US Domestic Airline Delay Data Analysis

Data-220 DB Systems of Analytics

Raveena Kagne
MS in Data Analytics
San Jose State University
San Jose, California
017427761

Omkar Satish Aurangabadkar
MS in Data Analytics
San Jose State University
San Jose, California
017428437

Harsh Manish Somaiya
MS in Data Analytics
San Jose State University
San Jose, California
017402294

Yesheswini Lakshmi Spandana Potti
MS in Data Analytics
San Jose State University
San Jose, California
017422990

***Abstract*—With rising air travel, airspace congestion causes flight delays. The US domestic airline industry provides employment opportunities for hundreds of thousands of people every year, while at the same time generating tens of billions in revenues yearly. These issues need to be resolved through improving efficiency, which can be done through using an August 2018 dataset of airport and flight information for analysing airline systems. First of all, we created a MongoDB Atlas cluster, then conducted sharding and made connection to an unormalized dataset by means of MongoDB Compass, after that we built aggregations pipelines for queries, imported a CSV dataset through Mongoshell and runned these We developed a four table relational database schema, normalized our data, and queried the database in sql to get good information. We constructed graphs in tableau and tied this with the database. We finished by explaining the visualization process including the schema design, queries, and visualization details.**

***Index Terms*—Analysis, Data Cleaning, MongoDB, Snowflake, Visualization, Sharding, Normalization, Tableau, Python.**

I. Introduction

A. Motivation

It should be noted that the airline industry has grown immensely in the last few years with its own advantages and disadvantages. U.S. airlines transported 169.8 million additional passengers between 2022 and 2021, demonstrating an increase of approximately 30

Delayed flights may result from different causes such as air traffic control restriction, weather, inefficient operations within an airline, etc… If one delay happens to a single flight, it has a domino effect on rest of the schedule as well, which may lead to more delays and disturbances. In 2023, it is recorded that 20.83 percentage of the aircrafts are delayed for more than fifteen minutes, which is the largest figure since 2013.

The US domestic airlines industry that annually generates billions of dollars and employs hundreds of thousands experiences the expensive and inconveniencing challenges of air traffic congestion and delay. As such, addressing these issues involves having a full appreciation of the operational processes in the airline system. Analyzing information from US airports and airlines will enable us uncover operational failures which if attended to could boost efficiency and cost reduction.

B. Causes of delay

Airlines document the reasons behind the delay in general categories created by the Air Carrier On-Time Reporting Advisory Committee. The categories include Air carrier; national aviation system; weather; late-arriving aircraft, and also security. except the category of late arriving aircraft the reasons for the cancelations are the same.

- Air Carrier: This airline bears responsibility of any delay or cancellation due to issues such as maintenance, crew, baggage loading, fueling, and so on.
- Extreme Weather: Meteorological situations which might hinder flying activities, for example tornados, blizards, or hurricanes.
- National Aviation System(NAS): Part of the national aviation system comprises delays and cancellations not caused by extreme weather, heavy traffic, airport operations, or air traffic control.

C. Technology

MongoDB Atlas and Sharding: The cloud database service, mongodb atlas provides storage, data analysis, and retrievable features. Shardig is a method of partitioning information among different machines in order to increase speed as well as scalability.

Data Connection and Aggregation Pipeline: MongoDB Compass (graphical user interface for MongoDB) connected the data set. In MongoDB, processing of data and queries was implemented via the use of the Aggregation Pipeline.

Snowflake Data Warehousing: Snowflake is a cloud software that can store information related to on-time performance statistic data for an airline.

Normalization and Schema Design: The unnormalized dataset was divided into four tables: Airlines, airports, flights, and flight delays. The process was arranged such that there would be minimum repetition as well as dependencies.

SQL Queries and Data Analysis: Several SQL queries were used in order to retrieve meaningful data out of this structured database. The responses to these questions could contribute immensely to the analysis of airline and airport performance.

