Technical Test

Context

A **taxi company** running in a metropolitan city is looking at **generating analytics** to get **insights into revenue** and the **customer base.**

Two types of taxis: **Yellow** & **Green** taxi

Attributes captured by the taxis

- Vendor (taxi type)
- Pickup & drop time
- Passenger count
- Pickup & drop location details
- Trip distance
- Payment – Rate code, type, amount, tax, tip, toll, surcharge & total

Ask from IT:

Develop **Self servicing** data analytics capability for business users allowing them to *create* **insights** on **Revenue** & **Customer base**

Raw data file (attached separately in email):

taxi_raw_data.csv

Exercise

1. Ingest raw data & identify quality issues
2. Build a simple data pipeline to model the data with quality checks where needed

- You can use any language you like (Python, Scala etc.)
- You can use any IDE – if you don't have one setup right now, you can try this one online:
  - Online Jupyter notebook: https://jupyter.org/try

Understandably you have short time for this, try to complete the basics – ***Data quality has a higher priority.***