Domain: Transportation
Dataset: Airlines – Airport, Airlines and Route data (http://openflights.org/data.html)
Airlines OD data (http://stat-computing.org/dataexpo/2009/the-data.html)
http://stat-computing.org/dataexpo/2009/
Customer Reviews: https://github.com/quankiquanki/skytrax-reviews-dataset
Airline Delays: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

About the Dataset and Data: (2003 – 2017)
The airline dataset consists of all details about the airlines, airports and the routes data from a source airport to the destination along with the airline taken on the route.

The data also consists of airline delays and takes into consideration different factors of flight delays such as Arrival Delay, Carrier Delay, Weather Delay, NAS Delay, Security Delay, and Late Aircraft Delay.

Also, customer reviews have also been taken into consideration for different carriers, so as to know the carrier/airline review.

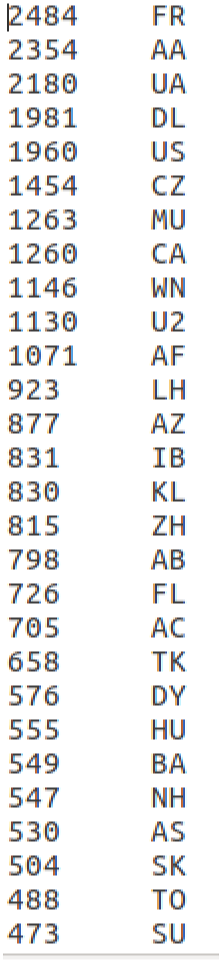
Analysis Performed:

1) **Most Preferred Airline:**

   Operations Performed: **Chaining and Secondary Sorting** using Comparator Class.

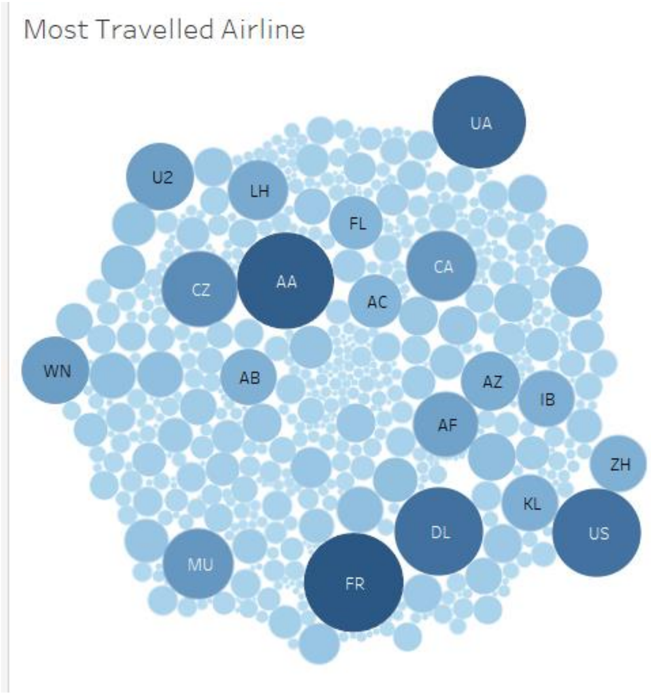
   Analysis: The routes database has details about routes taken from a source airport to destination airport with several airlines. This analysis is done by performing chaining operations, by taking the output from the first job (getting the airline and its count) and giving it to the second job where sorting is performed in descending order using a comparator class.

   Result/Output:



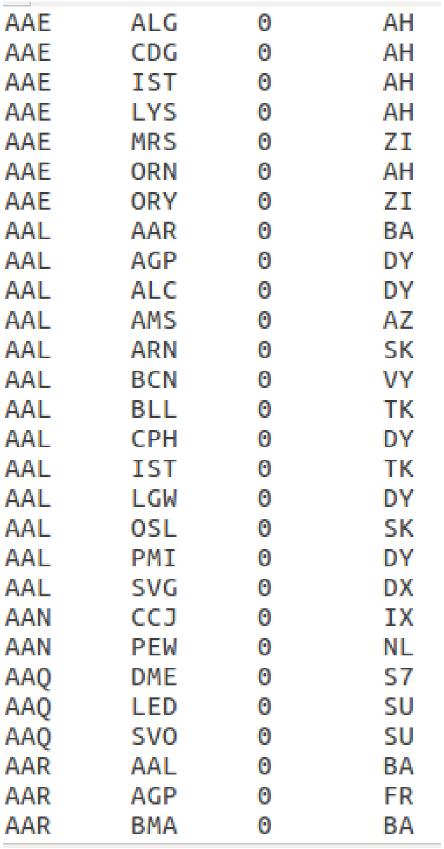
```
2484    FR
2354    AA
2180    UA
1981    DL
1960    US
1454    CZ
1263    MU
1260    CA
1146    WN
1130    U2
1071    AF
923     LH
877     AZ
831     IB
830     KL
815     ZH
798     AB
726     FL
705     AC
658     TK
576     DY
555     HU
549     BA
547     NH
530     AS
504     SK
488     TO
473     SU
```

   Conclusion: It can be concluded that FR, AA, UA, DL are among the most travelled/preferred airlines.



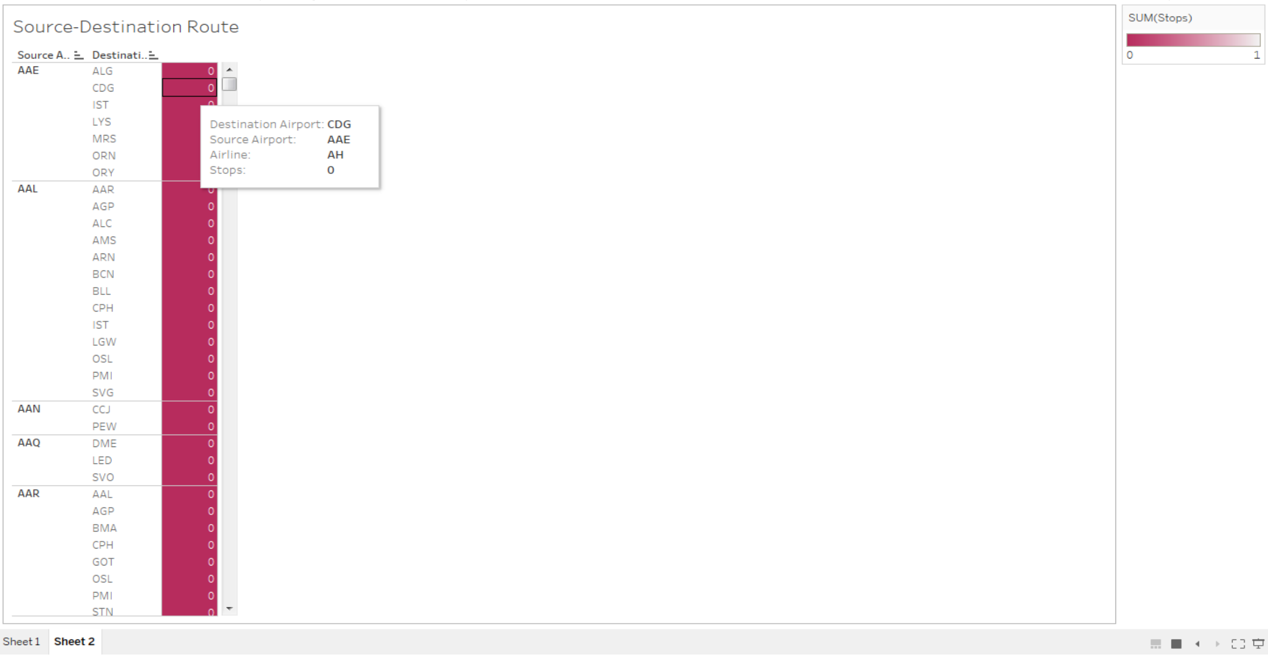
2) **Source – Destination Analysis:**
Use of **Custom Writable** Objects.
The routes dataset allows to perform analysis to find out the different routes that can be taken from a particular airport (Source Airport). Also, information such as the airline taken and the number of stops from travelling Source -> Destination has also been displayed.
Two custom writable classes have been created, one to store and sort the ke y (source and destination) and the other (value) to store and display the airline and number of stops details.

