# Data Engineer Technical Assessment (GA)

This technical assessment aims to measure a candidate's ability to solve data engineering problems in a business context.

This first page of this document outlines the parameters of the assessment, the second page is the assessment, subsequent pages will be appendix or supplementary materials.

You **must** use Python/SQL to perform this assessment.
We expect that this assessment will take a couple of hours to complete.

## Materials to submit

You MUST provide the following in a zipped archive by email, if not provided, it may result in your assessment being rejected:

1. The code that you wrote to perform the task
2. Two slides summarising your approach and results of your work so that the business analyst can understand and interpret the results for the business
3. A short README outlining the contents of your submission

## Expectations

This exercise is for us to get to know how you think about problems, and provide us with discussion material in a technical interview rather than an exam with right or wrong answers.

Your code should be well structured, suitably commented, have error handling and unit tests where appropriate.

Any comments should help give context about why you have made a particular decision, as opposed to simply repeating what the code already says. Likewise, log messages you include should indicate forethought about supporting / debugging in a live environment.

100% test coverage, exhaustive error handling, and full logging implementations are not required. We only wish to see that you know how to write testable data pipelines, how you deal with errors, and how you use logging to help with debugging in a production setting.

## Post-assessment

At the interview, we will use your submission as the basis for discussion.

Additionally, you may be expected to walk through your outputs with the assessor. For example, justifying the design decisions you made and walking them through your thought process.

You should be prepared to discuss how you would scale the code you've written, including the trade-offs of your choices.

# Assessment

## Problem Statement

An online retailer has asked us to analyse their transactional data over the past decade.

They have asked us to build out the initial data pipelines to support this. They have sent some sample data with the expectation that your code will be run against the larger dataset.

You have received an email from one of the business analysts detailing how the pipelines should work.

## Email

Morning, a few different requests have come in from across the team after yesterday's meeting. We'll need a few pipelines to analyse the data.

The head of sales is interested in the total value of cancelled orders, as well as the total value of orders currently on hold. Do you think we could get these segmented by year as well?

One of the supplier leads wanted to know how many unique products per product line they have. On top of that, they floated that they'd like to understand the sales trend in the number of classic cars they've shipped. Admittedly, I didn't capture how that was defined but could you get something going in the meantime and I'll try to find out what they meant - something more informative than total value would be great.

Finally, there was some talk about reviewing the discounting scheme that they have. It works by applying bulk discounts against the MSRP. I got one of the sales team to send me an illustrative copy of the discounting scheme which I've attached below. For this one they're only interested in vintage or classic cars, motorbikes, trucks and buses.

### Requirements

1. A pipeline that transforms the data into daily partitions in the Parquet format
2. A pipeline (or pipelines) that outputs the analysis suggested in the email
3. A pipeline that computes the volume-based discount suggested in the email (see Appendix)

# Appendix

## Quantity Discounts

| **Quantity Range** | **Discount** |
|---|---|
| 0-30 | 0.0% |
| 30-60 | 2.5% |
| 60-80 | 4.0% |
| 80-100 | 6.0% |
| >100 | 10% |