Descriptive Analysis for domestic flights in the United States

Arnav Somani, Fernando Melchor, Hrafnkell Hjorleifsson, and Subhansu Gupta.
GitHub

***Abstract*—With the increasing rate at which data is being generated, it is necessary to develop skills and tools to better manage these large quantities of data. The value of these big data-sets will depend on the ability to perform on-time analysis, inference and the scalability of the developed pipeline. In this project, we use big data tools to analysis the USA domestic flight system and get facts and trends of interest to general public and stakeholders.**

***Index Terms*—Big Data, Spark, Flights, USA, Networks, Time-series, Airports, Airlines.**

I. Introduction

THOUSANDS of domestically flying airplanes take off and land everyday in America (around 31,460 to be precise). Rich data on each and everyone of these flights is publicly available and they make up for a very interesting dataset. Even though millions of people fly domestically every year, and might therefore have a vested interest in knowing more about this market, no publicly available analysis is to be found online. Knowing the amount of delay that is to be expected for a certain flight route might for example help someone make an informed decision on the layover time to allow for when booking a flight. Our goal is to draw an informative picture of domestic flights in America for a chosen number of years. To do so we will apply big data methods, running spark scripts in a cluster environment, to extract chosen information for output and visualization. Furthermore we will be looking for trends over the time period of the data and looking into airport and state connectivity through network analysis. All this extracted information will make up for a static factbook on the domestic aviation industry in America. A factbook such as this could quench the information thirst of curious readers and hopefully help inform customers about the performance of airports and airlines in more effective way.

A. Data

The data used was threefold; aviation data, geospatial information for airports and geospatial information on states.

- The aviation data comes from the United States Department of Transportations Bureau of Transportation Statistics. [1]
  The data is filtered by month with the option of downloading data for individual states or the whole United States at once. For this project data was collected for the whole year of 2014 through 2016 for the whole country. A month of data is between 200 and 250 mb so this whole dataset was a little less than 8 gb in size. Each row of data represents a flight and for each flight there are 110 different attributes such as airline, airport of origin, destination airport, scheduled arrival time, actual arrival time, reason for delay and so forth. Total number of rows/flights are around 17 million.
- The geospatial information for the different airports, it contains a list of all identified airports in the USA and their Airport Codes, information about the longitude and latitude is included in the file and as well information about the city where the airport is located.[2]
- The geospatial information for the different states comes from the USA Census Bureau, it contains the geographical political boundaries of all the USA territories. [3]

B. Big Data Challenges

Filtering and aggregating 17 million rows of data in various ways is computationally heavy and time consuming. By filtering and aggregating the data with spark scripts running on a cluster environment we minimize it by a magnitude of $10^4 - 10^6$ depending on the granularity of the outputted csv files. Pivot tables, that are computationally heavy to process, were created in order to analyze the daily flights like time-series grouping them by airline, airport or city. This output csv files are very useful for different kinds of analysis and they can be processed without the need of a big computational power. We designed the outputs of each script to be efficient and easy to handle for analysis proposes. We enrich our outputs by merging them with more specific information about the airports, like geo-location, names and cities. We divided our scripts depending on the size of the tasks, when performing the pivot table for daily flights per airport it was not possible to add other tasks inside that script because of memory problems. Bellow we show a table 1 with some statistics of our scripts.

Script Name	jobs	time	outputs
1_Scripts_collection_nopivot	16	1.3	3
2_Scripts_collection_delays	10	1.7	5
3_busiest_airport_day	8	1.3	1
4_Script_collection_others	4	1.2	1
Total	38	5.5	10

Fig. 1. Running time of 4 principal scripts.

As it is possible to observe, the speed of processing the almost 8GB of data is great. This could encourage other people to run the analysis for the last 10 years instead of 3, if they see some value in going beyond the last 3 years. Some of the challenges that we faced during this project was the

version tracking of files, when working with several persons in the cluster. These could be mitigated by using a service like GitHub since the beginning. Another important thing to consider is to keep track of changes when updating files and troubleshooting inside the terminal. It is important to highlight that some well-known numbers or counts are needed to verify the output of any given script given the complexity detecting an error in the output is not easy. As final comment we believe that the interface could be improved to make it easier to follow relevant printed out comments when submitting a job to the cluster.