Result/Output:



```
AAE     ALG     0       AH
AAE     CDG     0       AH
AAE     IST     0       AH
AAE     LYS     0       AH
AAE     MRS     0       ZI
AAE     ORN     0       AH
AAE     ORY     0       ZI
AAL     AAR     0       BA
AAL     AGP     0       DY
AAL     ALC     0       DY
AAL     AMS     0       AZ
AAL     ARN     0       SK
AAL     BCN     0       VY
AAL     BLL     0       TK
AAL     CPH     0       DY
AAL     IST     0       TK
AAL     LGW     0       DY
AAL     OSL     0       SK
AAL     PMI     0       DY
AAL     SVG     0       DX
AAN     CCJ     0       IX
AAN     PEW     0       NL
AAQ     DME     0       S7
AAQ     LED     0       SU
AAQ     SVO     0       SU
AAR     AAL     0       BA
AAR     AGP     0       FR
AAR     BMA     0       BA
```

Conclusion: This analysis gives a clear picture about the different destinations from a source airport.



Source-Destination Route

Source A..	Destinati..	
AAE	ALG	0
	CDG	0
	IST	0
	LYS	
	MRS	
	ORN	
	ORY	
AAL	AAR	0
	AGP	0
	ALC	0
	AMS	0
	ARN	0
	BCN	0
	BLL	0
	CPH	0
	IST	0
	LGW	0
	OSL	0
	PMI	0
	SVG	0
AAN	CCJ	0
	PEW	0
AAQ	DME	0
	LED	0
	SVO	0
AAR	AAL	0
	AGP	0
	BMA	0
	CPH	0
	GOT	0
	OSL	0
	PMI	0
	STN	0

Destination Airport: CDG
Source Airport: AAE
Airline: AH
Stops: 0

3) Airline Delay Reason: (Summarization Pattern)

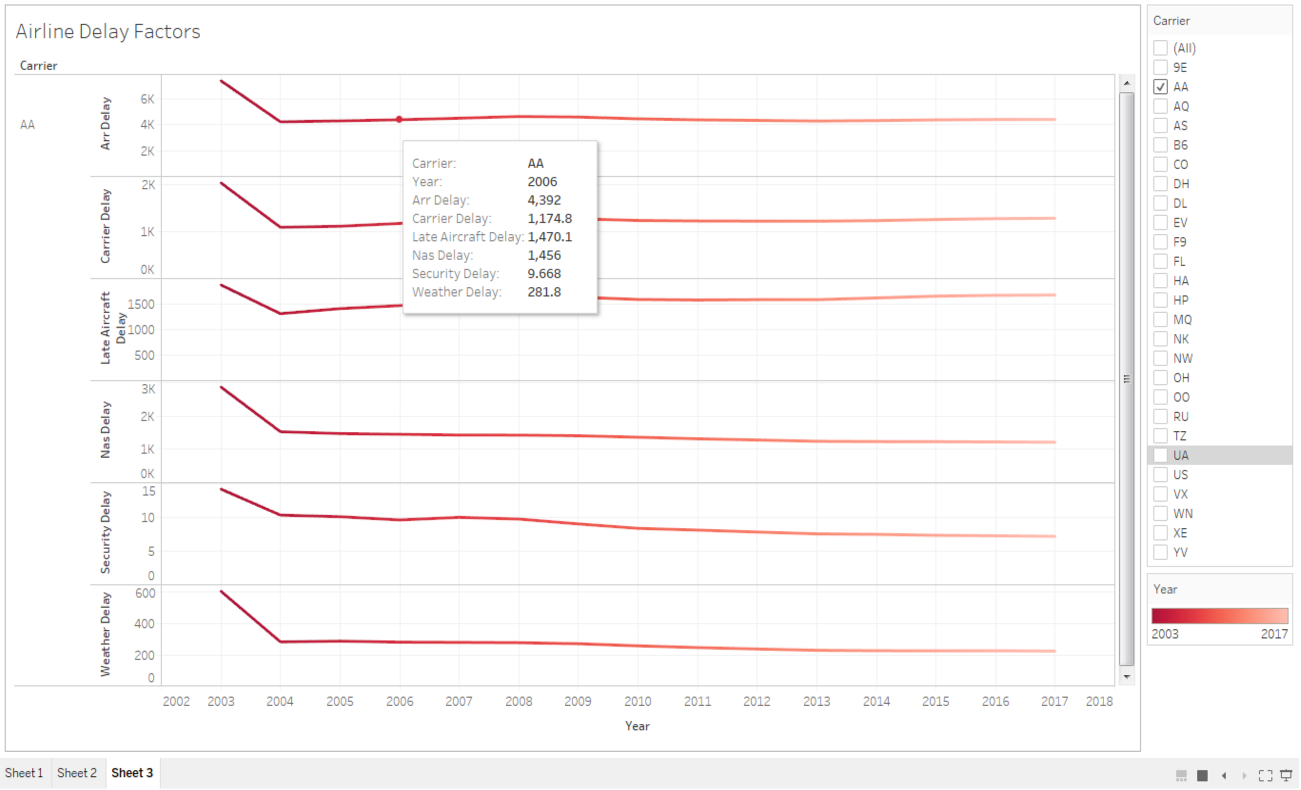
Use of **Numerical Summarization** Pattern.

The airline delays dataset is used to analyse the major factors on flight delays. The delays can be Arrival Delay, Carrier Delay, Weather Delay, NAS Delay, Security Delay, and Late Aircraft Delay. This analysis is performed by taking the average of these delays for each carrier/airline. For making use of combiner optimizations, this pattern is done by calculating the average by dividing the running sum to the running count.

