US Domestic Flight Analysis

Can we predict flight delays?

Jaskaran Garha
214314439
garhajas
garhajas@my.yorku.ca

David Geller
214404255
dgeller
dgeller@my.yorku.ca

Gilbert Andze Tsoungui
211262656
christ18
chris18@my.yorku.ca

William Zhen
214305171
will636
will636@my.yorku.ca

ABSTRACT

Every few seconds a flight will take off or land from an airport in the United States. With the increasing amount of flights taking place, it is more likely for there to be delays amongst these flights. Using flight data, we can potentially predict these delays using Apache Spark and distributed computing. This information can be passed on to inform the consumer and help companies improve services.

1 Domain Description and Motivate

1.1 *What is the data domain?*

The data domain for our project is air transportation. We will be examining US domestic flight data from recent years provided by the Bureau of Transportation Statistics (BTS) [1]. We will be using the csv data files from BTS as our primary source of analysis. Some of the key information the dataset includes are destination and origin of flight, airport names, elapsed, arrival and departure times, and lastly the causes of delay or cancellation of the flight along with other miscellaneous information.

1.2 *What is the goal of your project?*

The goal of this project is to find a way to analyze several years of recent flight data in order to come up with a system of determining which airline(s) to recommend to customers.

1.3 *What is the motivation for rigorous data analytics?*

The motivation for rigorous data analytics is the demand for customers of airlines to reach their destinations in a timely matter. By extracting this data, it can be used to provide insight on the delay patterns of airlines. This analysis will help us to make predictions about these delays that can be passed on as knowledge to the consumer so that they can make a well-informed decision when they are choosing which airlines to purchase tickets from.

1.4 *What are the questions you want to answer?*

- What is the most common cause of flight delays and cancellations?
- What airline companies have the most delays and cancellations over time?
- Which geographical regions are most common to cause/have delays at their airport?
- What times of the year has a high probability of delays/cancellations?
- Has there been an improvement or decline in the quality of service of an airline over a period?
- What is the turnaround time of popular airports vs unpopular airports when dealing with delays?

1.5 *Why is the analysis important?*

The results of the data from this analysis can be used to motivate airlines to improve their services and provide valuable insight to customers when choosing flights. Additionally, customers will be informed of common delays and cancelations based of airlines on our results.

1.6 *What are a few potential applications?*

- The processed data set can be used to help travelers know which airline carriers and routes are usually on time.
- Planners for airline routes can use historic data to understand what routes and airports are most likely delayed due to unforeseen circumstances and use that knowledge to better schedule flight times (very important for busy airports).
- The findings can be sent to the worst airlines in order to help them improve their services, which will result in improved customer satisfaction and more options of flights for the consumer.

2 Architecture of Proposed Solution

2.1 *Architecture for Data Analytics*

Data Ingestion:

- CSV files that were obtained from the BTS database [1]
- Remove any unnecessary information from the CSV files and clean the data

Storage:

- For optimal performance, Apache Parquet will be used for columnar distributed storage of data

Processing & Analysis:

- Using Spark API and tools, such as dataframes and MLlib, to find patterns and make predictions

Representation:

- Using graphs, plot charts and maps to display information on delays and locations of delays

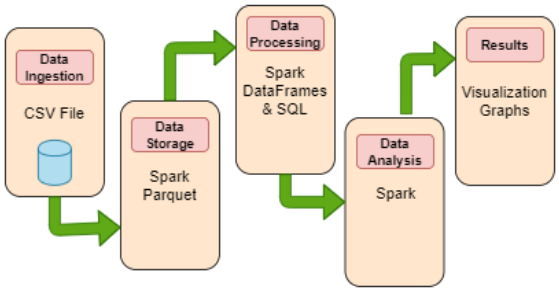

Figure 1 Depiction of the architecture and data flow in the system.

2.2 *Limitations and Difficulties to this Approach*

- Overtime the code and/or name of the airline carrier can change and may either no longer exist or be used by another airline carrier, making aggregating data from multiple years a potential challenge
- Mergers of airline carriers, where the name of each airline doesn't change, over the years make it a potential challenge when collecting data and figuring out how to report the data findings (i.e. do we report each of the airlines or do we report them together)

3 System Evaluation and Data Analysis

3.1 *How will you evaluate your system and architecture?*

- Evaluating whether the results are accurate through manual calculation of part of the data set.
- Testing whether the system can handle increasingly larger data sets to test the limits of the system.

3.2 *What results do you plan to obtain?*

- Graph/plot charts of the delays across airlines for a specific time period, location or flight route.
- A map showing at which airports delays are most likely to occur.

3.3 *What type of data analysis will you perform?*

- Predictive analysis on weather data and flight route with respect to delays.
- Statistical analysis on average/total delays with respect to aircraft, airports, carrier, weather, time, date.

3.4 *How is this type of analysis adequate for the data, problems and questions posed?*

- The data captures the quantitative value of how delayed the flights are, wind speeds, throughput of airports which enables Statistical Analysis to look for patterns, averages and categorizations after normalization across a large dataset.
- Predictive Analysis, along with machine learning, can be adequate for answering questions like "Will this flight be delayed" for streamed data. The Analysis can be trained on the large dataset of flight records and weather records.

3.5 *What other datasets can be used?*

- Specific weather data: Wind patterns, Rain, Fog, Temperature
- Finance and operations of airline carriers
- Airport/airspace rules and regulations

3.6 *What are the steps you need to take to scale your solution?*

- Hadoop can be used to distribute the data storage so we can efficiently query and access large datasets.
- Reduce the search space by only querying/running analysis on a fixed sized Frame at a time.
- Split the data based on things like regions, carrier, etc. and merging the result after the analysis of each is done so we can run everything in parallel.
- Increasing the computational power and storage capabilities by adding new hardware.

REFERENCES

[1] *Bureau of Transportation Statistics*, Sep-2019. [Online]. Available: https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline On-Time Performance Data&DB_Short_Name=On-Time.