Tableau Visualization: Data was made available and used in creation of tableaus that were hooked to the database. They

will provide insights for understanding the trends, patterns and exceptions found within the dataset .

Python and Pandas: Data analysis and visualization is commonly done using Python and Pandas. These tools can be used in cleaning of data, data manipulation, and analytics.

II. Methodology

A. Data Source

The data used for this project was obtained from Data World which got it from the Bureau of Transportation Statistics, which is managed by the United States Department of Transportation (https://data.world/dot/airline-on-time-performance-statistics)

B. Data Cleaning

Data cleaned using python libraries like pandas and NumPy. Several preprocessing steps were conducted prior to analysis of the provided dataset. The analysis included the exclusion of six columns first. To this end, delay outliers were left as they exerted a huge influence on this parameter.ederbörd: Baguio is an area receiving very heavy rainfall. Missing values were replaced by zeros, while the time columns have been formatted consistently. An addition of the values in CRS Elapsed Time and ARR Delay New columns created a new column known as "ACTUAL ELAPSED TIME." We generated a new column CRS ARR TIME using the formula ARR TIME – ARR DELAY NEW. Finally, a day characteristic was also carefully included in the data set that could be used to improve the analysis process.

C. Architecture

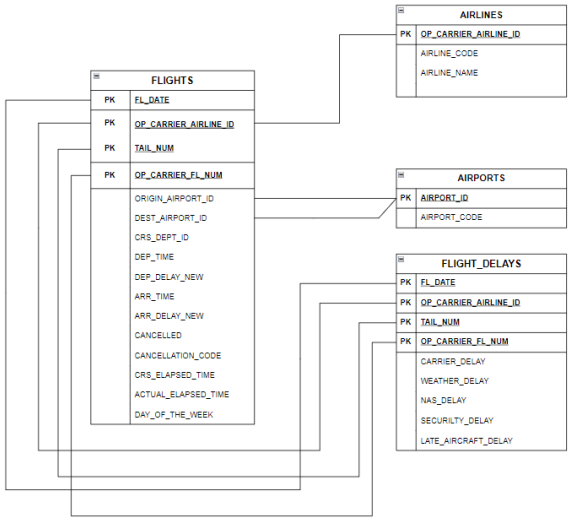
First, we designed a cluster in MongoDB Atlas and linked it to our data set. Next, we opted for sharding as a strategy to spread the burden of the denormalized dataset. Therefore, we linked with MongoDB compass through cluster and localhost after that. Write queries accordingly using the aggregation pipeline. Finally, we used the mongo import command in the mongoshell to connect our CSV dataset: mongoimport -d flights -c delay, from "cleaneddataflights.csv" as type csv with headerline. Finally, we ran the queries against the mongoshell.

The collection of on-time performance statistics data for aviation we did based on Snowflake data warehousing tools. The creation of dataset "flightinfo" entailed the inclusion of information such as flight number, delays, airlines, and airport names. We later formulated a database schema that comprised of four tables of which were Airlines, Airports, Flights and Flight delays. These are the tables that we normalized the flightinfo based upon its attributes.

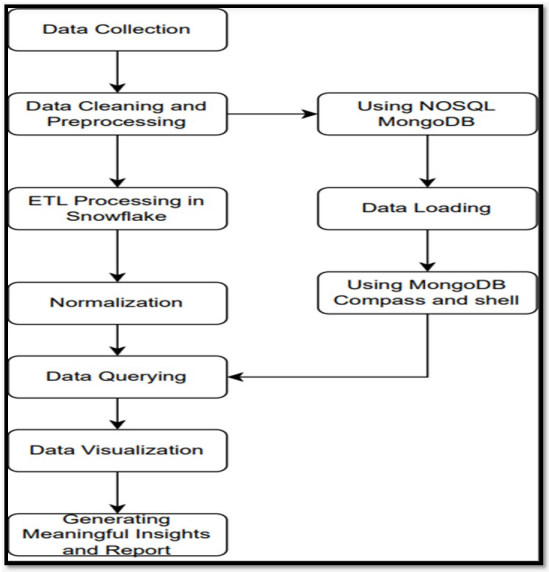
Once normalized, we created sample SQL for analysis and reporting of these data. These questions can fetch relevant details from the organized database. Then, we linked Tableau with the AIRLINEDELAY database so as to come up with the visuals using this data.