Result/Output:

```
2003    AA    7410.879     2040.8988    602.70483    2877.3481    14.268657    1875.6583
2003    AS    5746.5005    1655.7291    410.3532     2032.7118    17.176598    1630.5298
2003    B6    5192.2925    1505.0669    364.1467     1823.3398    18.698236    1481.0409
2003    CO    4485.209     1163.3135    290.86307    1799.2452    14.472307    1217.3147
2003    DH    4149.6235    1121.0243    308.08777    1629.5819    11.576508    1079.3533
2003    DL    4145.1978    1118.6294    263.03992    1700.0676    10.000339    1053.4606
2003    EV    3804.7678    1065.2067    264.95526    1518.0345    8.813392     947.75793
2003    FL    3803.6538    1030.0128    252.78297    1511.5199    8.209612     1001.1286
2003    HA    3779.1516    1026.334     250.77028    1499.4553    8.178246     994.4138
2003    HP    3652.9727    1025.3403    234.05067    1431.4419    8.761739     953.378
2003    MQ    3666.2192    1018.4687    226.52007    1395.3483    7.963268     1017.9201
2003    NW    3543.2698    1027.3937    228.1533     1342.7717    7.6623693    937.29144
2003    OO    3352.4614    1003.4031    234.58763    1241.2139    8.912444     864.3502
2003    RU    3279.0283    944.02716    226.58308    1251.6135    8.402171     848.40735
2003    TZ    3250.3105    931.0641     221.66527    1242.358     8.724194     846.5027
2003    UA    3417.1594    946.7296     220.23909    1324.3438    8.502144     917.34576
2003    US    3502.495     947.74054    219.05392    1356.6512    8.080297     970.97266
2003    WN    3715.4033    978.2293     226.9179     1354.384     9.149366     1146.7233
2004    AA    4232.6343    1092.6854    283.40222    1533.3772    10.388815    1312.7793
2004    AS    4180.479     1099.2058    271.42288    1477.4537    11.245343    1321.1547
2004    B6    4126.918     1082.6498    265.92865    1457.8413    11.942873    1308.5609
2004    CO    4076.0322    1046.2085    258.3729     1491.943     12.231789    1267.2794
```

Conclusion: This analysis gives a clear detail about the flight delays over the years and can be seen that over the years the delays have lessened.



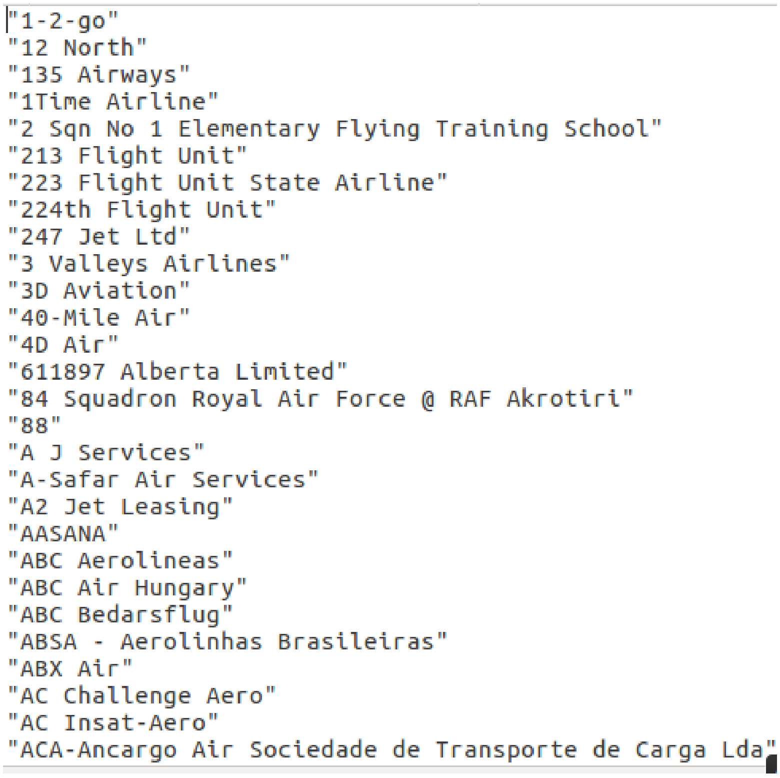
Carrier: AA
Year: 2006
Arr Delay: 4,392
Carrier Delay: 1,174.8
Late Aircraft Delay: 1,470.1
Nas Delay: 1,456
Security Delay: 9.668
Weather Delay: 281.8

4) **Distinct Airline: (Filtering Pattern)**

Use of **Distinct Filtering Pattern**.

This analysis gives us the distinct/unique airline names, so that we can know the different carriers/airlinesfor which we have the details of. Also, with having such large dataset, this pattern helps to get set of unique airlines and to understand better about the data that we are dealing with.

Result/Output:



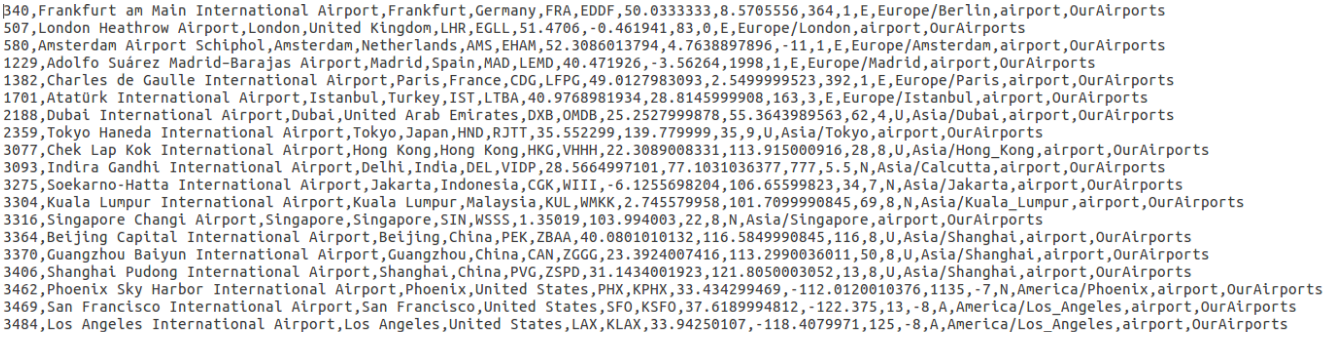
```
"1-2-go"
"12 North"
"135 Airways"
"1Time Airline"
"2 Sqn No 1 Elementary Flying Training School"
"213 Flight Unit"
"223 Flight Unit State Airline"
"224th Flight Unit"
"247 Jet Ltd"
"3 Valleys Airlines"
"3D Aviation"
"40-Mile Air"
"4D Air"
"611897 Alberta Limited"
"84 Squadron Royal Air Force @ RAF Akrotiri"
"88"
"A J Services"
"A-Safar Air Services"
"A2 Jet Leasing"
"AASANA"
"ABC Aerolineas"
"ABC Air Hungary"
"ABC Bedarsflug"
"ABSA - Aerolinhas Brasileiras"
"ABX Air"
"AC Challenge Aero"
"AC Insat-Aero"
"ACA-Ancargo Air Sociedade de Transporte de Carga Lda"
```

5) Top 30 Busiest Airport Details: (Filtering Pattern)

Use of **Bloom Filter** and **Distributed Cache.**

Bloom Filter helps to filter out the data by comparing to a set of hot values. In this analysis we make use of a bloom filter which consists of IATA code for 30 top most/busiest airports. The input data is compared to this bloom filter and only those records are emitted which are present in the set of hot values (bloom filter). The bloom filter is loaded into distributed cache and the mapper loads this files from distributed cache and compares the input file.

