Airline On-Time Streaming Data Analysis

Arwa EL-Hawwat

Rahul Dev Ellezhuthil

Jaini Patel

Supervised by Professor James Abello Mondero, Harshini Bonam, & Haoyang Zhang
for CS543 Group 6, Fall 2021

ABSTRACT

- Airline consumers today are overwhelmed by the sheer number of carriers and flight patterns offered to them when choosing flights.

- While they can easily compare costs and layovers at booking time, it can be much more difficult to predict delays on the day of departure.

- Utilizing past airline data, we can extract a model to formulate possible delays and when they are most likely to occur. In addition, we are able to determine the likelihood of certain carriers or trip patterns to experience difficulties.

OVERVIEW OF PROJECT

- Database querying in a streaming fashion.

- Data is sourced from the 2009 Data Expo - Airline On-Time Performance **[1]**
  - Harvard Dataverse modified version **[2]**
  - Publicly available data on domestic United States flights provided by the American Statistical Association (AMSTAT)

- Uses Pyspark as the querying background for data.

- Uses Flask web application hosted on a Rutgers iLab machine for user interface and data visualization.

PROJECT OBJECTIVES

To query the data in a streaming fashion to be able to answer the following questions:

1) **Which airline is the most reliable in terms of flight punctuality?**

2) **What were the worst months to fly historically?**

3) **What are the busiest airports in the United States?**

4) **What are the busiest flight paths in the United States?**

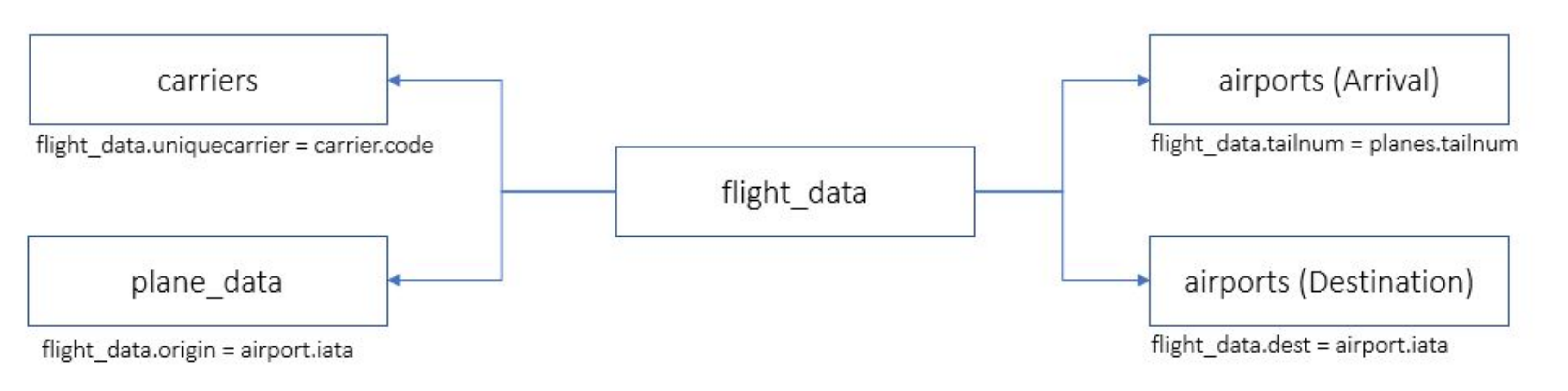
OVERVIEW OF DATASET

- Approximately 120 million records from October 1987 - April 2008. Data at rest.

- 30 columns:
    - Year, Month, Day
    - Airline Carrier Information
    - Airport Information
    - Timings and Delay information

- 4 additional files with detailed information about airplanes, carrier, airports and metadata for the dataset.