Lastly, we also noted the procedure down, that is, design schema and tables, queries used alongside their purposes and interpretations in case they are expressed visually using Tableau. Structured data base and visualizations may provide helpful information about airline performance analyses.

D. ER Diagram



E. Process Diagram



III. RELATED WORK

[1] The paper's primary focus is on anticipating flight delays in the aviation industry. It emphasizes the significance of data preprocessing, feature engineering, and the use of machine learning models to address real-world problems. This study emphasizes the power of data-driven decision-making and predictive analytics, which may be used to improve database system capabilities and optimize operations.

[2] The study pre-processes the dataset by restricting the time period to rule out the influence of Covid-19 on the irregularity of airline schedules. Analysis shows that the correlation between airline carriers, departure times, and airline delay is strong. A weak correlation between weather conditions and

airline delays and no correlation between months and plane age to an airline delay.

[3] The technique presented in this paper might be a useful reference for database system initiatives, showcasing the ability of data mining in unearthing hidden insights and helping decision-making processes. The authors discover a number of intriguing trends in the data. The authors also employ data mining techniques to identify groups of travelers who exhibit similar behavior. Its emphasis on knowing client behavior and preferences is consistent with database systems.

[4] Logistic regression to predict delay in departure times of aircraft. Microsoft Azure Learning Studio platform is used for integrating machine learning for training and testing the model on the cloud. 80 percent accuracy in predicting whether a given aircraft would be delayed or not based on the training using past data.

[5] This paper develops a machine learning approach for forecasting US domestic airline delay based on data. This model is trained on more than 10 million flights' data set and provides a forecast of over 90 per cent accuracy for flight delays.

[6] Behavior analytics of the US domestic airline passengers is evaluated by means of data mining tools. Passengers prefer direct flights, shorter time intervals of the flight, and convenient time for departure and arrival.

IV. RESULTS AND USE CASE ANALYSIS

BUSINESS CASE 1 - Find the total number of flights for each airline, showing the airline code, name, and the count of flights.

Technology Used – Snowflake data warehouse tool.

```
-- Find the total number of flights for each airline, showing the airline code, name, and the count of flights.
SELECT
    A.Airline_Code,
    A.Airline_Name,
    COUNT(F.OP_CARRIER_AIRLINE_ID) AS Total_Flights
FROM
    Airlines A
JOIN
    Flights F ON A.OP_CARRIER_AIRLINE_ID = F.OP_CARRIER_AIRLINE_ID
GROUP BY
    A.Airline_Code, A.Airline_Name;
```

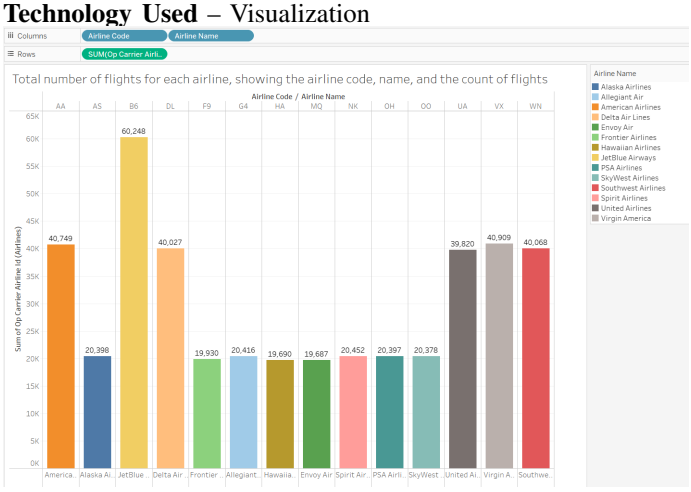
Results | Chart

	AIRLINE_CODE	AIRLINE_NAME	TOTAL_FLIGHTS
1	WN	Southwest Airlines	39829
2	UA	United Airlines	60799
3	NK	Spirit Airlines	13803
4	DL	Delta Air Lines	47307
5	VX	Virgin America	17011
6	HA	Hawaiian Airlines	3677
7	OH	PSA Airlines	11957
8	AA	American Airlines	37287

Query Details — Query duration 75ms — Rows 14 — Query ID 01b090da-0001-e7a0-...