Result/Output:



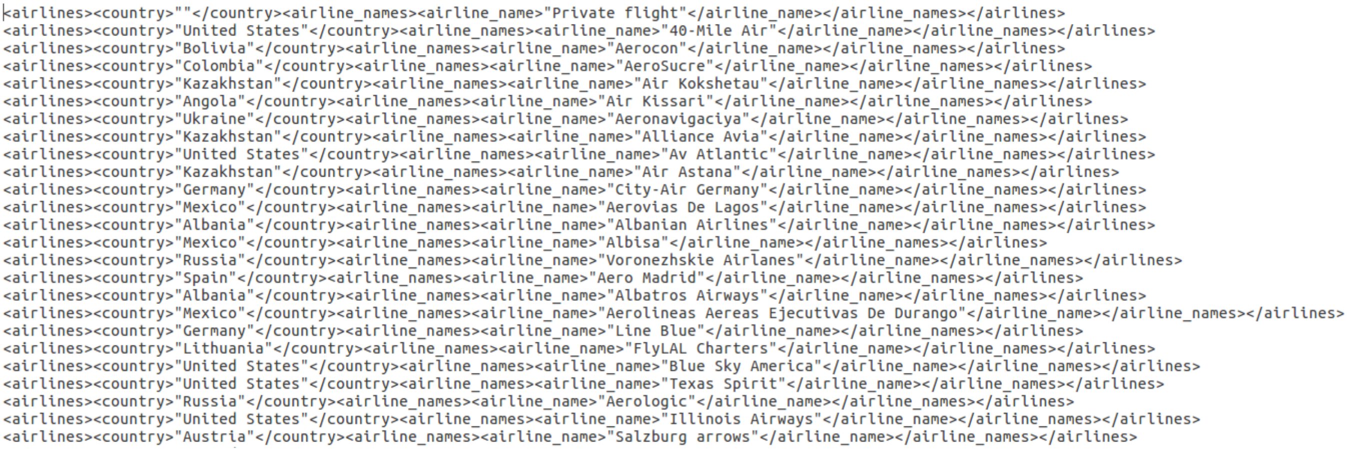
```
340,Frankfurt am Main International Airport,Frankfurt,Germany,FRA,EDDF,50.0333333,8.5705556,364,1,E,Europe/Berlin,airport,OurAirports
507,London Heathrow Airport,London,United Kingdom,LHR,EGLL,51.4706,-0.461941,83,0,E,Europe/London,airport,OurAirports
580,Amsterdam Airport Schiphol,Amsterdam,Netherlands,AMS,EHAM,52.3086013794,4.7638897896,-11,1,E,Europe/Amsterdam,airport,OurAirports
1229,Adolfo Suárez Madrid-Barajas Airport,Madrid,Spain,MAD,LEMD,40.471926,-3.56264,1998,1,E,Europe/Madrid,airport,OurAirports
1382,Charles de Gaulle International Airport,Paris,France,CDG,LFPG,49.0127983093,2.5499999523,392,1,E,Europe/Paris,airport,OurAirports
1701,Atatürk International Airport,Istanbul,Turkey,IST,LTBA,40.9768981934,28.8145999908,163,3,E,Europe/Istanbul,airport,OurAirports
2188,Dubai International Airport,Dubai,United Arab Emirates,DXB,OMDB,25.2527999878,55.3643989563,62,4,U,Asia/Dubai,airport,OurAirports
2359,Tokyo Haneda International Airport,Tokyo,Japan,HND,RJTT,35.552299,139.779999,35,9,U,Asia/Tokyo,airport,OurAirports
3077,Chek Lap Kok International Airport,Hong Kong,Hong Kong,HKG,VHHH,22.3089008331,113.915000916,28,8,U,Asia/Hong_Kong,airport,OurAirports
3093,Indira Gandhi International Airport,Delhi,India,DEL,VIDP,28.5664997101,77.1031036377,777,5.5,N,Asia/Calcutta,airport,OurAirports
3275,Soekarno-Hatta International Airport,Jakarta,Indonesia,CGK,WIII,-6.1255698204,106.65599823,34,7,N,Asia/Jakarta,airport,OurAirports
3304,Kuala Lumpur International Airport,Kuala Lumpur,Malaysia,KUL,WMKK,2.745579958,101.7099990845,69,8,N,Asia/Kuala_Lumpur,airport,OurAirports
3316,Singapore Changi Airport,Singapore,Singapore,SIN,WSSS,1.35019,103.994003,22,8,N,Asia/Singapore,airport,OurAirports
3364,Beijing Capital International Airport,Beijing,China,PEK,ZBAA,40.0801010132,116.5849990845,116,8,U,Asia/Shanghai,airport,OurAirports
3370,Guangzhou Baiyun International Airport,Guangzhou,China,CAN,ZGGG,23.3924007416,113.2990036011,50,8,U,Asia/Shanghai,airport,OurAirports
3406,Shanghai Pudong International Airport,Shanghai,China,PVG,ZSPD,31.1434001923,121.8050003052,13,8,U,Asia/Shanghai,airport,OurAirports
3462,Phoenix Sky Harbor International Airport,Phoenix,United States,PHX,KPHX,33.434299469,-112.0120010376,1135,-7,N,America/Phoenix,airport,OurAirports
3469,San Francisco International Airport,San Francisco,United States,SFO,KSFO,37.6189994812,-122.375,13,-8,A,America/Los_Angeles,airport,OurAirports
3484,Los Angeles International Airport,Los Angeles,United States,LAX,KLAX,33.94250107,-118.4079971,125,-8,A,America/Los_Angeles,airport,OurAirports
```

6) Country – Airline Organization: (Organization Pattern)

Use of **Structured to Hierarchical Pattern**.

This organization pattern allows to have a hierarchical representation (XML format) of airlines by their country. Here the country is the parent element and airline name is the child node.

Result/Output:



