-Big-Data/Analy6_Pigout"

Counters:
Total records written : 409
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local751352462_0001 ->      job_local1141842868_0002,
job_local1141842868_0002        ->      job_local70254493_0003,
job_local70254493_0003


2020-08-14 01:49:40,492 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,493 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,493 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,498 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,498 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,498 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,502 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,504 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,505 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - a
lready initialized
2020-08-14 01:49:40,510 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```



Open ▾ part-r-00000 ~/Desktop/Engineering-Big-Data/Analy6_Pigout Save

```
"ATL"   4076452.0
"DFW"   3373023.0
"ORD"   3193117.0
"DEN"   1740563.0
"LGA"   1707207.0
"IAH"   1684888.0
"CVG"   1586656.0
"EWR"   1536671.0
"DTW"   1510877.0
"MSP"   1498889.0
"LAX"   1451107.0
"SFO"   1300852.0
"BOS"   1269060.0
"PHL"   1151910.0
"JFK"   1148180.0
"LAS"   984159.0
"BWI"   944493.0
"CLT"   943573.0
"SLC"   930864.0
"MDW"   913036.0
"MCO"   888559.0
"DCA"   883222.0
"PHX"   882094.0
"MIA"   712470.0
"SEA"   672810.0
"IAD"   631180.0
"RDU"   587283.0
"TPA"   560339.0
"FLL"   558167.0
"BNA"   524345.0
"CLE"   511373.0
"SAN"   500489.0
"STL"   487491.0
"MCI"   432493.0
"HOU"   427806.0
"AUS"   424521.0
"DAL"   405462.0
```

Plain Text ▾ Tab Width: 8 ▾ Ln 1, Col 1 ▾ INS

- **Top 5 (most NAS Delay)**
ATL: Atlanta Airport
DFW: Dallas
ORD: Chicago
DEN: Denver
LGA: Queens, NY

```
2020-08-14 01:58:43,625 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
("ATL",4076452.0)
("DFW",3373023.0)
("ORD",3193117.0)
("DEN",1740563.0)
("LGA",1707207.0)
("IAH",1684888.0)
("CVG",1586656.0)
("EWR",1536671.0)
("DTW",1510877.0)
("MSP",1498889.0)
grunt>
```

```
grunt> top10 = LIMIT sortedNasDelayCount 10;
grunt> describe top10;
top10: {group: bytearray,totalNasDelay: double}
grunt> dump top10;
```

g) Which carrier had the least number of diverted flights? (mongoDb)

This analysis will help airline companies analyze how many of its flights were diverted over the years

```
mit@mit-Lenovo-ideapad-720S-14IKB:~$ mongoimport --db finalProject --collection airlineData --type csv --headerline --file /home/mit/Desktop/Engineering-Big-Data/datasets-projects/AirlineData_Full/AirLine_Data.csv
2020-08-14T21:52:44.147-0400    connected to: localhost
2020-08-14T21:52:47.142-0400    [#############.........] finalProject.airlineData     26.6MB/47.3MB (56.2%)
2020-08-14T21:52:49.434-0400    [######################] finalProject.airlineData     47.3MB/47.3MB (100.0%)
2020-08-14T21:52:49.434-0400    imported 280833 documents
mit@mit-Lenovo-ideapad-720S-14IKB:~$
```

```
> show dbs;
ContactManagementSystem  0.000GB
accesslog_Assignment3    0.002GB
admin                    0.000GB
config                   0.000GB
finalProject             0.042GB
gamesDb                  0.000GB
local                    0.000GB
mitdb                    0.000GB
movielens                0.031GB
movielens_Assign3        0.380GB
nyse_indexing_2          0.743GB
nyse_indexing_3          0.742GB
nysedb                   0.694GB
test                     0.000GB
userdb                   0.000GB
> use finalProject;
switched to db finalProject
> show collections;
airlineData
```

```
> db.airlineData.find().limit(1).pretty();
{
        "_id" : ObjectId("5f373fec5e048df510d2ec43"),
        "year" : 2004,
        " month" : 1,
        "carrier" : "DL",
        "carrier_name" : "Delta Air Lines Inc.",
        "airport" : "PBI",
        "airport_name" : "West Palm Beach/Palm Beach, FL: Palm Beach International",
        "arr_flights" : 650,
        "arr_del15" : 126,
        "carrier_ct" : 21.06,
        " weather_ct" : 6.44,
        "nas_ct" : 51.58,
        "security_ct" : 1,
        "late_aircraft_ct" : 45.92,
        "arr_cancelled" : 4,
        "arr_diverted" : 0,
        " arr_delay" : 5425,
        " carrier_delay" : 881,
        "weather_delay" : 397,
        "nas_delay" : 2016,
        "security_delay" : 15,
        "late_aircraft_delay" : 2116,
        "" : ""
}
>
```

