Airport Delays Database Analytics
(March 2023)

Daniel Chang | James Khamthung | Robert Blue | Santiago Gutierrez

***Abstract* — Flight delays have become increasingly prevalent in the aviation industry, causing economic losses for airlines and passengers alike. This study aims to identify the causes of flight delays and provide effective solutions to improve airport service quality and minimize delays. Our analysis examines various types of delay, including late aircraft arrival, airline delays, weather, and security, among others. Ultimately, the results of this study can provide valuable insights for airlines, airports, and transportation authorities to optimize their operations, enhance the passenger experience, and minimize flight delays.**

***Keywords*— DBMS, SQL Server, MySQL, etc.* For a list of suggested keywords, send an email to gutierrezs@smu.edu**

I. INTRODUCTION

Flight delays are a significant issue in civil aviation. They incur direct and indirect costs, such as gate maintenance, crew fees, food service, and lodging. They also have an impact on passenger satisfaction. An airport is a maintenance and transit hub where flights begin and end. Inaccurate flight delay forecasts will result in losses for aviation-dependent industries and passengers, while delays will harm the transportation network's service capacity and cause delays at other airports.

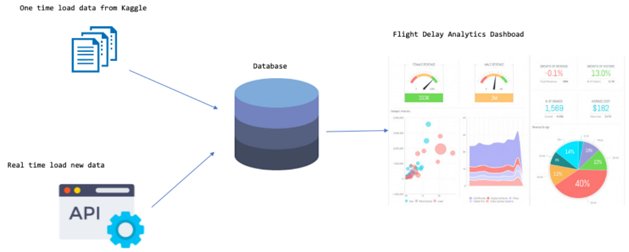
The untimely arrival of aircraft (both earlier and later than expected) has significant implications for airport management, including air-system delays, security delays, etc . As a result, forecasting and analyzing flight delays is critical for airlines, passengers, and airports. Flight punctuality is an important criterion for evaluating airlines and airports. Predicting delays allows airports to adjust resource allocations, quickly analyze causes, and implement measures to reduce or eliminate delays.

We want to help analyze flight delays that can be avoided by combining flight, air traffic control, and weather data. This study decomposes and locates the causes of real-time delays to provide a foundation for managers to make decisions.

Delays caused by direct and indirect factors should be treated differently in forecast results. Delays caused by direct factors, such as weather or holidays, can only be planned beforehand. However, the management can avoid the delay caused by indirect factors by taking some precautions, such as ensuring the quick exit of previous flights.

Meanwhile, because the number of influencing factors that managers can deal with over time is limited, a model that can output the most critical factors (or time points) can save humans and equipment during the delay-handling process. As a result, managers must identify critical time points or factors to reduce or eliminate tactical delays.

II. METHODOLOGY



A. *Database and Dataset*

We will begin by developing a database and loading the dataset into the database. Our dataset is mostly taken from Kaggle and contains flight delays and cancellations collected and published by the DOT's Bureau of Transportation Statistics. More information is provided in the Data Collection section.

B. *API Data*

We will then create a script to pull data from API and insert new data into the database. We use the aviationstack API. It is a RESTful API that allows developers to access real-time data and historical flight data from all the airlines around the world. It provides us with a lots of information, including flight status, departure times and arrival times, flight routes, etc.

The API returns data in JSON format that we can transform and extract the data using Python in Jupyter notebook, then we will extract the data involves parsing the JSON response and specific pieces of information that we need. That includes flight numbers, departure and arrival times, airport codes, and other information similar to what we have from Kaggle data.

From there, after we manage to insert extracted information into our MySQL database, we will create queries and data pulls into separate SQL tables in order to display the real-time data. Finally, we can use these specific queries in order to split our dataset into factors that airports can directly improve on vs extraneous factors.