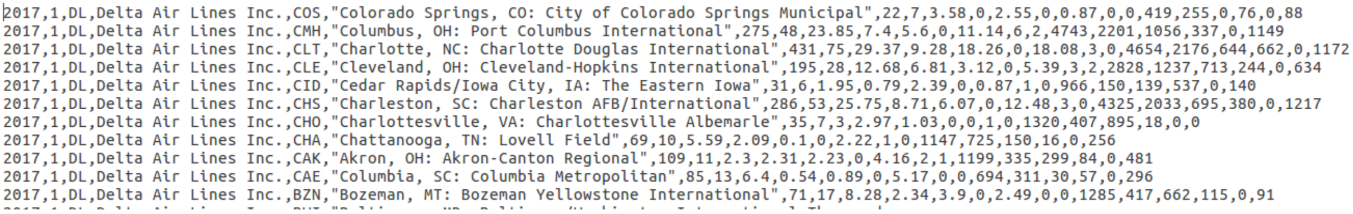
```
<airlines><country>""</country><airline_names><airline_name>"Private flight"</airline_name></airline_names></airlines>
<airlines><country>"United States"</country><airline_names><airline_name>"40-Mile Air"</airline_name></airline_names></airlines>
<airlines><country>"Bolivia"</country><airline_names><airline_name>"Aerocon"</airline_name></airline_names></airlines>
<airlines><country>"Colombia"</country><airline_names><airline_name>"AeroSucre"</airline_name></airline_names></airlines>
<airlines><country>"Kazakhstan"</country><airline_names><airline_name>"Air Kokshetau"</airline_name></airline_names></airlines>
<airlines><country>"Angola"</country><airline_names><airline_name>"Air Kissari"</airline_name></airline_names></airlines>
<airlines><country>"Ukraine"</country><airline_names><airline_name>"Aeronavigaciya"</airline_name></airline_names></airlines>
<airlines><country>"Kazakhstan"</country><airline_names><airline_name>"Alliance Avia"</airline_name></airline_names></airlines>
<airlines><country>"United States"</country><airline_names><airline_name>"Av Atlantic"</airline_name></airline_names></airlines>
<airlines><country>"Kazakhstan"</country><airline_names><airline_name>"Air Astana"</airline_name></airline_names></airlines>
<airlines><country>"Germany"</country><airline_names><airline_name>"City-Air Germany"</airline_name></airline_names></airlines>
<airlines><country>"Mexico"</country><airline_names><airline_name>"Aerovias De Lagos"</airline_name></airline_names></airlines>
<airlines><country>"Albania"</country><airline_names><airline_name>"Albanian Airlines"</airline_name></airline_names></airlines>
<airlines><country>"Mexico"</country><airline_names><airline_name>"Albisa"</airline_name></airline_names></airlines>
<airlines><country>"Russia"</country><airline_names><airline_name>"Voronezhskie Airlanes"</airline_name></airline_names></airlines>
<airlines><country>"Spain"</country><airline_names><airline_name>"Aero Madrid"</airline_name></airline_names></airlines>
<airlines><country>"Albania"</country><airline_names><airline_name>"Albatros Airways"</airline_name></airline_names></airlines>
<airlines><country>"Mexico"</country><airline_names><airline_name>"Aerolineas Aereas Ejecutivas De Durango"</airline_name></airline_names></airlines>
<airlines><country>"Germany"</country><airline_names><airline_name>"Line Blue"</airline_name></airline_names></airlines>
<airlines><country>"Lithuania"</country><airline_names><airline_name>"FlyLAL Charters"</airline_name></airline_names></airlines>
<airlines><country>"United States"</country><airline_names><airline_name>"Blue Sky America"</airline_name></airline_names></airlines>
<airlines><country>"United States"</country><airline_names><airline_name>"Texas Spirit"</airline_name></airline_names></airlines>
<airlines><country>"Russia"</country><airline_names><airline_name>"Aerologic"</airline_name></airline_names></airlines>
<airlines><country>"United States"</country><airline_names><airline_name>"Illinois Airways"</airline_name></airline_names></airlines>
<airlines><country>"Austria"</country><airline_names><airline_name>"Salzburg arrows"</airline_name></airline_names></airlines>
```

7) Airline Delay Details by Year (Organization Pattern)

Use of **Partitioning Pattern**.

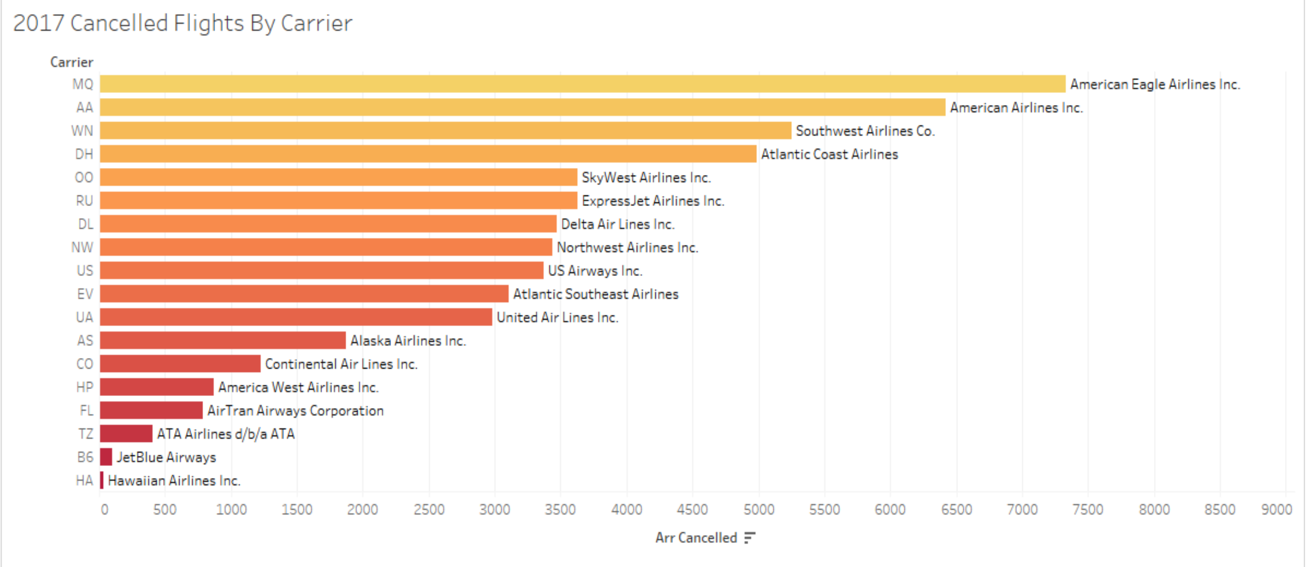
The airline delays data has data for years 2003-2017. This pattern allows us to partition the data by year so that we get a better picture about the airline delays for a particular year.

Result/Output:



```
2017,1,DL,Delta Air Lines Inc.,COS,"Colorado Springs, CO: City of Colorado Springs Municipal",22,7,3.58,0,2.55,0,0.87,0,0,419,255,0,76,0,88
2017,1,DL,Delta Air Lines Inc.,CMH,"Columbus, OH: Port Columbus International",275,48,23.85,7.4,5.6,0,11.14,6,2,4743,2201,1056,337,0,1149
2017,1,DL,Delta Air Lines Inc.,CLT,"Charlotte, NC: Charlotte Douglas International",431,75,29.37,9.28,18.26,0,18.08,3,0,4654,2176,644,662,0,1172
2017,1,DL,Delta Air Lines Inc.,CLE,"Cleveland, OH: Cleveland-Hopkins International",195,28,12.68,6.81,3.12,0,5.39,3,2,2828,1237,713,244,0,634
2017,1,DL,Delta Air Lines Inc.,CID,"Cedar Rapids/Iowa City, IA: The Eastern Iowa",31,6,1.95,0.79,2.39,0,0.87,1,0,966,150,139,537,0,140
2017,1,DL,Delta Air Lines Inc.,CHS,"Charleston, SC: Charleston AFB/International",286,53,25.75,8.71,6.07,0,12.48,3,0,4325,2033,695,380,0,1217
2017,1,DL,Delta Air Lines Inc.,CHO,"Charlottesville, VA: Charlottesville Albemarle",35,7,3,2.97,1.03,0,0,1,0,1320,407,895,18,0,0
2017,1,DL,Delta Air Lines Inc.,CHA,"Chattanooga, TN: Lovell Field",69,10,5.59,2.09,0.1,0,2.22,1,0,1147,725,150,16,0,256
2017,1,DL,Delta Air Lines Inc.,CAK,"Akron, OH: Akron-Canton Regional",109,11,2.3,2.31,2.23,0,4.16,2,1,1199,335,299,84,0,481
2017,1,DL,Delta Air Lines Inc.,CAE,"Columbia, SC: Columbia Metropolitan",85,13,6.4,0.54,0.89,0,5.17,0,0,694,311,30,57,0,296
2017,1,DL,Delta Air Lines Inc.,BZN,"Bozeman, MT: Bozeman Yellowstone International",71,17,8.28,2.34,3.9,0,2.49,0,0,1285,417,662,115,0,91
```

Conclusion: This partitioned data for the year 2017 shows that MQ, AA are the carriers with maximum flight cancellations.