AIRLINE_CODE — 100% filled

Technology Used – Visualization



Technology Used – Queries executed via Mongo Compass

STAGE OUTPUT — OPTIONS

Sample of 10 documents

```
count: 8526
airlineCode: "B6"
airlineName: "JetBlue Airways"

count: 3805
airlineCode: "AS"
airlineName: "Alaska Airlines"

count: 9761
airlineCode: "AA"
airlineName: "American Airlines"

count: 5074
airlineCode: "VX"
airlineName: "Virgin America"

count: 3089
airlineCode: "F9"
airlineName: "Frontier Airlines"
```

BUSINESS CASE 2 - Calculate the average departure delay for each airline, showing the airline code, name, and the average delay.

Technology Used – Snowflake data warehouse tool.

```
-- Calculate the average departure delay for each airline, showing the airline code, name, and the average delay.
SELECT
    A.Airline_Code,
    A.Airline_Name,
    AVG(F.DEP_DELAY_NEW) AS Average_Departure_Delay
FROM
    Airlines A
JOIN
    Flights F ON A.OP_CARRIER_AIRLINE_ID = F.OP_CARRIER_AIRLINE_ID
GROUP BY
    A.Airline_Code, A.Airline_Name;
```

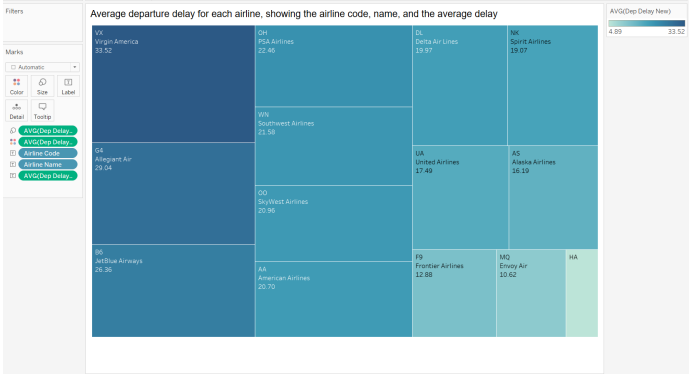
Results | Chart

	AIRLINE_CODE	AIRLINE_NAME	AVERAGE_DEPARTURE_DELAY
1	WN	Southwest Airlines	21.582440
2	UA	United Airlines	17.492179
3	NK	Spirit Airlines	19.065421
4	DL	Delta Air Lines	19.971590
5	VX	Virgin America	33.516901
6	HA	Hawaiian Airlines	4.894207
7	OH	PSA Airlines	22.463076
8	AA	American Airlines	20.702175