C. *Queries and Dashboards*

Followed by analytics queries and dashboards. These tools will allow us to more easily investigate new/all data through the use of pre-built, reproducible models and visualizations. We will use our queries to build dashboards in order to display a real-time visual of flight times and current delays. Our goal is to be able to provide a quick look and real-time view for both passengers as well as airport analysts.

III. DATA COLLECTION

The U.S. Department of Transportation's Bureau of Transportation Statistics is responsible for gathering and analyzing data related to air travel within the United States. As part of this work, the Bureau tracks the on-time performance of domestic flights operated by large air carriers. By collecting this data, the Bureau is able to gain insights into the reliability and efficiency of the nation's air travel system.

The monthly Air Travel Consumer Report published by the DOT provides a comprehensive summary of the on-time performance of domestic flights operated by large airlines. In addition to providing information on the number of on-time, delayed, canceled, and diverted flights, the report also includes data on the causes of delays and cancellations, such as weather, equipment problems, and air traffic control issues. This information can be used by airlines, airports, and policymakers to identify areas for improvement in the air travel system.

In 2015, the DOT's Bureau of Transportation Statistics published a dataset of flight delays and cancellations. This dataset contains information on over 460,000 flights that were operated by large airlines within the United States during that year. By analyzing this data, researchers and analysts can gain insights into trends and patterns related to flight delays and cancellations. For example, they may be able to identify which airlines or airports are more likely to experience delays, and what factors contribute to these delays.

In order to analyze this data effectively, it is often necessary to inspect, clean, and organize it in a way that is useful for a particular research or analysis project. This is the role of the speaker in the second paragraph - to take the flight delay and cancellation data that has been collected and published by the DOT's Bureau of Transportation Statistics, and to process it in a way that meets their specific needs. Some of the data used by the speaker may have been gathered from an API that provides real-time information about aviation flight times, which can help to ensure that the analysis is based on the most up-to-date and accurate data available.

VI. DATA ANALYSIS

Our analysis began by first taking a large-eye view of what the data was telling us. We wanted to give our airport managers, flight crews, and airline executives the right tools and information in order to make the correct decisions that would have the biggest impact on future business.

Right from the start, we were able to show that late aircraft delays and airline delays made up the bulk majority of the overall delays by almost 2 to 1. This already shows a good value in our analysis that most delays can be monitored and have significant process improvement to their workflows. The two lowest, security and weather delays, are affected by outside forces that the airlines and airports have no control over; however, we still were able to find relevant data that could lend an insight on how to minimize these issues.

The most impactful analysis we were able to find came from breaking up the airlines by delay type and major airports by delay type. This was able to give a much more clear and focused picture on the major hot-spots around the country. For example, Southwest by far led the pack with the most delays out of any airlines. Cross-referencing this data with the fact that Orlando, Atlanta, and Dallas Fort-Worth, and Denver were the top cities, which all coincidentally are major Southwest hubs, gives a very strong correlation to Southwest's delays. In fact, 6 of the top 10 airports that showed major delays all are also major Southwest hubs. This is more than enough of a cause to investigate and potentially create a domino effect that could reduce delays even further.

Finally, breaking out each of the top airline and airport delays by type gives an extremely granular view on what areas need to be focused on in order to reduce delays. For example, showing that ATL has 1.2 million minutes in delay time solely on aircraft delays (and this is just in 2015) is a major pain point that cascades across all airlines and airports. Even focusing on a mere 10% reduction in this factor could reduce delays across the board by nearly 45% and increase airline reliability.