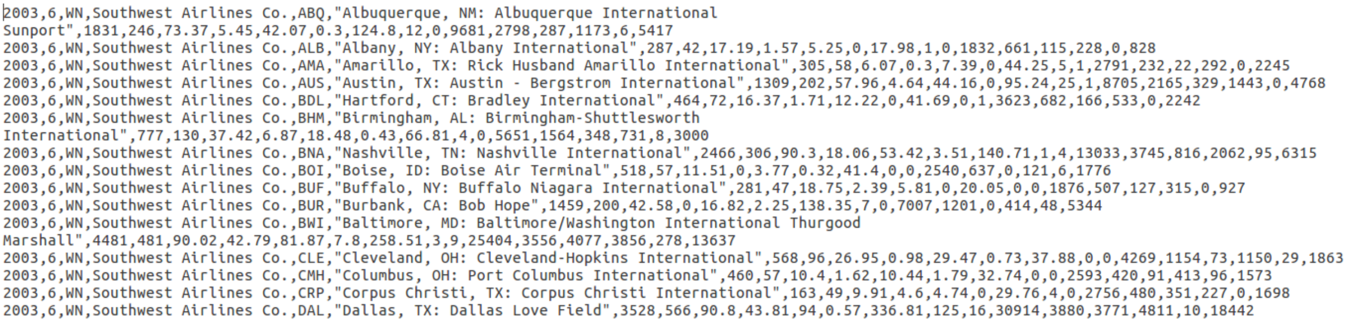


8) **Categorized Airline Delay Data by Carrier: (Organization Pattern)**

Use of **Binning Pattern**.

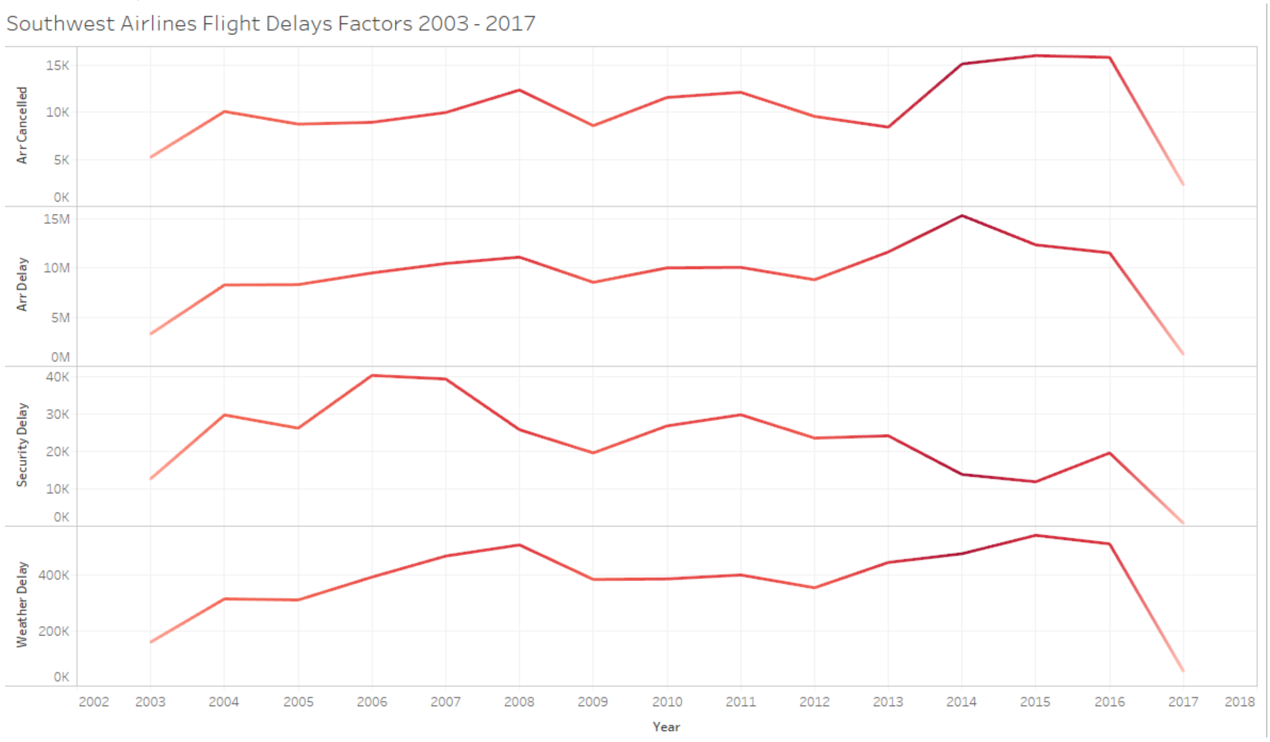
Create bins of data by categorizing the delay data by carrier. We make use of **MultipleOutputFormat** to write the different bins into HDFS.

Result/Output:



```
2003,6,WN,Southwest Airlines Co.,ABQ,"Albuquerque, NM: Albuquerque International
Sunport",1831,246,73.37,5.45,42.07,0.3,124.8,12,0,9681,2798,287,1173,6,5417
2003,6,WN,Southwest Airlines Co.,ALB,"Albany, NY: Albany International",287,42,17.19,1.57,5.25,0,17.98,1,0,1832,661,115,228,0,828
2003,6,WN,Southwest Airlines Co.,AMA,"Amarillo, TX: Rick Husband Amarillo International",305,58,6.07,0.3,7.39,0,44.25,5,1,2791,232,22,292,0,2245
2003,6,WN,Southwest Airlines Co.,AUS,"Austin, TX: Austin - Bergstrom International",1309,202,57.96,4.64,44.16,0,95.24,25,1,8705,2165,329,1443,0,4768
2003,6,WN,Southwest Airlines Co.,BDL,"Hartford, CT: Bradley International",464,72,16.37,1.71,12.22,0,41.69,0,1,3623,682,166,533,0,2242
2003,6,WN,Southwest Airlines Co.,BHM,"Birmingham, AL: Birmingham-Shuttlesworth
International",777,130,37.42,6.87,18.48,0.43,66.81,4,0,5651,1564,348,731,8,3000
2003,6,WN,Southwest Airlines Co.,BNA,"Nashville, TN: Nashville International",2466,306,90.3,18.06,53.42,3.51,140.71,1,4,13033,3745,816,2062,95,6315
2003,6,WN,Southwest Airlines Co.,BOI,"Boise, ID: Boise Air Terminal",518,57,11.51,0,3.77,0.32,41.4,0,0,2540,637,0,121,6,1776
2003,6,WN,Southwest Airlines Co.,BUF,"Buffalo, NY: Buffalo Niagara International",281,47,18.75,2.39,5.81,0,20.05,0,0,1876,507,127,315,0,927
2003,6,WN,Southwest Airlines Co.,BUR,"Burbank, CA: Bob Hope",1459,200,42.58,0,16.82,2.25,138.35,7,0,7007,1201,0,414,48,5344
2003,6,WN,Southwest Airlines Co.,BWI,"Baltimore, MD: Baltimore/Washington International Thurgood
Marshall",4481,481,90.02,42.79,81.87,7.8,258.51,3,9,25404,3556,4077,3856,278,13637
2003,6,WN,Southwest Airlines Co.,CLE,"Cleveland, OH: Cleveland-Hopkins International",568,96,26.95,0.98,29.47,0.73,37.88,0,0,4269,1154,73,1150,29,1863
2003,6,WN,Southwest Airlines Co.,CMH,"Columbus, OH: Port Columbus International",460,57,10.4,1.62,10.44,1.79,32.74,0,0,2593,420,91,413,96,1573
2003,6,WN,Southwest Airlines Co.,CRP,"Corpus Christi, TX: Corpus Christi International",163,49,9.91,4.6,4.74,0,29.76,4,0,2756,480,351,227,0,1698
2003,6,WN,Southwest Airlines Co.,DAL,"Dallas, TX: Dallas Love Field",3528,566,90.8,43.81,94,0.57,336.81,125,16,30914,3880,3771,4811,10,18442
```

