PROJECT PROPOSAL

MODULE 3: DATA ENGINEERING

TASK: TO BUILD AN END-TO-END DATA PIPELINE

Project Back-end:

My project seeks to feed a large raw, static dataset into a data pipeline that should clean the data, perform EDA and then feed the clean dataset to a web app for users.

The Data:
The project will use a delayed flight dataset of not less than 3m records.
The purpose of the project is to analyze and visualize the flight dataset to illustrate:

1. The major factors causing flight departure delays. Are they within the carriers' control or not ?
2. Most airports in the US are quite large, with multiple run-ways. This is a good thing. But aircraft then have to travel a considerable distance before accessing the runway to become airborne and actually start on their route - or start de-planing at the end of it. So are taxi-durations therefore an additional cause of delays for instance?
3. What are the major contributors to flight arrival delay?
4. Planes only make money once airborne. Extended ground time therefore, prior to departure does pose an overhead on the carrier - which eats into the profit margin of a given route. How can we illustrate this overhead?

While there is lots of live flight data along with APIs, live flight data is quite restricted and its access is closely monitored for safety and security. As such, none of the APIs are free (reliable search).

Project Front End:
On the front end, a user should be able to interact with the dataset by extracting pre-existig statistics AS well as feeding direct inputs into the dataset and extractign an outcome. For instance:

- Which airport had the worst arrival delay times last December?
- Which airline had the worst departure delay record last summer?
- Based on existing data, how much delay is a passenger travelling from Boston to LAX likely to encounter in winter? In summer?

Tools to use:
Psql, PySpark, Pandas, scikit learn and Streamlit.

Dennis Ssekamaanya
Metis Engineering - May17th to May28th 2021.