V. CONCLUSION

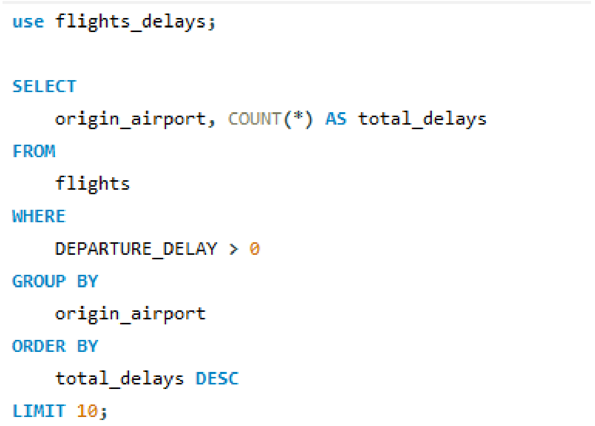
The analysis aimed to provide airport managers, flight crews, and airline executives with tools and information to make better decisions for future business. The study revealed that late aircraft delays and airline delays were the primary causes of delays, while security and weather delays were influenced by external factors. The most significant insights came from analyzing delays by airline and airport, which highlighted major hotspots around the country. For example, Southwest Airlines had the most delays, and six out of the top ten airports with significant delays were Southwest hubs. By breaking down delays by type, the analysis provided a granular view of the areas that needed attention to reduce delays, such as focusing on a 10% reduction in aircraft delays at Atlanta's airport, which could reduce delays across the board by nearly 45%.

Overall, the study showed that most delays were manageable and could be improved with process improvements. The analysis provided insights that could help airlines and airports reduce delays and increase reliability, such as identifying the root cause of delays and focusing on reducing delays by type. These findings could help stakeholders make informed decisions to improve their operations and enhance the passenger experience.

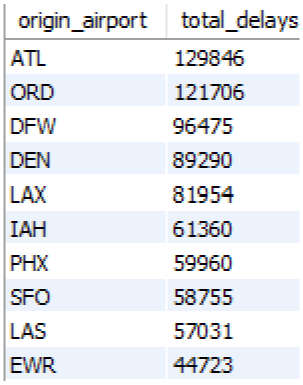
The flight delay and cancellation data collected and published by the DOT's Bureau of Transportation Statistics is a valuable resource for researchers, analysts, and policymakers who are interested in understanding and improving the efficiency and reliability of the U.S. air travel system. The U.S. Department of Transportation's Bureau of Transportation Statistics plays an important role in tracking the on-time performance of domestic flights operated by large air carriers within the United States.

By collecting and publishing data on flight delays and cancellations, the Bureau provides valuable insights into the reliability and efficiency of the nation's air travel system. While the speaker's specific role is to analyze this data based on their own needs, the data itself has a broader value in terms of informing research, analysis, and policymaking related to air travel.By continuing to track and publish data related to flight delays and cancellations, the DOT's Bureau of Transportation Statistics can help to ensure that the U.S. air travel system operates as smoothly and efficiently as possible.

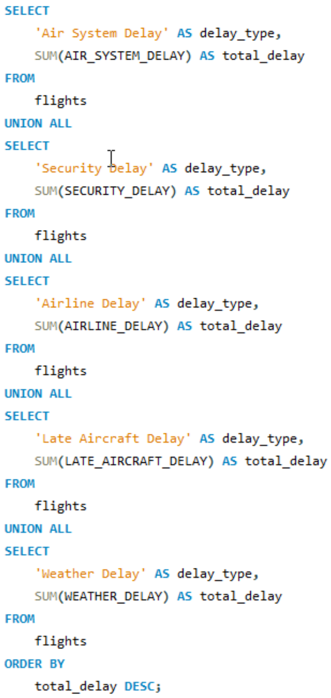
VI. APPENDIX



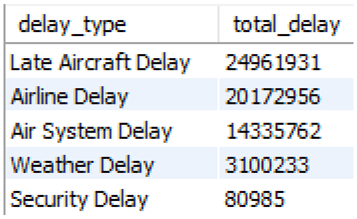
```
use flights_delays;

SELECT
    origin_airport, COUNT(*) AS total_delays
FROM
    flights
WHERE
    DEPARTURE_DELAY > 0
GROUP BY
    origin_airport
ORDER BY
    total_delays DESC
LIMIT 10;
```



