Aviation Performance Metrics: An in-depth analysis CS526 Spring 2020

Kaushal Patel, Qilun Lyu, Tathagata Dutta
Rutgers University
Piscataway, NJ, USA
Email: kpp98@rutgers.edu, ql193@rutgers.edu, td425@rutgers.edu

Abstract— Aviation Performance Metrics (APM) provides tourists or passengers, airport authorities and airline companies with a platform for being better aware of the aviation industry through some common benchmarks. It answers when, where and why the delays, as well as cancellations happen the most. Moreover, it bridges the gap between humans and machines by interactions such as hovering, zooming, brushing, panning and many others.
We provide the user with a complete set of visualization elements to better understand the performance of different components in the aviation industry. These include Time series of delay and cancellation rates, a map where users can interact with different airports(and get further information), a chord diagram to show inter airport delays, fully interconnected statistical modules to analyse airline delays and cancellations and many other utilities. All of these being interactive helps the user to visualize his analysis.

I. Introduction

Delay or cancellation of flights is annoying not only for the passengers but also for the airport and airline crew. Such delay or cancellation leads to loss for all the involved parties. The passenger loses time and money and consequently the airport/airline crew have to face their complains and whining, which is not a pleasant experience for them. Moreover, their reputation also gets a hit. Many people consider delay as the primary indicator of the rating of an airline (Quality and Price being the other indicators). Delays due to unavoidable circumstances like hazardous weather conditions or matters of national security cannot be prevented, but there might be other reasons which can be managed and administered to avert flight delays. The first step towards such a task requires visualization and analysis of large-scale flight data. Our project provides an interactive and visual platform for the above analysis using the flight data starting from January 2015 till December 2019 (latest available data during the start of the project) as found in Bureau of Transportation Statistics website. The data covers the all domestic flights in the United States of America. This project can greatly help passengers, airport authorities and airline companies, be it to choose the best from available choices or to improve the performance/punctuality of flights.

The project has four important Stages: The Requirement Gathering Stage, Design Stage, Implementation Stage and User interface.

II. The Requirement Gathering Stage

The data set[1] we are going to analyse is directly downloaded from Bureau of Transportation Statistics, a website of United States Department of Transportation. It consists of 60 Months of Flight data, departing from or arriving at the United States. The data used for our project is at rest. It covers all the 50 states of United States including Puerto Rico US virgin Islands and also other US possessions and territories but excluding Washington D.C. The data is time variant as it ranges from January 2015 to December 2019. The total size of these CSV(comma-separated-value) files is around 13 gigabytes with more than 31 million records.

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

ID	FlightDate	IATA_Airline	Tail_Number	Flight_Number	Origin	Dest	CRSDepTime	CRSArrTime	CRSElapsedTime	Distance	Canc
1	2015-01-01	AA	N787AA	1	JFK	LAX	0900	1230	390	2475	0
2	2015-01-02	AA	N795AA	1	JFK	LAX	0900	1230	390	2475	0
3	2015-01-03	AA	N788AA	1	JFK	LAX	0900	1230	390	2475	0
4	2015-01-04	AA	N791AA	1	JFK	LAX	0900	1230	390	2475	0
5	2015-01-05	AA	N783AA	1	JFK	LAX	0900	1230	390	2475	0
6	2015-01-06	AA	N799AA	1	JFK	LAX	0900	1235	395	2475	0
7	2015-01-07	AA	N784AA	1	JFK	LAX	0900	1235	395	2475	0
8	2015-01-08	AA	N787AA	1	JFK	LAX	0900	1235	395	2475	0
9	2015-01-09	AA	N795AA	1	JFK	LAX	0900	1235	395	2475	0
10	2015-01-10	AA	N790AA	1	JFK	LAX	0900	1235	395	2475	0
11	2015-01-11	AA	N786AA	1	JFK	LAX	0900	1235	395	2475	0
12	2015-01-12	AA	N790AA	1	JFK	LAX	0900	1235	395	2475	0
13	2015-01-13	AA	N796AA	1	JFK	LAX	0900	1235	395	2475	0
14	2015-01-14	AA	N793AA	1	JFK	LAX	0900	1235	395	2475	0
15	2015-01-15	AA	N790AA	1	JFK	LAX	0900	1235	395	2475	0
16	2015-01-16	AA	N797AA	1	JFK	LAX	0900	1235	395	2475	0
17	2015-01-17	AA	N795AA	1	JFK	LAX	0900	1235	395	2475	0
18	2015-01-18	AA	N786AA	1	JFK	LAX	0900	1235	395	2475	0
19	2015-01-19	AA	N785AA	1	JFK	LAX	0900	1235	395	2475	0
20	2015-01-20	AA	N797AA	1	JFK	LAX	0900	1235	395	2475	0
21	2015-01-21	AA	N789AA	1	JFK	LAX	0900	1235	395	2475	0
22	2015-01-22	AA	N790AA	1	JFK	LAX	0900	1235	395	2475	0
23	2015-01-23	AA	N790AA	1	JFK	LAX	0900	1235	395	2475	0
24	2015-01-24	AA	N787AA	1	JFK	LAX	0900	1235	395	2475	0
25	2015-01-25	AA	N791AA	1	JFK	LAX	0900	1235	395	2475	0