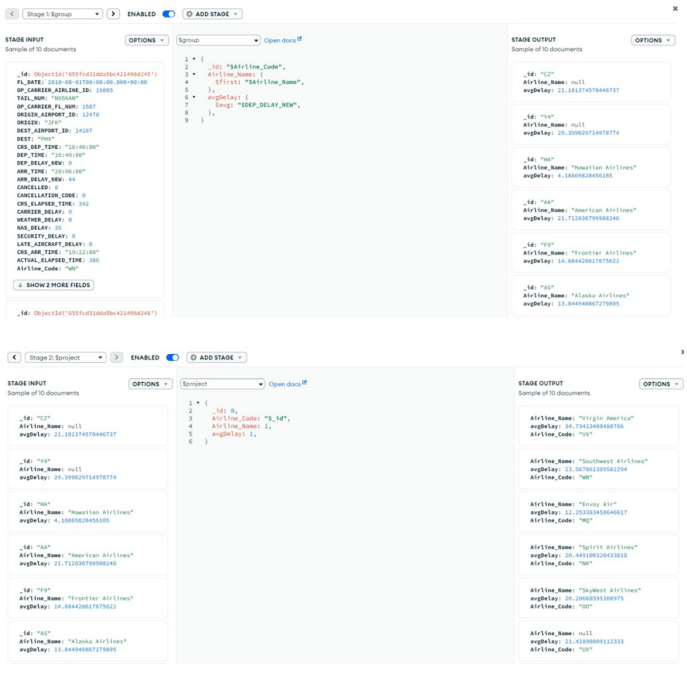
Query Details — Query duration 71ms — Rows 14 — Query ID 01b090e8-0001-e7ed-...

AIRLINE_CODE — 100% filled

Technology Used – Visualization



Technology Used – Queries executed via Mongo Compass



BUSINESS CASE 3 - Show the count of canceled flights and the cancellation reasons.

Technology Used – Snowflake data warehouse tool.

```
-- Show the count of canceled flights and the cancellation reasons.
SELECT
    CASE
        WHEN cancellation_code = 'A' THEN 'Maintenance issues'
        WHEN cancellation_code = 'B'THEN 'Weather Issues'
        WHEN cancellation_code = 'C' THEN 'National Transporation Issue'
        ELSE 'Severe Delay'
    END AS CancelLation_Reason,
    COUNT(*) AS Canceled_Flights
FROM
    Flights
WHERE
    CANCELLED = 1
GROUP BY
    CANCELLATION_CODE;
```

Results | Chart

	CANCELLATION_REASON	CANCELED_FLIGHTS
1	Maintenance issues	2257
2	Weather Issues	2311
3	National Transporation Issue	4069

Query Details — Query duration 50ms — Rows 3 — Query ID 01b090e8-0001-e803-...

CANCELLATION_REASON — 100% filled

Technology Used – Visualization

Technology Used – Queries executed via Mongo Compass

BUSINESS CASE 4 - List the top 5 busiest airports based on the total number of departures and arrivals.

Technology Used – Snowflake data warehouse tool.