OVERVIEW OF DATASET

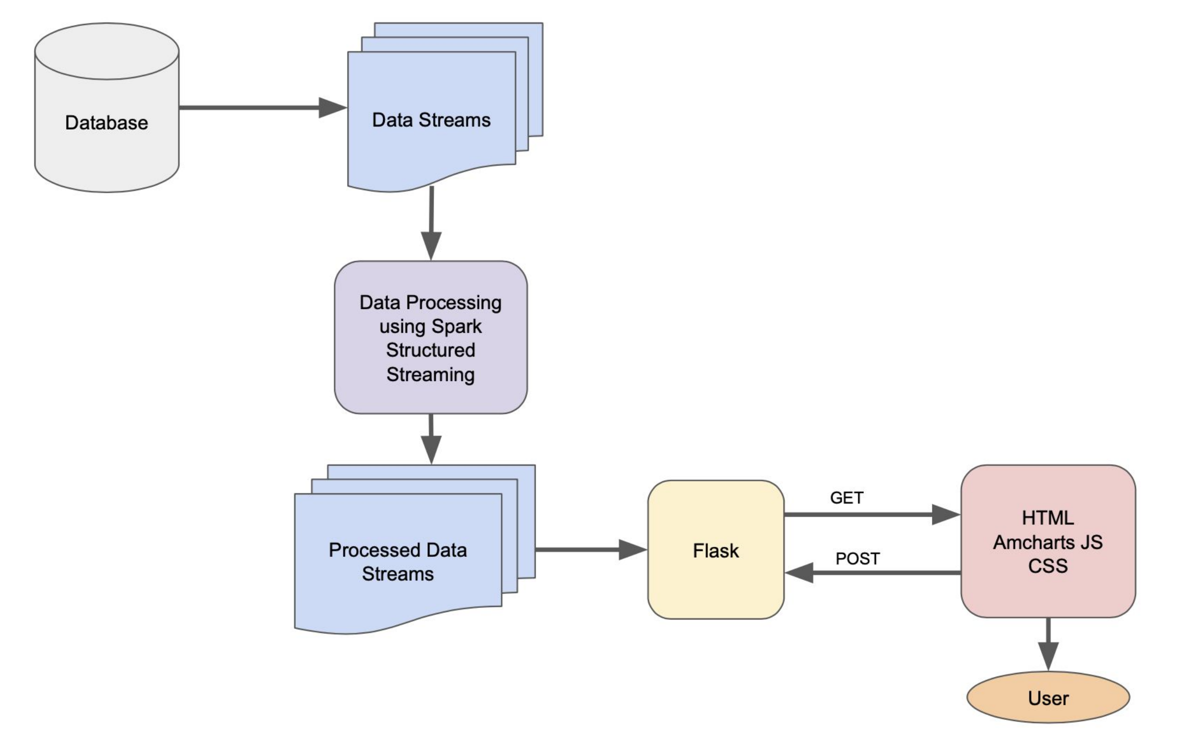

carriers
flight_data.uniquecarrier = carrier.code
plane_data
flight_data.origin = airport.iata
flight_data
airports (Arrival)
flight_data.tailnum = planes.tailnum
airports (Destination)
flight_data.dest = airport.iata

IMPLEMENTATION

- Processing Raw Data
  - To simulate streaming processing, we decided to divide the original data by year using PySpark
  - Stored in a local Hive file system in Rutgers iLab machine
  - Streaming and processing operations were written as functions and queries in Python

- Visualization
  - Web application using Python Flask framework
  - Connected to Apache Spark backend via Flask where data is queried then passed as JSON
  - Graph visualizations then computed through Python libraries like amCharts and Matplotlib

Data Flow Diagram



TOOLS

- Spark Structured Streaming
- Flask Web Application

USERS

- Airline staff, route planners, and pilots
- US domestic travellers

INTERACTIVITY

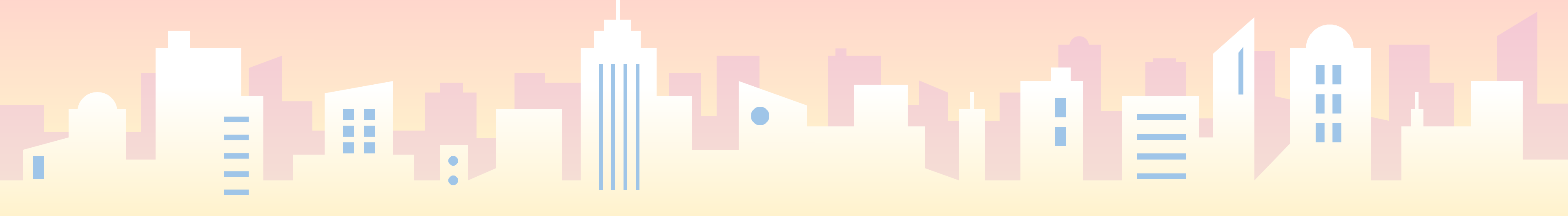
- Web Application
- Drop Down Menu
- Start Button
- Interactive Plots:
    - ○ Label selection
    - ○ Tooltip