Fig. 1. Data Snippet

Our web application will provide the analysis related to route delay and cancellations of airports and airlines with relevant visual plots and also user interaction to analyse the delay as well as cancellation and also the factors behind them.

The different types of users for our applications are:

- Passengers/Consumers: A person who want to travel from one destination to another in US using Domestic flights will find out application useful as he/she can look at the delays of airlines,airports and cancellation rate of airlines and airports to plan his bookings.Using

[1]See https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236& DB_Short_Name=On-Time

our application passenger can know the delay trend ,delay in routes,On time rate of airports and airlines which will help him to know from which Airport he/she should book the flight and in which airline:

- Scenario 1 Description: Passenger A wants to know the delay between LAX and EWR.
- System input for Scenario 1: Go to the Route Delay page under Airport Menu. In the chord diagram on the left, select the desired airports.
- System output for Scenario 1: A path will appear between them. This path shows the delay between the chosen airports (both directions)
- Scenario 2 Description: Passenger B wants to know the reason behind cancellations for a particular date.
- System input for Scenario 2: Go to the cancellation page under Airline Menu and select the desired filters Finally click on any of the bars in the bottom bar chart and subsequently click on the Reason for Cancellation button.
- System output for Scenario 2: The various reasons that may have caused the cancellations will appear in a google page, from where the user can select the ones he wants to delve into further.
- Scenario 3 Description: Passenger C wants to know how to go to a specific airport or to get more information about that airport
- System input for Scenario 3: Go to Map page under Airports and either select the desired airport from the map itself or search for it using the search bar. Hover of the highlighted airport and an info box will appear. Click on "How to get there?" or "More Info".
- System output for Scenario 3: "How to get there?" - Google Maps with directions from current location to the desired airport. "More Info" - Wiki Page containing further information.

- Airport Authorities: Airport Authorities can use our application to get know the On-time performance of the Airport and also get useful and relevant statistics and plots for the Airport. Also they can visualize important trends for delay and cancellation for their Airports.
  - Scenario 1 Description: Airport Authority of JFK wants to know the Average Delay, On-Time Rate and/or Cancellation Rate.
  - System input for Scenario 1: Go to Map page under Airports and either select JFK from the map itself or search for it using the search bar.
  - System output for Scenario 1: On the right panel, all the required information will be provided.
  - Scenario 2 Description: Airport Authority of LAX wants to know the pattern of delays over time (to better manage the airlines and reduce delay)
  - System input for Scenario 2: Go to Map page under Airports and either select LAX from the map itself or search for it using the search bar.
  - System output for Scenario 2: Various statistical data will appear on the bottom half of the page. From there, they can gather information regarding the pattern of delays, which years, months and days have higher traffic and delays.
- Airline Companies: Airline Companies can find our application useful to get causes for delay in their airlines, to compare the performance with other Airlines. They can also make analysis regarding the delays as well as the cancellation and also reasons and factors behind that using the functionalities provided in our project.
  - Scenario 1 Description: Alaska Airlines want to know the reason for cancellation behind their cancellation in February 2015.
  - System input for Scenario 1: Go to Cancellation page under Airlines menu. Select Alaska Airlines from airlines bar chart, select Feb from month bar chart and select 2015 from year bar chart. Click on reason for cancellation button on the right.
  - System output for Scenario 1: The various reasons that may have caused the cancellations will appear in a google page, from where they can select the ones they want to delve into further.
  - Scenario 2 Description: American Airlines want to know which Month and Year had the highest delay for LateAircraftDelay.
  - System input for Scenario 2: Go to Overall Comparison page under Airlines menu. Select American Airlines from airlines bar chart and LateAircraftDelay from the pie chart.
  - System output for Scenario 2: The bar chart at the bottom will show the distribution of delays and the user can view which months and years had higher delays compared to the rest.