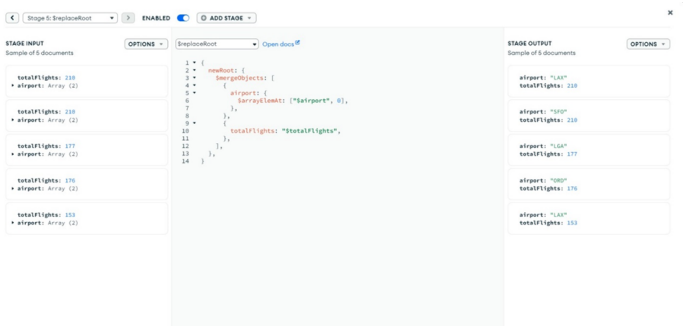
AIRLINEDELAY.PUBLIC | Settings

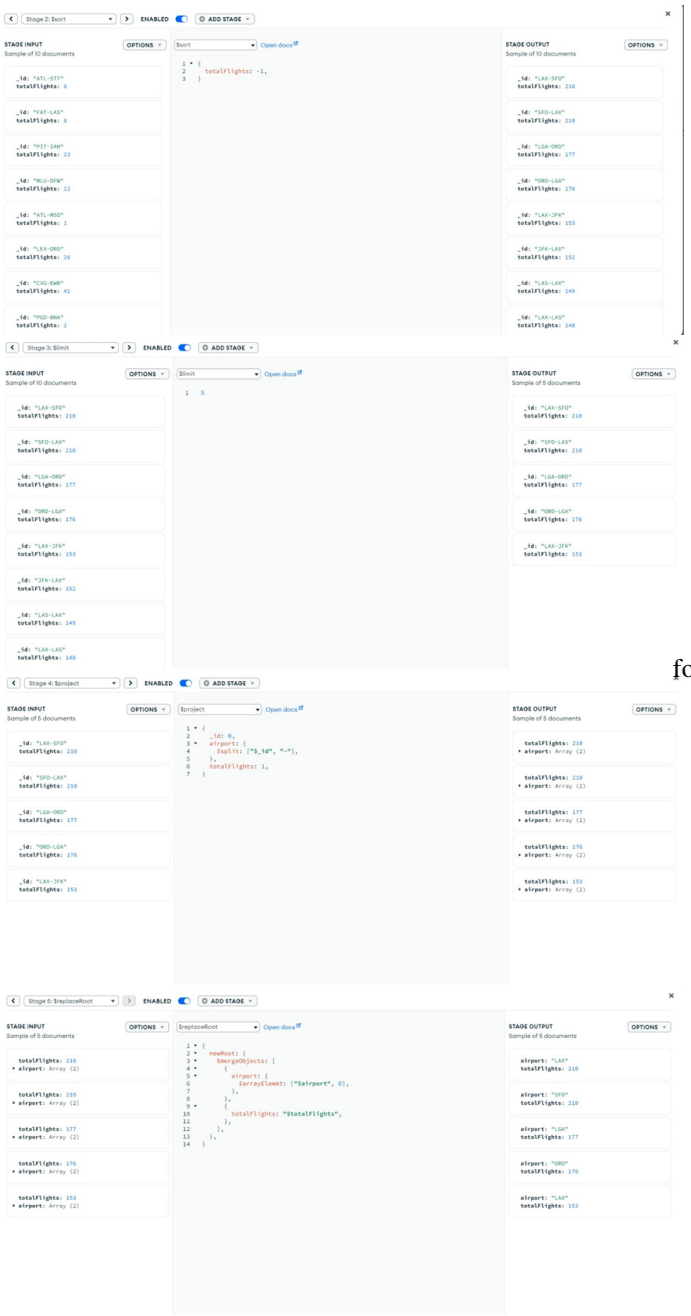
```
SELECT
    Airport_ID,
    Airport_Code,
    SUM(CASE WHEN Origin_Airport_ID = Airport_ID THEN 1 ELSE 0 END) AS Departures,
    SUM(CASE WHEN Dest_Airport_ID = Airport_ID THEN 1 ELSE 0 END) AS Arrivals,
    SUM(1) AS Total_Flights
FROM
    Airports A
JOIN
    Flights F ON A.Airport_ID IN (F.Origin_Airport_ID, F.Dest_Airport_ID)
GROUP BY
    Airport_ID, Airport_Code
ORDER BY
    Total_Flights DESC
LIMIT 5;
```

Results | Chart

	AIRPORT_ID	AIRPORT_CODE	DEPARTURES	ARRIVALS	TOTAL_FLIGHTS
1	13930	ORD	16751	16723	33474
2	10397	ATL	15347	15354	30701
3	11298	DFW	11777	11779	23556
4	11292	DEN	10937	10936	21873
5	11057	CLT	10080	10082	20162

Technology Used – Queries executed via Mongo Compass





BUSINESS CASE 5 - Show the distribution of departure delays with categories (e.g., early, on time, minor delay, moderate delay, severe delay).

Technology Used – Snowflake data warehouse tool.

```
-- Show the distribution of departure delays with categories (e.g., early, on time, minor delay, moderate delay, severe delay).

SELECT
    CASE
        WHEN DEP_DELAY_NEW <= 0 THEN 'On Time'
        WHEN DEP_DELAY_NEW <= 15 THEN 'Minor Delay'
        WHEN DEP_DELAY_NEW <= 60 THEN 'Moderate Delay'
        ELSE 'Severe Delay'
    END AS Delay_Category,
    COUNT(*) AS Flights_Count
FROM
    Flights
GROUP BY
    Delay_Category
ORDER BY
    Delay_Category;
```