Output:

Aloha airlines had least number of diverted flights while Southwest airlines had maximum diverted flights from 2003 to 2020

```
> db.airlineData.aggregate([{"$group": {_id:"$carrier_name", numOfDivertedFlights:{$sum:"$arr_diverted"}}} , {$sort: {numOfDivertedFlights:1}}])
{ "_id" : "Aloha Airlines Inc.", "numOfDivertedFlights" : 36 }
{ "_id" : "ATA Airlines d/b/a ATA", "numOfDivertedFlights" : 80 }
{ "_id" : "Independence Air", "numOfDivertedFlights" : 238 }
{ "_id" : "ExpressJet Airlines LLC", "numOfDivertedFlights" : 390 }
{ "_id" : "Allegiant Air", "numOfDivertedFlights" : 690 }
{ "_id" : "America West Airlines Inc.", "numOfDivertedFlights" : 691 }
{ "_id" : "Atlantic Coast Airlines", "numOfDivertedFlights" : 729 }
{ "_id" : "Hawaiian Airlines Inc.", "numOfDivertedFlights" : 825 }
{ "_id" : "Virgin America", "numOfDivertedFlights" : 952 }
{ "_id" : "Spirit Air Lines", "numOfDivertedFlights" : 1341 }
{ "_id" : "Endeavor Air Inc.", "numOfDivertedFlights" : 1378 }
{ "_id" : "Republic Airline", "numOfDivertedFlights" : 1732 }
{ "_id" : "PSA Airlines Inc.", "numOfDivertedFlights" : 1822 }
{ "_id" : "Frontier Airlines Inc.", "numOfDivertedFlights" : 1990 }
{ "_id" : "Pinnacle Airlines Inc.", "numOfDivertedFlights" : 3461 }
{ "_id" : "Envoy Air", "numOfDivertedFlights" : 3506 }
{ "_id" : "Comair Inc.", "numOfDivertedFlights" : 3584 }
{ "_id" : "Mesa Airlines Inc.", "numOfDivertedFlights" : 4517 }
{ "_id" : "Atlantic Southeast Airlines", "numOfDivertedFlights" : 5346 }
{ "_id" : "Northwest Airlines Inc.", "numOfDivertedFlights" : 5558 }
Type "it" for more
> it
{ "_id" : "AirTran Airways Corporation", "numOfDivertedFlights" : 5577 }
{ "_id" : "Continental Air Lines Inc.", "numOfDivertedFlights" : 6949 }
{ "_id" : "Alaska Airlines Inc.", "numOfDivertedFlights" : 7910 }
{ "_id" : "US Airways Inc.", "numOfDivertedFlights" : 9194 }
{ "_id" : "JetBlue Airways", "numOfDivertedFlights" : 10352 }
{ "_id" : "American Eagle Airlines Inc.", "numOfDivertedFlights" : 11901 }
{ "_id" : "United Air Lines Inc.", "numOfDivertedFlights" : 19009 }
{ "_id" : "ExpressJet Airlines Inc.", "numOfDivertedFlights" : 22404 }
{ "_id" : "Delta Air Lines Inc.", "numOfDivertedFlights" : 24163 }
{ "_id" : "SkyWest Airlines Inc.", "numOfDivertedFlights" : 25722 }
{ "_id" : "American Airlines Inc.", "numOfDivertedFlights" : 32939 }
{ "_id" : "Southwest Airlines Co.", "numOfDivertedFlights" : 38381 }
>
```