II. Methodologies

Aviation data was downloaded from the United States Department of Transportations Bureau of Transportation Statistics. The data came in 36 separate csv files for the 36 months that were under inspection. These files were then moved to and stored on the NYU dumbo cluster. Separate scripts were written for each targeted output and ran on the cluster. These scripts would then output the desired results as csv files. For example one script would output the count of flights departing and arriving at each airport for a time period of a month (busiest airport per month) and another one would output that count for each day of the dataset (busiest airport per day). The aim being to have all data aggregation and grouping done on the cluster so to produce results those tables could be presented as is or visualized in a very easy manner. The structure of the scripts were all the same: yielding the desired datacolumns in their raw form into a pyspark.sql.dataframe. The columns were then grouped, summed, counted or normalized depending on what output was needed. Run time was around 2 min depending on the granularity of the output. For the propose of delivery of the project we grouped some scripts together in other to facilitated reproduction. To perform network analysis airport and state aggregations from the aviation database were merged with airport lat/lon information and state shapefiles. This merging was done in the dumbo cluster. The network analysis itself was performed locally using networkx package [4]. Some time series analysis was performed locally with outputted data. Seasonal decomposition was done with the statsmodels.api [5] package in python.

1) Outcome: As is to be expected when constructing a factbook the outcomes were numerous. Some key findings:

- The busiest (no. landings and takeoffs) airport for domestic flights in America, by some margin, is Hartsfield-Jackson Atlanta International Airport in Atlanta Georgia, figure 2.
- Somewhat obviously the busiest city is Atlanta, Georgia, though Chicago, Illinois, recorded more flights for the period June 2014 through December 2015, figure 3. Interestingly a drop in flights through Chicago can be seen in January 2016. The reason for this is not clear though a quick internet search seems to suggest that at least a part of the reason might be due to employment of larger aircraft vessels. New York is the 7th busiest city in America but if Newark airport is included it rises to 3rd.
- The busiest route by some margin is Los Angeles - San Francisco, followed by Los Angeles - New York and Los Angeles - Las Vegas, figure 4. These are calculated irrespective of flight direction but there is a close to 1:1 matching between all legs.
- The airport that has seen the most increase in flights is SeattleTacoma International Airport and the airport that has seen the most decrease is Dallas/Fort Worth International Airport, figures 5 and 6. The same results are found when looking at the data on city level.
- The route that saw the most increase in traffic was Dallas - LaGuardia and the route that saw the most decline in traffic was Los Angeles - San Diego, figures 7 and 8.
- The Airline with the maximum flights is Southwest 14.
- Major reason for delay of Hawaiian Airline is Carrier delay with appx 60% as compared to other reasons. Weather effect results to be minimum 12.
- The major reason of overall delay is due to air control, late air-crafts and airline problems. We can infer that delayed aircraft is the effect of previous delays 13.

The airport that has seen the most increase in flights is SeattleTacoma International Airport and the airport that has seen the most decrease is Dallas/Fort Worth International Airport. The trend holds when looking at the data on city level. The route that saw the most increase in traffic was Dallas - LaGuardia and the route that saw the most decline in traffic was Los Angeles - San Diego. Time series analysis reveals cyclical seasonality of the data (describe further). Event detection (looking at airline traffic more than 2 standard deviations from the mean) revealed that the slowest day of the year, and the only one that can be labelled an anomaly, is thanksgiving.

For the network analysis, we used the total routes to get all the airports that were connected through flights, in this way a network with 334 airports/nodes was created to represent the whole system 15. To better analyze this system, the network was reduced by grouping the airports by states, this was done by using the states shapefiles [3] and performing the geometric operation of point with in a polygon to determine if certain airport was within a state. This was necessary because the airport data that we had did not have a feature identifying the state of the airport. After reducing the network by state the network had 55 airports/nodes 16. With this reduced network, we measured the average shortest path length from one origin to all the possible destinations. Each state had a measure of connectivity that could be interpreted as the average number of connecting flights needed to go from a particular state to all the others. 17. Most connected states are Illinois, Georgia, Texas, Colorado and Minnesota. The least connected states/territories are Northern Mariana Islands, American Samoa, Guam, Mississippi and Delaware.

III. Conclusion

For us dividing the problem in small tasks prove to be effective, even that running scripts one by one takes longer because the request time in the cluster, it was must easier to

troubleshoot and get the code working. At the end we grouped some scripts together to facilitate the review of the project. After this first experience with high performance computing it is possible to envision much more bigger projects, in size and complexity.

