The Air Traffic and the Air Pollution

Georgios Christodoulou

Background and Motivation:

Has been stated that the unforeseen outbreak due to COVID-19 in the 21st century, and the forced confinement of millions of inhabitants over the world, mainly during the first part of 2020, has shown a reduction in the temperature in the bigger capitals worldwide.

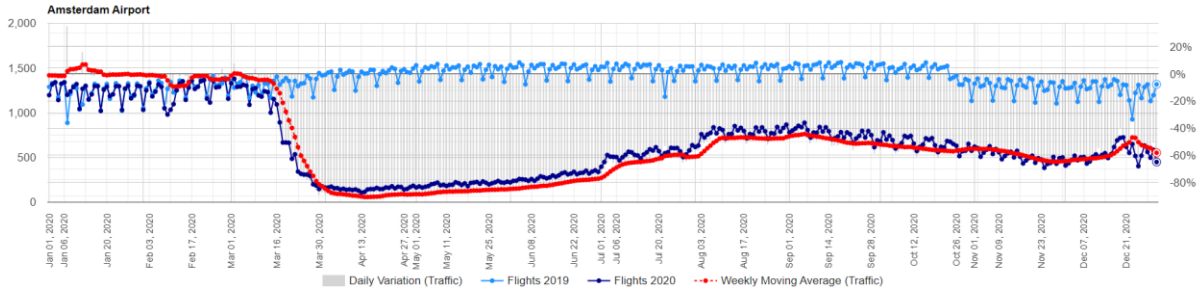
"Human activity is the cause of the climate problem."
António Guterres, COP27, Egypt, Nov. 2022

The idea of this project is to investigate how air traffic contributes to the increase the air pollution and how this may also affect global warming. We will try to utilize historical data for correlation and analysis, and real-time data with the aim to predict some measurements of the air pollutants that show the air quality such as the Ozone (O3), Carbon monoxide (CO), Sulfur dioxide (SO2), etc.

Project Objectives (Hypothesis):

The objective of this project is to investigate the correlation between airline traffic and air pollutants per area and period and identify how the increase in emissions degrades air quality.

The below graph from Euro Control data shows the traffic reduction during the second quarter of 2020 due to the pandemic measures in Amsterdam, the same decrease happened in all international airports during this period. Based on this we expect to see a similar decrease in air pollutants in the areas or cities of these airports.



Data Sources:

There is plenty of open data sources that we try to extract and include in our study such as:

- Airport, airline and route data from Eurocontrol.int, OpenFlights.org and AirLabs.co
- Weather and Environmental data from the government data sets such as data.gov.cy, www.ncei.noaa.gov, www.eea.europa.eu

Big Data Dimensions:

This project has to cope with at least three dimensions of Big Data problems. The first one is the variety of datasets, as we have to collect data from several sources and we need to extract, clean, and transform the data in a manner to be able to be used for our analysis.

Secondly, historic data and especially airline traffic are massive volumes of data that we need to process in batches, aggregate, and store for correlation and model training.

The third dimension that we have to deal with is the velocity of data as the scenario is to collect also real-time/streaming data, and use them for predictions and real-time visualization.

Solution Overview:

The difficult part of this project is to bring the data from all multiple sources, clean and uniform into a common database that will provide easy access for manipulation, joining, and analyzing them.

This may be required to build some MapReduce for processing, data, or/and use algorithms like Count-Min Sketch to keep track number of flights per airport and per hour, or the last value of an air pollutant for each station in the last hour.

Tools/libraries:

As we have observed up to now, most of the data are in CSV and JSON format, which needs to use some SQL such as Google BigQuery and NoSQL document databases such as Google Firestore.

However, the idea is to use also data streaming in order to collect real-time data from air quality sensors and air traffic and provide real-time view and analytics. For this purpose, Apache Spark and pySpark libraries looks basic tool for this.