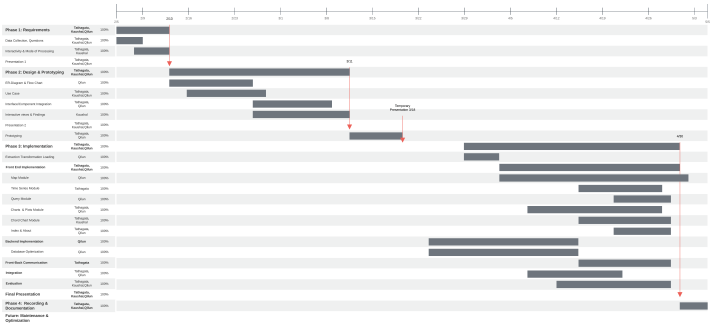

Fig. 2. Gantt Chart

III. The Design Stage

Following are some aspects of the design implementation of our application.

A. Data Processing

We have identified the following entities, attributes and relationships for given data as shown in Fig. 3 in the ideal Entity Relationship Diagram (ERD):

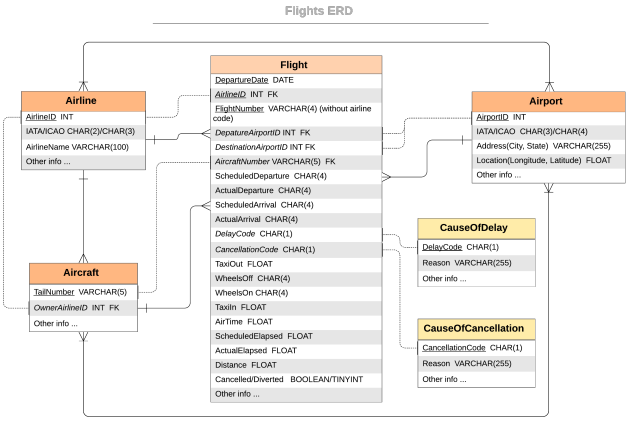

Fig. 3. ERD For The System

In the given figure, it clarifies the entities and the relations between them. Airline, Airport, Aircraft, Flight, CauseOfDelay and CauseOfCancellation are evident entities. For relations, Flight presents many:one with others, which is the same as Aircraft with Airline. Also, Airport indicates one:many with Aircraft and many:many with Airline, and similar is the case with Aircraft. Every chart contains no anatomic domain, at least one set of unique identifiers as a primary key, as well as no dependency between any other non-primary attributes. The whole diagram therefore satisfies Boyce-Codd NF[?] or 3.5NF.

Although function reflections seem to appear among attributes, we still cannot deal with them by simple arithmetic due to lack of "common-sense" information in the data set, like time zone. But when it comes to getting the results of time intervals at the same location, we can easily figure out whether the flight has been delayed or not, by computing the difference between scheduled departure time and actual departure time, as well as, similarly, between scheduled arrival time and actual arrival time.