The Domestic Flights data-set is pretty rich and a lot of information can be extracted from it. The files that we created as an output are useful for certain proposes and further analysis can be done, especially when thinking about regional economics. The impact of economic booms and declines can be spotted in the flows of passengers to different areas of the country.

In other perspective the data that indicates airline's performance is valuable for users and should be available in a more easy and interactive way.

APPENDIX
FIGURES AND PLOTS.

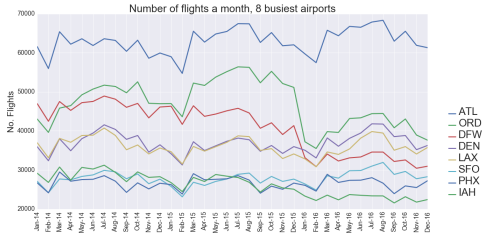

Fig. 2. The eight busiest airports in America for the time period examined.

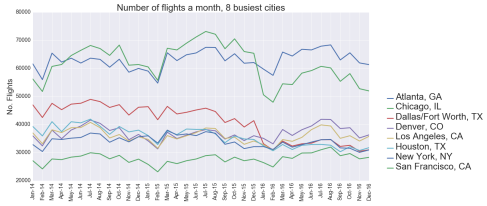

Fig. 3. The eight busiest cities in America for the time period examined.

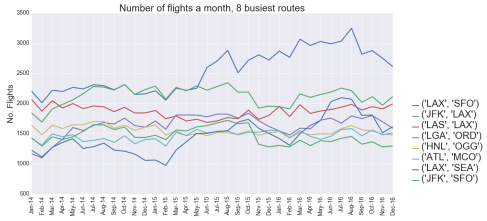

Fig. 4. The eight busiest routes in America for the time period examined.

REFERENCES

[1] "Bureau of transportation statistics." [Online]. Available: https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

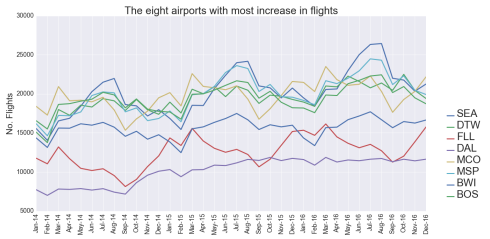

Fig. 5. The airports that saw the most increase in traffic for the time period examined.

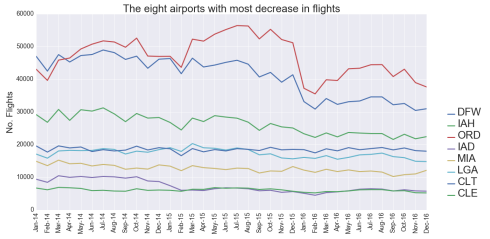

Fig. 6. The airports that saw the most decrease in traffic for the time period examined.

[2] "Airport location and codes around the world." [Online]. Available: https://community.tableau.com/thread/156711

[3] "United states census bureau, cartographic boundary shapefiles." [Online]. Available: https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html

[4] "Networkx documentation, python package." [Online]. Available: https://networkx.readthedocs.io/en/stable/

[5] "Statsmodelss documentation, python package." [Online]. Available: http://www.statsmodels.org/stable/index.html

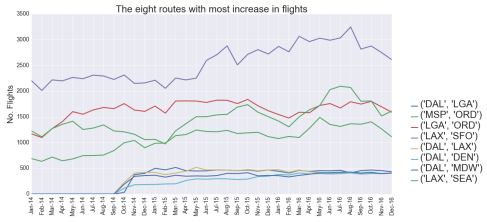

Fig. 7. The routes that saw the most increase in traffic for the time period examined.

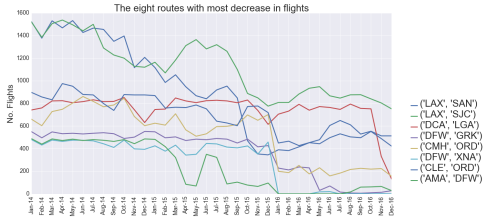

Fig. 8. The routes that saw the most decrease in traffic for the time period examined.

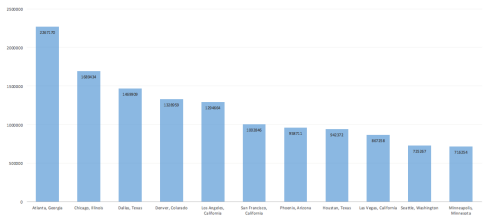

Fig. 9. The busiest airports.