PROJECT HIGHLIGHTS

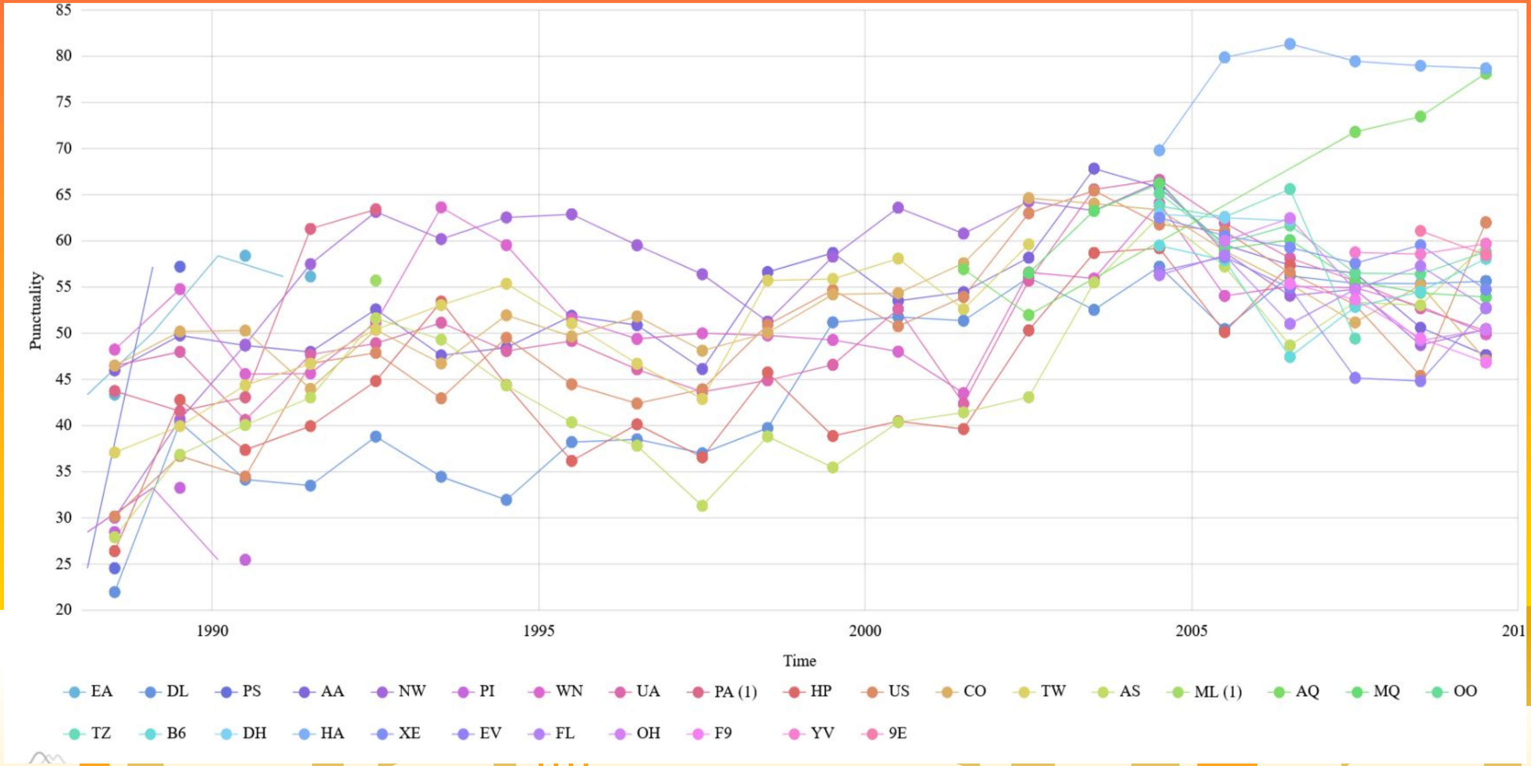
- The total number of flight records increased steadily from 1987 to 2008, which is understandable as travel increased as it became cheaper and more accessible.

- In studying the percentage punctuality of different airline carriers, we noticed various fluctuations throughout the years for a majority of them ranging from around 10-20%.

- Certain months, like during the holiday season or after global events, experienced an influx of delays and cancellations in reaction. These are most notably observed in the last quarter of the year.