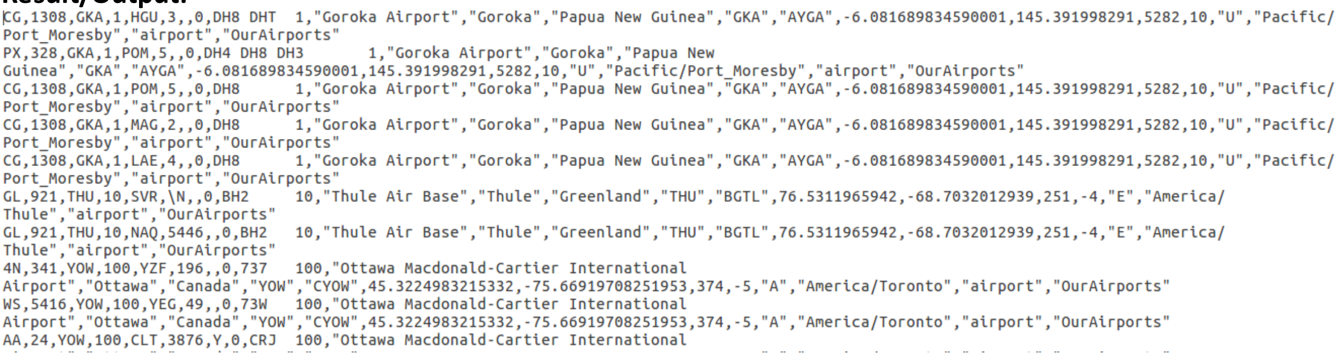
Conclusion: This pattern helps to analyse each carrier, and the data helps to get data over the years (2003-2017) for a particular airline (Southwest Airlines in the below case)



9) Inner Join: (Join Patterns)

Here we use the inner join pattern to merge two tables routes database with the airport database, so as to get one view for routes and airport details.

Result/Output:

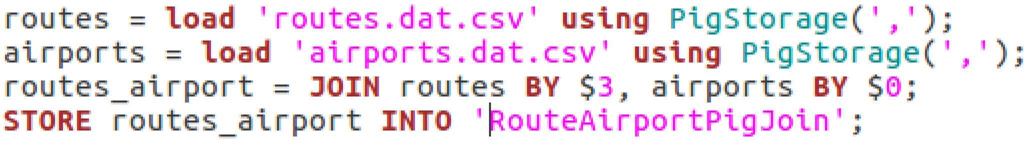


```
CG,1308,GKA,1,HGU,3,,0,DH8 DHT  1,"Goroka Airport","Goroka","Papua New Guinea","GKA","AYGA",-6.081689834590001,145.391998291,5282,10,"U","Pacific/
Port_Moresby","airport","OurAirports"
PX,328,GKA,1,POM,5,,0,DH4 DH8 DH3       1,"Goroka Airport","Goroka","Papua New
Guinea","GKA","AYGA",-6.081689834590001,145.391998291,5282,10,"U","Pacific/Port_Moresby","airport","OurAirports"
CG,1308,GKA,1,POM,5,,0,DH8      1,"Goroka Airport","Goroka","Papua New Guinea","GKA","AYGA",-6.081689834590001,145.391998291,5282,10,"U","Pacific/
Port_Moresby","airport","OurAirports"
CG,1308,GKA,1,MAG,2,,0,DH8      1,"Goroka Airport","Goroka","Papua New Guinea","GKA","AYGA",-6.081689834590001,145.391998291,5282,10,"U","Pacific/
Port_Moresby","airport","OurAirports"
CG,1308,GKA,1,LAE,4,,0,DH8      1,"Goroka Airport","Goroka","Papua New Guinea","GKA","AYGA",-6.081689834590001,145.391998291,5282,10,"U","Pacific/
Port_Moresby","airport","OurAirports"
GL,921,THU,10,SVR,\N,,0,BH2     10,"Thule Air Base","Thule","Greenland","THU","BGTL",76.5311965942,-68.7032012939,251,-4,"E","America/
Thule","airport","OurAirports"
GL,921,THU,10,NAQ,5446,,0,BH2   10,"Thule Air Base","Thule","Greenland","THU","BGTL",76.5311965942,-68.7032012939,251,-4,"E","America/
Thule","airport","OurAirports"
4N,341,YOW,100,YZF,196,,0,737   100,"Ottawa Macdonald-Cartier International
Airport","Ottawa","Canada","YOW","CYOW",45.3224983215332,-75.66919708251953,374,-5,"A","America/Toronto","airport","OurAirports"
WS,5416,YOW,100,YEG,49,,0,73W   100,"Ottawa Macdonald-Cartier International
Airport","Ottawa","Canada","YOW","CYOW",45.3224983215332,-75.66919708251953,374,-5,"A","America/Toronto","airport","OurAirports"
AA,24,YOW,100,CLT,3876,Y,0,CRJ  100,"Ottawa Macdonald-Cartier International
```

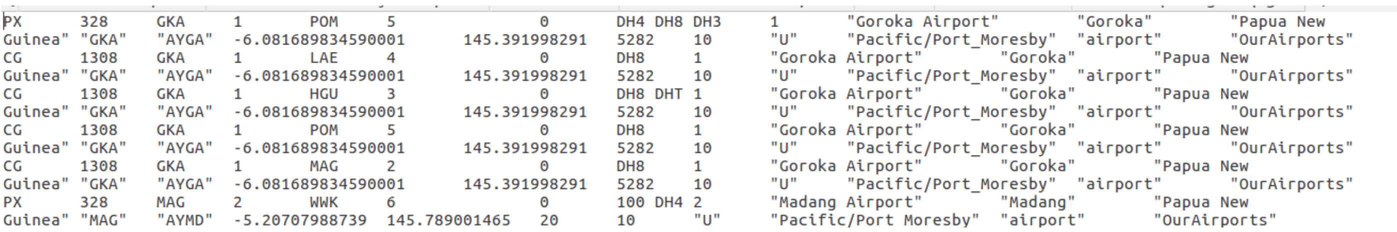
10) Pig (Inner Join):

This pig script performs an inner join on routes data and airports data. Pig is fast and easy to code. Pig is a Hadoop extension that simplifies Hadoop programming by giving you a high-level data processing language while keeping Hadoop's simple scalability and reliability.