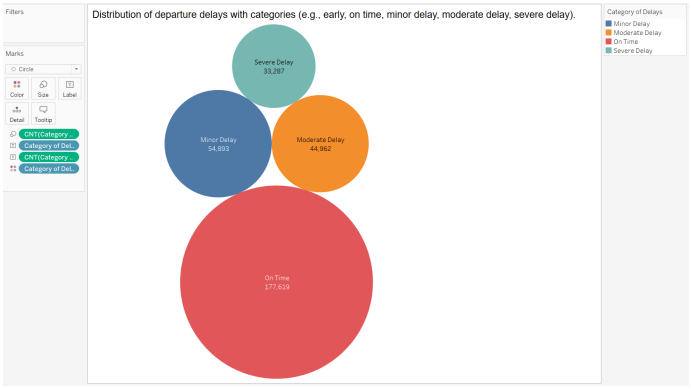
Results | Chart

	DELAY_CATEGORY	FLIGHTS_COUNT
1	Minor Delay	54893
2	Moderate Delay	44962
3	On Time	177619
4	Severe Delay	33287

Query Details
Query duration 22ms
Rows 4
Query ID 01b090f6-0001-e7ec-0...

DELAY_CATEGORY
100% filled

Technology Used – Visualization



BUSINESS CASE 6 - Calculate the average departure delay for each day of the week.

Technology Used – Snowflake data warehouse tool.

```
-- Calculate the average departure delay for each day of the week.

    SELECT
        DAY_OF_WEEK,
        AVG(DEP_DELAY_NEW) AS Average_Departure_Delay
FROM
        Flights
GROUP BY
        DAY_OF_WEEK
ORDER BY
        DAY_OF_WEEK;
```

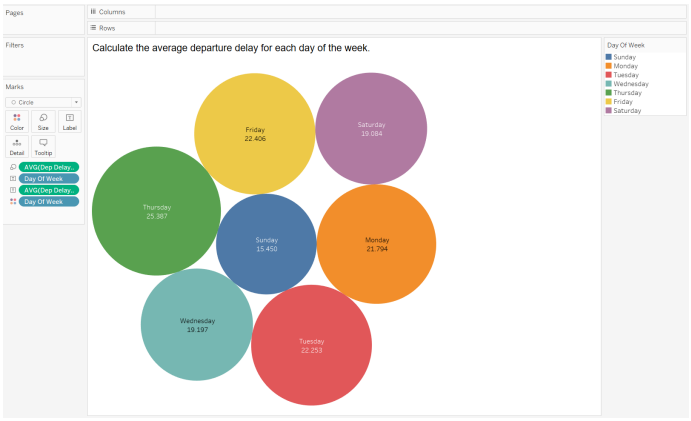
Results | Chart

	DAY_OF_WEEK	AVERAGE_DEPARTURE_DELAY
1	Friday	22.406151
2	Monday	21.793609
3	Saturday	19.084213
4	Sunday	15.449548
5	Thursday	25.386692
6	Tuesday	22.253007
7	Wednesday	19.197428

Query Details
Query duration 44ms
Rows 7
Query ID 01b090f7-0001-e830-0...

DAY_OF_WEEK
100% filled

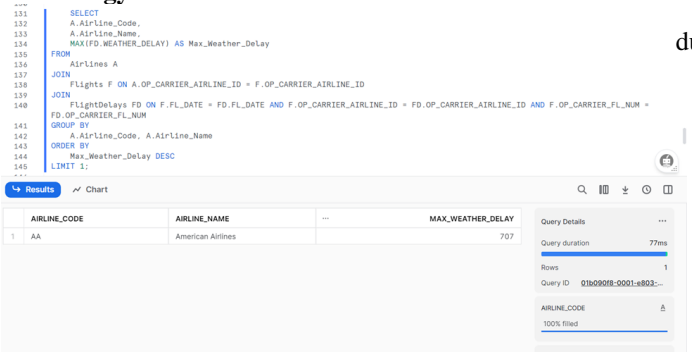
Technology Used – Visualization



Technology Used – Queries executed via Mongo Compass

BUSINESS CASE 7 - Find the airline with the maximum weather delay, showing the airline code and name.