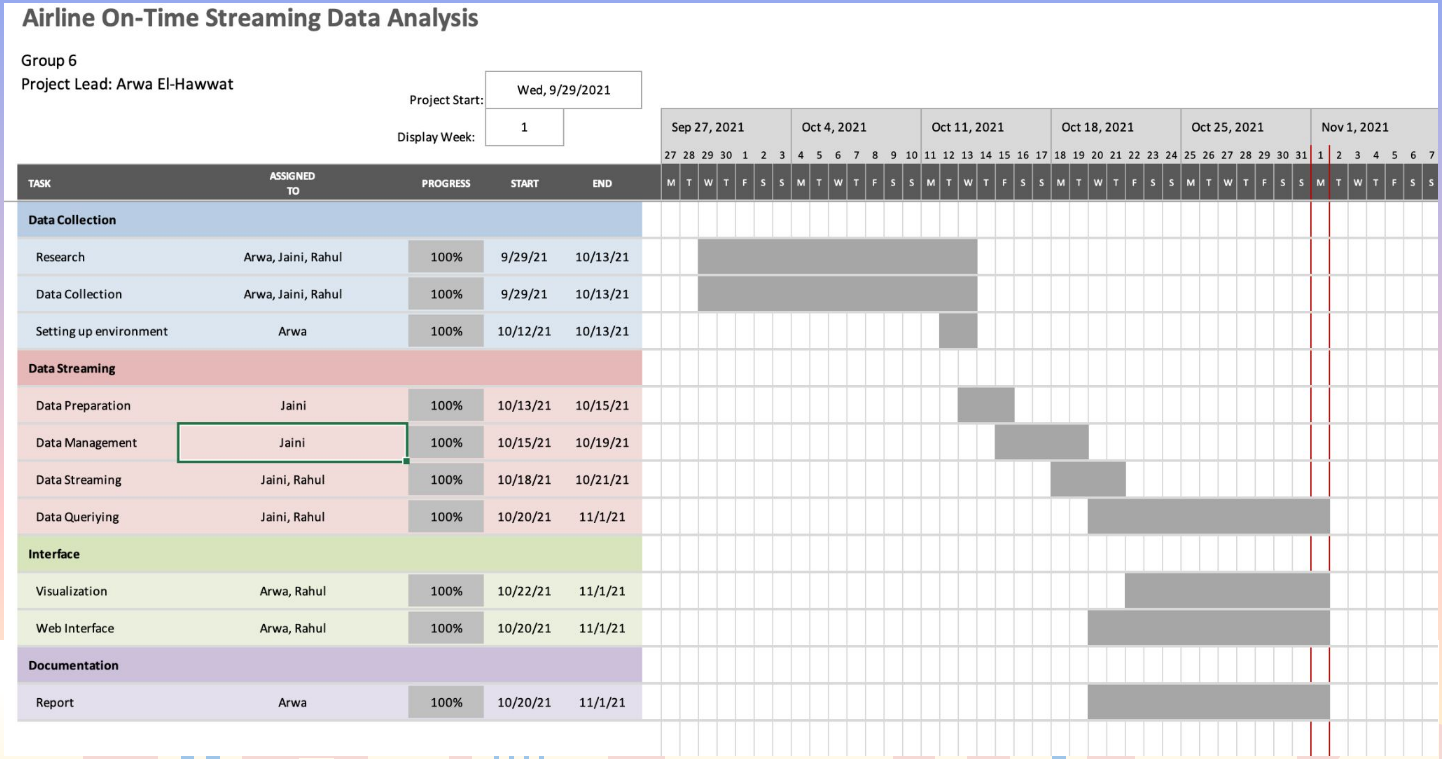


INSIGHTS



*Visualization of different US airline carriers and their calculated punctuality as a percentage from 1987 - 2008 using amChart

PROJECT TIMELINE



Airline On-Time Streaming Data Analysis

Group 6
Project Lead: Arwa El-Hawwat

Project Start: Wed, 9/29/2021

Display Week: 1

TASK	ASSIGNED TO	PROGRESS	START	END
Data Collection				
Research	Arwa, Jaini, Rahul	100%	9/29/21	10/13/21
Data Collection	Arwa, Jaini, Rahul	100%	9/29/21	10/13/21
Setting up environment	Arwa	100%	10/12/21	10/13/21
Data Streaming				
Data Preparation	Jaini	100%	10/13/21	10/15/21
Data Management	Jaini	100%	10/15/21	10/19/21
Data Streaming	Jaini, Rahul	100%	10/18/21	10/21/21
Data Queriying	Jaini, Rahul	100%	10/20/21	11/1/21
Interface				
Visualization	Arwa, Rahul	100%	10/22/21	11/1/21
Web Interface	Arwa, Rahul	100%	10/20/21	11/1/21
Documentation				
Report	Arwa	100%	10/20/21	11/1/21

REFERENCES

[1] 2008, Data Expo 2009 - Airline on-Time Performance, American Statistical Association, https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009

[2] 2008, Data Expo 2009: Airline on time data, Harvard Dataverse, V1, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

[3] Python Flask, https://flask.palletsprojects.com/en/2.0.x/

[4] amCharts JS library, https://www.amcharts.com/

THANK YOU!!