Status Report - Skyletics

Our Real-World Problem:
- We have SFO airport dataset in csv formats.
- How the dataset is created - Data is self-reported by airlines and is only available at a monthly level.
- Airport data is seasonal in nature therefore, any comparative analyses will be done on a period-to-period basis.
- With limited runway and gate capacity, SFO struggles to accommodate many flights arriving and departing within a short timeframe.
- High passenger volumes can lead to congestion in terminal buildings, security checkpoints, and baggage claim areas.
- Effective communication and coordination between the airport and airlines are crucial to managing passenger flows and minimizing delays.

Dataset Summary:

We have 2 datasets Air_Traffic_Passenger_Statistics.csv and Air_Traffic_Landings_Statistics.csv. Below is the description for both the files.

Air_Traffic_Landings_Statistics.csv -

The dataset provides details on aircraft landings. It lists three records associated with two airlines: ABX Air and ATA Airlines, all operating domestic flights within the US. ABX Air has two records; both indicate freighter flights using narrow-bodied McDonnell Douglas DC-9 aircrafts, with a combined 41 landings and a total landed weight of 4,168,000. One has 40 landings with a weight of 4,066,000 (version 30), and the other has 1 landing weighing 102,000 (version 41). On the other hand, ATA Airlines operates passenger flights with narrow-bodied Boeing 757 aircraft, recording 2 landings with a total weight of 396,000.

Air_Traffic_Passenger_Statistics.csv –

The dataset highlights passenger activity for ATA Airlines. Operating under the IATA code "TZ" and servicing domestic routes within the US, the airline recorded two distinct activities at Terminal 1, Boarding Area B. Specifically, 27,271 passengers deplaned and 29,131 passengers enplaned, all categorized under the "Low Fare" price category. In essence, the data provides insights into the volume of passengers both arriving and departing via ATA Airlines in the specified month, with all transactions being classified as low-cost.

Data Lifecycle:

- Data will be fetched from the above-mentioned Flat files. This will have data regarding Operating Airline , Published Airline, GEO Summary, GEO Region, Terminal Boarding Area, Passenger Count etc.
- We will perform ETL on the dataset where we will filter the data, apply aggregation functions and including tasks like handling missing values and data format standardization.
- We have stored the datasets on HADOOP server in a structured format and will use HIVE for performing analysis on the data to generate visualizations and reports for our stakeholders.
- Data will be processed using Machine Learning algorithms and Human review too.
- We will be applying prediction models for flight analysis, delays, flight path changes. Also, we will be providing data which will help clients understand the geological aspect of flight patterns such as which country/continent has high flight arrivals/departures.
- We will be adding a small segment to understand the passenger count for airlines to understand the capacity of airlines.
- We are planning to provide visualizations on a Dashboard so the client can directly check flight delay predictions and profit from this dashboard's insights and real-time access to metrics for airport efficiency, flight data, and safety information.
- We will also provide an option to provide these visualizations in a downloadable format so clients can have different reports in the manner of pdf versions.

High Level Diagram
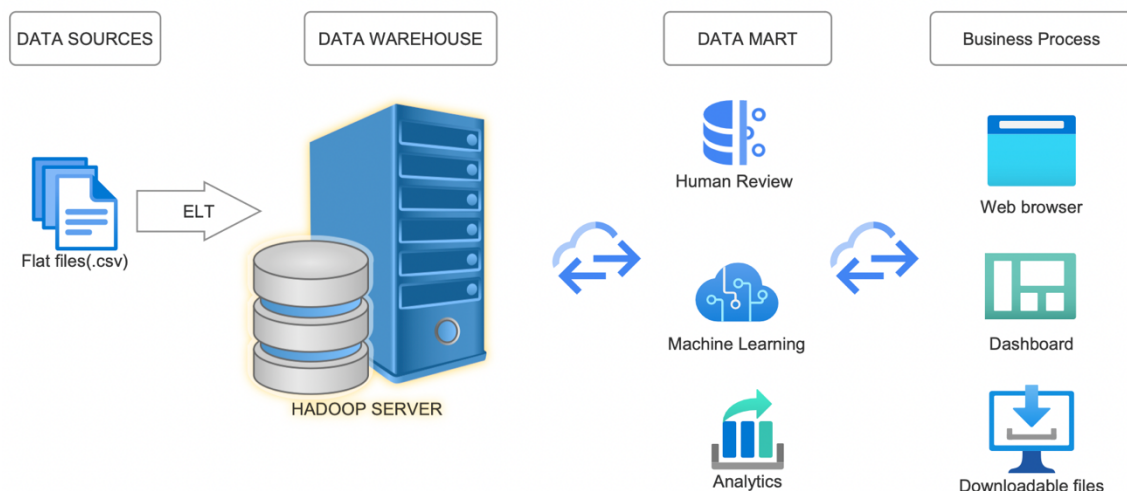
Shortcomings:
- Since it's a government generated dataset, we are figuring out a way to have the recently updated flat file at our end, so the prediction models function efficiently.
- We have tried to generate a flight analysis file which will help us in prediction of key metrics such as flight delays, but the size of the file is around 5GB which is creating an issue while storing it on the Hadoop Server.
- We also want to implement an Online Web Portal which can be accessible for the users along with the dashboard but are yet to finalise on how to use our data.