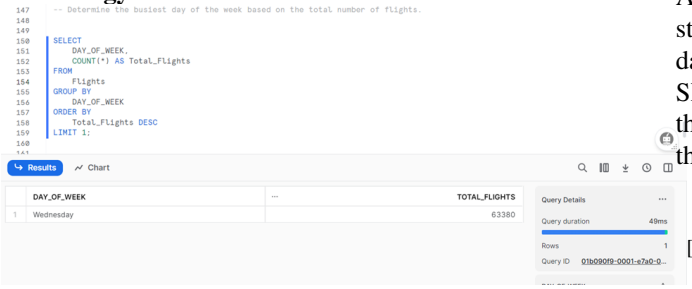
Technology Used – Snowflake data warehouse tool.



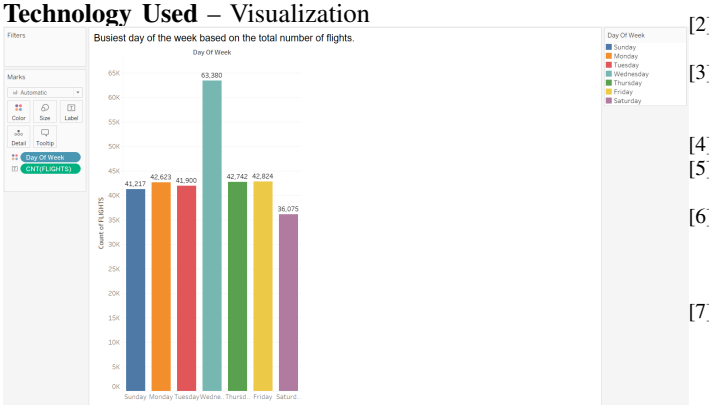
Technology Used – Queries executed via Mongo Compass

BUSINESS CASE 8 - Determine the busiest day of the week based on the total number of flights.

Technology Used – Snowflake data warehouse tool.



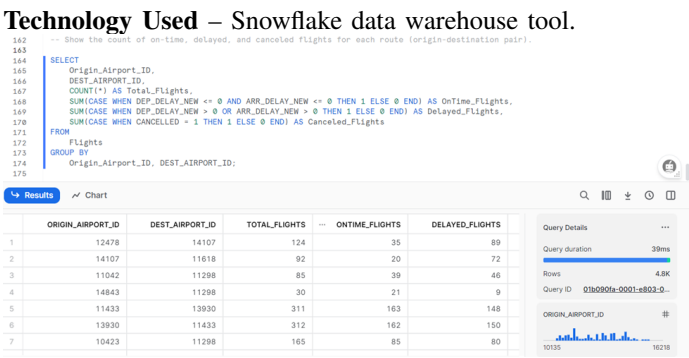
Technology Used – Visualization



Technology Used – Queries executed via Mongo Shell

BUSINESS CASE 9 - Show the count of on-time, delayed, and canceled flights for each route (origin-destination pair).

Technology Used – Snowflake data warehouse tool.



Technology Used – Queries executed via Mongo Compass

BUSINESS CASE 10 - Identify the peak departure times during the day.

Technology Used – Queries executed via Mongo Shell

V. Conclusion

Using data for August 2018, we studied airline's performance. The carrier with most weather-related delays was American Aires. With an average delay of 34 mins, Virgin America was the leader. The majority of cancelled transactions stemmed around issues related to National security. Wednesday was the busiest day, and the top 5 routes were: JFK, LAX, SFO, LAX, ORD, LGA. On averages, Thursday experienced the hightest delay. The result of our research is informative to the operation of airlines, and reduces delays.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

VI. Appendices

Check Team Project Rubrics.pdf for links.