In the project, however, we decided to lower the paradigm (i.e. Normal Form) to 3NF for less computations and faster retrieval, since the scale of data for presenting is significantly large, which in other word we keep preserving dependent non-primary attributes (e.g. delay) that have already existed in our data. After all, we can easily calculate the duration such as elapsed time, airtime, etc. Note that we only focus on flights departing from or arriving in US and Canada, and only care about normal flights, delayed flights and cancelled flights. The situation of diverted ones is not taken into consideration for this study.

Later on, we'll be using basic statistics for visual analytics, such as average delay of airlines, on-time rate of airports, frequency of flights of the two, and density of airports within a region, etc.

B. Context Diagram

Fig. 4 represents the Context Diagram of our Project.

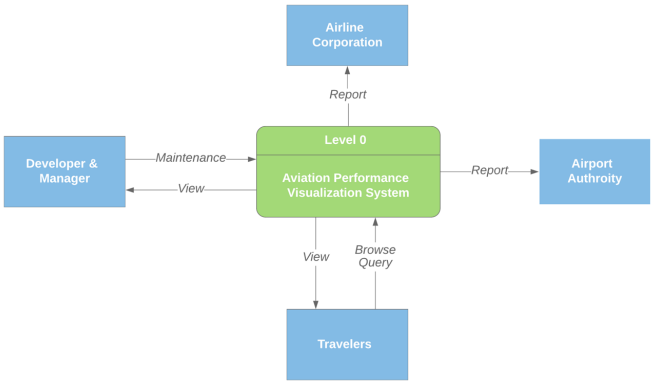

Fig. 4. Context Diagram

The diagram is the level 0 context diagram of our project.This includes the four major entities i.e. Developers which maintains the applications while Airline Authorities, Travellers , Airport Authorities can perform interaction and Queries to get suitable information from the data.

C. System Flow Diagram

Following is the System Flow Diagram of our Project.

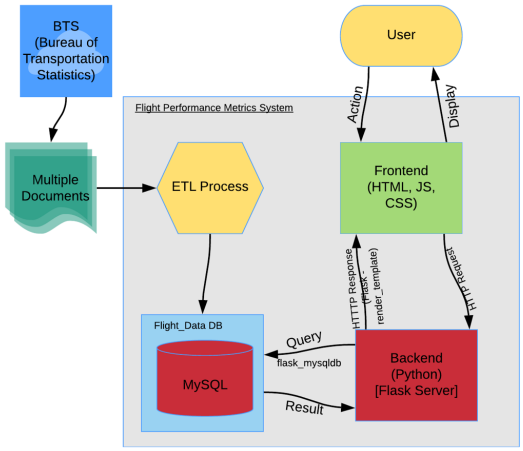

Fig. 5. System Flow Diagram

Fig. 5 shows the System Flow diagram of our project. User can perform actions on our application which will be updated in the front-end based on the actions by the users which will request back-end to get data from the database. Database will give result to back-end which will update the front-end view for the user. The data is stored

in our MySQL database using ETL process(Extraction Transformation and Loading).

IV. The Implementation Stage

The implementation was done in three phases: 1) Front-End implementation 2) Back-end implementation 3) Integration of Front-end and Back-end

A. Mode of Processing

For front end majorly we used HTML,CSS,and Javascripts. Major libraries used for various plots are:

1) D3.js
2) DC.js
3) crossfilter.js
4) Amcharts
5) Google Maps Platform

For back-end and integration with the front-end we used following tools and technologies:

1) Flask
2) Pyspark
3) JQuery
4) Bootstrap

B. Interesting Findings

There are many questions that can be answered using our application. From that some of the interesting findings were as follows:

- Airports in the east coast have more connections compared to west coast (Fig. 7).
- There is a positive correlation between Departure Delay and Arrival Delay, which is expected.
- There is a negative correlation between Cancellation and Delay, which is expected (Fig 6).
- Late Aircraft Delay, Carrier Delay and NAS delay are the main reasons of delay.
- Among the big name airlines, Alaska Airlines performs the best.

