**T1 - What tools would you use for data engineering here in general?**

NOT DONE

**T2 - Which tools would you apply for a, b and c individually considering each would have individual requirements in terms of how the data is used? You are free to assume a certain constraint from your own experience and solve this question.**

1. It requires a continuous data flow ingestion (arguably made by Airflow) to be analyzed with pandas. I used gps-tracking systems to find the distance and I would use time-stamp information to extrapolate the number of trips (eg: how long the truck stays without moving)

b-c) To evaluate the impact of CO2OPT other data are required: either data from trucks-trips before CO2OPT guidelines implementation (preferred), or from other general data regarding tracks not in CO2OPT. This would give us a ‘counterfactual’ to evaluate the difference in terms of emission.

For c, in particular, I’d use panel information (same-truck data over time) and plot the difference in CO2 emission over time.

**T3 - Write a sample code using Airflow to enable:**

**a. Data injestion from a sample endpoint: http://localhost:3000/random-endpoint-1 with a JSON payload into a dataware connected to Snowflake**

**b. Orchestrating data analytics to enable KPI calculations for either a/b/c (only 1)**

You can find the code for a) attached in the email.

**T4 - Explain in 5-7 sentences, what kind of pre-processing strategies and tools have you used in order to ensure the trueness of your analytics?**

If focused more on tasks 1, 2 and 5. The implementation of n 3 and 4 would come easy with multiple time data. For the km consumption, I calculated the distance travelled by the truck (using latitude and longitude) and put it in relation to the fuel used. With multiple time data, the result would be more precise, but the code gives an idea of the strategy used. For the comparison with other vehicles, I used a clustering method to group the vehicles. The main function at the end of the code allows to compare trucks from the same group. For the usage profile, I used random C02-consumption thresholds and divided the trucks in 3 categories.

e