Result/Output:



```
routes = load 'routes.dat.csv' using PigStorage(',');
airports = load 'airports.dat.csv' using PigStorage(',');
routes_airport = JOIN routes BY $3, airports BY $0;
STORE routes_airport INTO 'RouteAirportPigJoin';
```



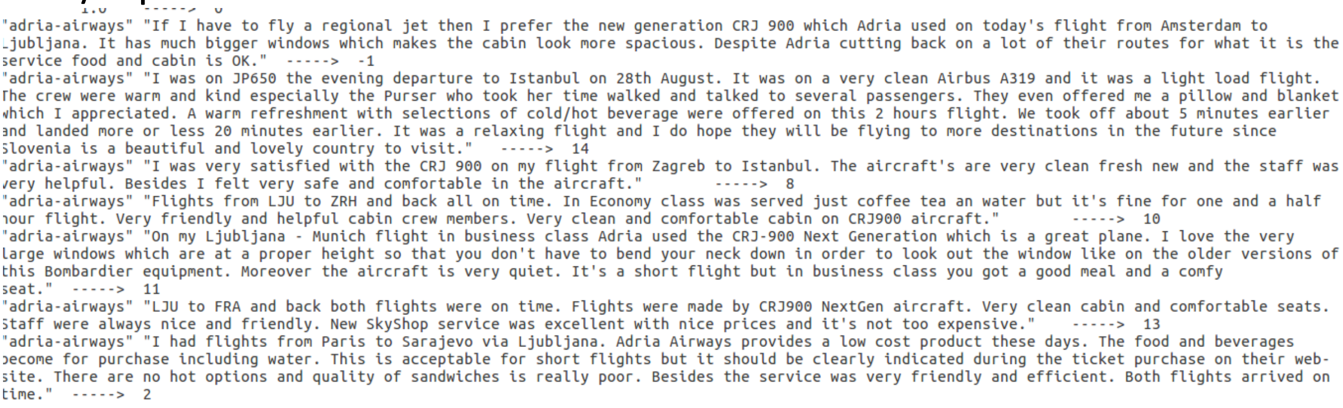
```
PX      328     GKA     1       POM     5               0       DH4 DH8 DH3     1       "Goroka Airport"        "Goroka"        "Papua New
Guinea" "GKA"   "AYGA"  -6.081689834590001      145.391998291   5282    10      "U"     "Pacific/Port_Moresby"  "airport"       "OurAirports"
CG      1308    GKA     1       LAE     4               0       DH8     1       "Goroka Airport"        "Goroka"        "Papua New
Guinea" "GKA"   "AYGA"  -6.081689834590001      145.391998291   5282    10      "U"     "Pacific/Port_Moresby"  "airport"       "OurAirports"
CG      1308    GKA     1       HGU     3               0       DH8 DHT 1       "Goroka Airport"        "Goroka"        "Papua New
Guinea" "GKA"   "AYGA"  -6.081689834590001      145.391998291   5282    10      "U"     "Pacific/Port_Moresby"  "airport"       "OurAirports"
CG      1308    GKA     1       POM     5               0       DH8     1       "Goroka Airport"        "Goroka"        "Papua New
Guinea" "GKA"   "AYGA"  -6.081689834590001      145.391998291   5282    10      "U"     "Pacific/Port_Moresby"  "airport"       "OurAirports"
CG      1308    GKA     1       MAG     2               0       DH8     1       "Goroka Airport"        "Goroka"        "Papua New
Guinea" "GKA"   "AYGA"  -6.081689834590001      145.391998291   5282    10      "U"     "Pacific/Port_Moresby"  "airport"       "OurAirports"
PX      328     MAG     2       WWK     6               0       100 DH4 2       "Madang Airport"        "Madang"        "Papua New
Guinea" "MAG"   "AYMD"  -5.20707988739  145.789001465   20      10      "U"     "Pacific/Port Moresby"  "airport"       "OurAirports"
```

11) Sentiment Analysis on Customer Reviews about Airline:

The customer reviews about the airline are analysed. Here we make use of **Distributed Cache** to perform sentiment analysis. By using distributed cache, we can perform map side joins. So, here we will join the dictionary dataset containing the sentiment values of each word. In order to perform Sentiment Analysis,we will be using a dictionary called AFINN.

AFINN is a dictionary, which consists of 2500 words rated from +5 to -5, depending on their meaning.

Result/Output:



```
   1.0    -----> 0
"adria-airways" "If I have to fly a regional jet then I prefer the new generation CRJ 900 which Adria used on today's flight from Amsterdam to
Ljubljana. It has much bigger windows which makes the cabin look more spacious. Despite Adria cutting back on a lot of their routes for what it is the
service food and cabin is OK."  ----->  -1
"adria-airways" "I was on JP650 the evening departure to Istanbul on 28th August. It was on a very clean Airbus A319 and it was a light load flight.
The crew were warm and kind especially the Purser who took her time walked and talked to several passengers. They even offered me a pillow and blanket
which I appreciated. A warm refreshment with selections of cold/hot beverage were offered on this 2 hours flight. We took off about 5 minutes earlier
and landed more or less 20 minutes earlier. It was a relaxing flight and I do hope they will be flying to more destinations in the future since
Slovenia is a beautiful and lovely country to visit."   ----->  14
"adria-airways" "I was very satisfied with the CRJ 900 on my flight from Zagreb to Istanbul. The aircraft's are very clean fresh new and the staff was
very helpful. Besides I felt very safe and comfortable in the aircraft."        ----->  8
"adria-airways" "Flights from LJU to ZRH and back all on time. In Economy class was served just coffee tea an water but it's fine for one and a half
hour flight. Very friendly and helpful cabin crew members. Very clean and comfortable cabin on CRJ900 aircraft."        ----->  10
"adria-airways" "On my Ljubljana - Munich flight in business class Adria used the CRJ-900 Next Generation which is a great plane. I love the very
large windows which are at a proper height so that you don't have to bend your neck down in order to look out the window like on the older versions of
this Bombardier equipment. Moreover the aircraft is very quiet. It's a short flight but in business class you got a good meal and a comfy
seat."  ----->  11
"adria-airways" "LJU to FRA and back both flights were on time. Flights were made by CRJ900 NextGen aircraft. Very clean cabin and comfortable seats.
Staff were always nice and friendly. New SkyShop service was excellent with nice prices and it's not too expensive."    ----->  13
"adria-airways" "I had flights from Paris to Sarajevo via Ljubljana. Adria Airways provides a low cost product these days. The food and beverages
become for purchase including water. This is acceptable for short flights but it should be clearly indicated during the ticket purchase on their web-
site. There are no hot options and quality of sandwiches is really poor. Besides the service was very friendly and efficient. Both flights arrived on
time."  ----->  2
```

Pig Analysis:

```
routes = load 'routes.dat.csv' using PigStorage(',');
airports = load 'airports.dat.csv' using PigStorage(',');
routes_airport = JOIN routes BY $3, airports BY $0;
STORE routes_airport INTO 'RouteAirportPigJoin';
```