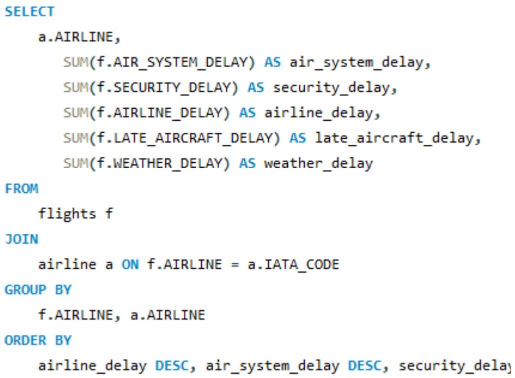
origin_airport	total_delays
ATL	129846
ORD	121706
DFW	96475
DEN	89290
LAX	81954
IAH	61360
PHX	59960
SFO	58755
LAS	57031
EWR	44723



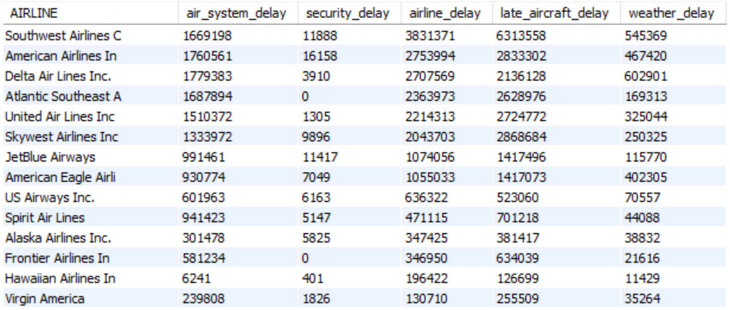
```
SELECT
    'Air System Delay' AS delay_type,
    SUM(AIR_SYSTEM_DELAY) AS total_delay
FROM
    flights
UNION ALL
SELECT
    'Security Delay' AS delay_type,
    SUM(SECURITY_DELAY) AS total_delay
FROM
    flights
UNION ALL
SELECT
    'Airline Delay' AS delay_type,
    SUM(AIRLINE_DELAY) AS total_delay
FROM
    flights
UNION ALL
SELECT
    'Late Aircraft Delay' AS delay_type,
    SUM(LATE_AIRCRAFT_DELAY) AS total_delay
FROM
    flights
UNION ALL
SELECT
    'Weather Delay' AS delay_type,
    SUM(WEATHER_DELAY) AS total_delay
FROM
    flights
ORDER BY
    total_delay DESC;
```



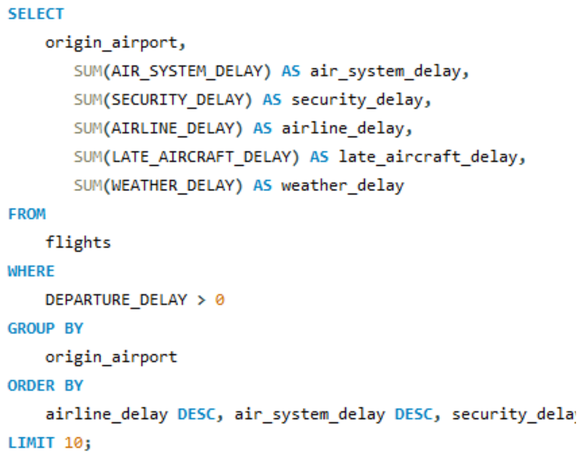
delay_type	total_delay
Late Aircraft Delay	24961931
Airline Delay	20172956
Air System Delay	14335762
Weather Delay	3100233
Security Delay	80985



