Technical Data Exercise #3
Pipeline Assembly

Every day, a report is sent to your company in the form of a CSV file with driving operations for company vehicles and drivers. The reports include information about the trip and an ID which links the trip to an employee of the company. The files are placed in a Google Cloud Storage bucket (*"gv-data-logs-[id]"*) by a company employee every evening. Today's file, left in the bucket, is named "*gv_trips*_[date]_*europe_report1.csv*". These files must be processed to make key business decisions.

Elaborations are enabled by company data hosted in the data warehouse BigQuery in the dataset "*dictionaries*". The dictionaries of company data enabling this elaboration include tables with vehicle information, vehicle operator information, and earnings multipliers used to calculate the wage payment expected by the operator for his trip. The elaborated data should be stored in a manner that allows for future access, helping to inform and contextualize decisions and analysis of future daily files.

Specifically, the cost modelling table contains the raw multipliers applied for hours that occur between 22:00 and 07:00 (local time), labelled as **night**, and the remaining window labelled as **day**. The trip's operator cost should be calculated using these multipliers.

It is crucial to track the extraction and processing of the received files from the first extraction onward. This tracking enables monitoring and evaluation of pipeline runs with metrics like runtime, received timestamp, and other relevant metadata. The storage of these metrics is at the discretion of the project owner (you).

At this initial phase of the project, at the end of each day, a notebook is run which gives and present insights into the day's operations. The notebook (written in your language of preference with any libraries necessary) must address at least 5 of the following 8 questions:

a. What is the total labour cost of each team of operators?
b. Comparatively, what was the total distance travelled by operators in each country?
c. What was the daily average speed of each type (Vehicle + Brand) of vehicle?
d. Which contractor companies were most active?
e. Are there any key outlier trips worthy of alerting a local country manager? On what grounds?
f. What was the closest city to the starting point of the operators with the four longest trips?
g. How does average trip labour pay vary per country [of trip origin]?
h. How much must be billed to each company to cover operator salaries?

All results must be automatically reproducible daily for each new file with no manual intervention required (except for running the notebook)**.** If you consider an alternative visual/result presentation of the data insights superior to a notebook, you are encouraged to construct your alternative instead – using any cloud ecosystem or locally hosted program you would like. The goal is to minimize manual intervention as much as possible.

At every step, attempt to use the strengths of every layer of the pipeline; a purely pythonic solution is less useful and maintainable then a system with properly distributed responsibilities.

The final project should not take more than **8-10 hours** to construct, depending on your familiarity with the tools at your disposition. If it takes longer, do not hesitate to submit the components built without worrying about the total completion of the assignment. In any case the solution proposal and completed components will be discussed and reviewed.

Importantly, a diagram of the pipeline solution in its entirety should be drawn and submitted when the assignment is reviewed, regardless of whether the assignment was successfully completed.

GCP Account Information:

Important. The Cloud Project is not connected to the internet. For any external data requirements, download and upload them manually.

i. *To be inserted… alongside practical instructions*

Please refer to acrossley@govoltmobility.com with any questions or concerns!