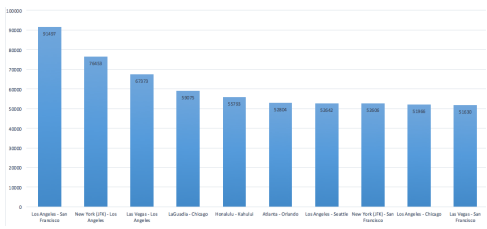

Fig. 10. The busiest routes.

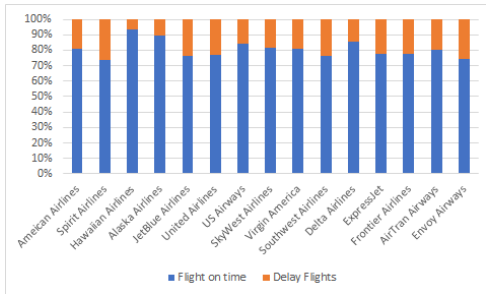

Fig. 11. Delays per airline.

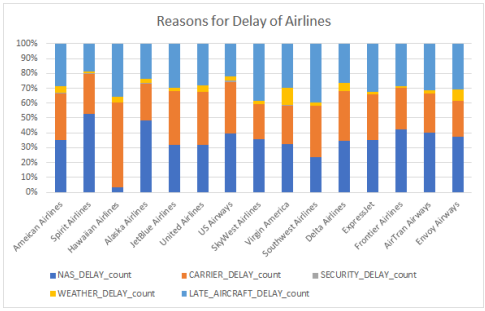

Fig. 12. Delays reasons per airline.

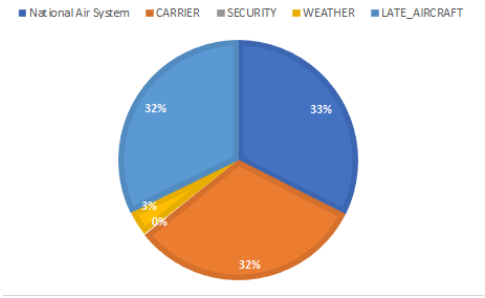

Fig. 13. Total delays reasons.

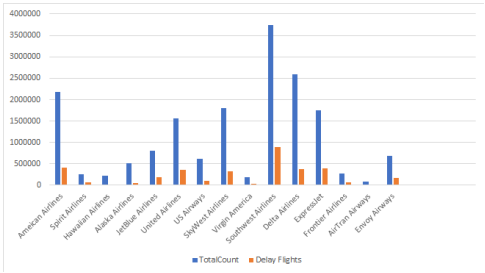

Fig. 14. Total flights per airline and delays.

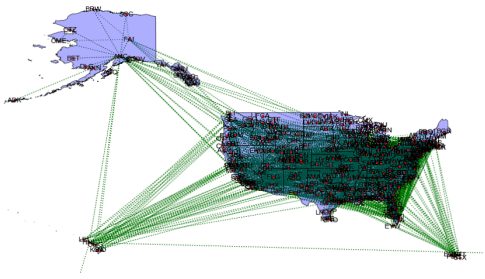

Fig. 15. Complete USA domestic flights network with 334 airports/nodes

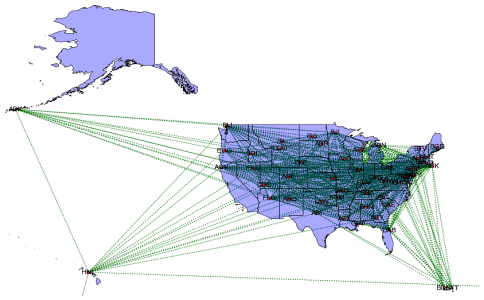

Fig. 16. Reduced USA domestic flights network with 55 airports/nodes

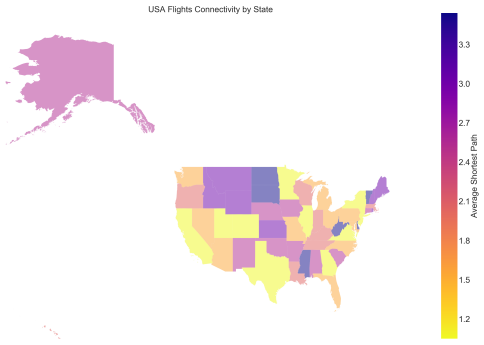

Fig. 17. USA States connectivity score based on the flights to all the other states.