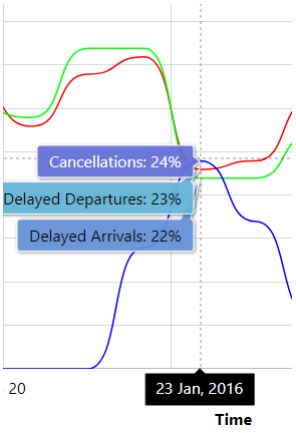

Fig. 6. Interesting Findings

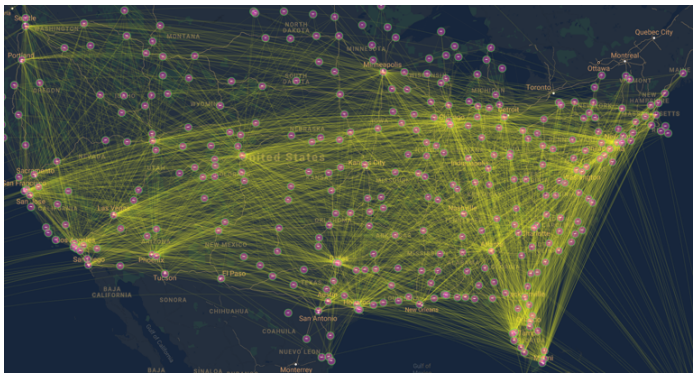

Fig. 7. Interesting Findings

V. User Interface

As you open our application you will see our homepage with an overview of what is being offered. There are links to the dashboard on the homepage. As you enter the dashboard you will have five tabs:

- Time Series: We have used line charts using amcharts. Delays and cancellation rates are visible for each day upon hovering. Also user can zoom in using the horizontal and vertical sliders. Zooming can also be achieved by selecting an area from the screen. The sliders can also be used as a brush to pan over the entire time series. There is also a reset button to reset the view to its original state.
- Map Module: Basic Visualization used here are bar plots, Line plots, Stacked Area Plots and A Map. The darkness of a dot represents the punctuality of the corresponding airport. And the size indicates its scale, measured by the number of destinations that are visible after clicking on an airport. In the Flight Route Mode, size of the plot becomes less important; however, detectable paths are suitable for experts to recognize its traffic pattern of the network especially in a dimmer background. But, as for a single direction of a particular path, one may refer to inter airport delays, where enables one to research it in depth.
- Route Delay: We have used Chord diagram to represent the route delay between two or more Airports. We used amcharts library. Here, the user can select any number of airports and paths appear between them. Hovering over a path will show the delay between the airports. Also user can click on more info while hovering over an airport node to view more details about that airport.
- Airline Module: Here there are three sub modules:
  - Airline Delay Analysis
  - Airline Delay Analysis Month-wise
  - Airline Cancellation Analysis

  The basic visualization used here are Pie charts,Bar charts and Row charts. We have used D3.js, dc.js and crossfilter.js library for this module where d3.js is used

for plotting, crossfilter.js is used for manipulating data and dc.js to connect d3.js and crossfilter.js. This module provides interactivity like mouse hovering, mouse selection, mouse clicks, linking of different views, User annotations. Colors used are independent of their usual meaning.

- Misc Plots: Here, one can see various statistical plots to identify trends in delays over:
  - Locations (state, city)
  - Airports
  - Airlines
  - Time (year, month, day of week, day of month)

VI. Acknowledgements

We would like to express our appreciation to Prof. James Abello for guiding and helping our team to a successful completion of the project. Assistance given by TA Nishant Kumar for JavaScript Libraries is greatly appreciated. We would like to offer our special thanks to TA Fangda Han for important suggestions during different phases of our project. We would also like to thank Enkai Ji for his input in different UI elements of the front-end.

References

[1] https://www.transtats.bts.gov/ONTIME/
[2] https://flask.palletsprojects.com/en/1.1.x/
[3] https://spark.apache.org/docs/latest/api/python/index.html
[4] https://developers.google.com/maps/documentation/javascript/tutorial
[5] https://api.jquery.com/
[6] https://getbootstrap.com/
[7] https://www.amcharts.com/demos/chord-diagram/
[8] https://d3js.org/
[9] https://dc-js.github.io/dc.js/
[10] https://square.github.io/crossfilter/
[11] http://dimplejs.org/