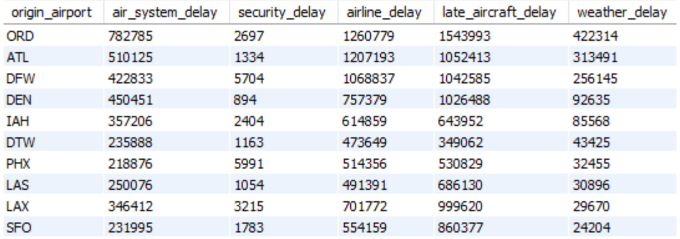
```
SELECT
    a.AIRLINE,
        SUM(f.AIR_SYSTEM_DELAY) AS air_system_delay,
        SUM(f.SECURITY_DELAY) AS security_delay,
        SUM(f.AIRLINE_DELAY) AS airline_delay,
        SUM(f.LATE_AIRCRAFT_DELAY) AS late_aircraft_delay,
        SUM(f.WEATHER_DELAY) AS weather_delay
FROM
    flights f
JOIN
    airline a ON f.AIRLINE = a.IATA_CODE
GROUP BY
    f.AIRLINE, a.AIRLINE
ORDER BY
    airline_delay DESC, air_system_delay DESC, security_delay
```



AIRLINE	air_system_delay	security_delay	airline_delay	late_aircraft_delay	weather_delay
Southwest Airlines C	1669198	11888	3831371	6313558	545369
American Airlines In	1760561	16158	2753994	2833302	467420
Delta Air Lines Inc.	1779383	3910	2707569	2136128	602901
Atlantic Southeast A	1687894	0	2363973	2628976	169313
United Air Lines Inc	1510372	1305	2214313	2724772	325044
Skywest Airlines Inc	1333972	9896	2043703	2868684	250325
JetBlue Airways	991461	11417	1074056	1417496	115770
American Eagle Airli	930774	7049	1055033	1417073	402305
US Airways Inc.	601963	6163	636322	523060	70557
Spirit Air Lines	941423	5147	471115	701218	44088
Alaska Airlines Inc.	301478	5825	347425	381417	38832
Frontier Airlines In	581234	0	346950	634039	21616
Hawaiian Airlines In	6241	401	196422	126699	11429
Virgin America	239808	1826	130710	255509	35264



```
SELECT
    origin_airport,
        SUM(AIR_SYSTEM_DELAY) AS air_system_delay,
        SUM(SECURITY_DELAY) AS security_delay,
        SUM(AIRLINE_DELAY) AS airline_delay,
        SUM(LATE_AIRCRAFT_DELAY) AS late_aircraft_delay,
        SUM(WEATHER_DELAY) AS weather_delay
FROM
    flights
WHERE
    DEPARTURE_DELAY > 0
GROUP BY
    origin_airport
ORDER BY
    airline_delay DESC, air_system_delay DESC, security_dela
LIMIT 10;
```



origin_airport	air_system_delay	security_delay	airline_delay	late_aircraft_delay	weather_delay
ORD	782785	2697	1260779	1543993	422314
ATL	510125	1334	1207193	1052413	313491
DFW	422833	5704	1068837	1042585	256145
DEN	450451	894	757379	1026488	92635
IAH	357206	2404	614859	643952	85568
DTW	235888	1163	473649	349062	43425
PHX	218876	5991	514356	530829	32455
LAS	250076	1054	491391	686130	30896
LAX	346412	3215	701772	999620	29670
SFO	231995	1783	554159	860377	24204

VII. ACKNOWLEDGEMENT

The preferred spelling of “Blue” in American English is with an “e” after the “u” (ie., not “Blu”).

The authors of this analysis paper would like to thank Professor Rafiqi of Southern Methodist University’s MSDS program for a fantastic semester.

VIII. REFERENCES

[1] Understanding the Reporting of Causes of Flight Delays and Cancellations <https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations>

[2] Flight delay forecasting and analysis of direct and indirect factors <https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/itr2.12183>

[3] Ranking different factors influencing flight delay<https://www.researchgate.net/publication/271068267_Ranking_different_factors_influencing_flight_delay>

[4] Airline On-Time Statistics and Delay Causes<https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E>