Data Engineer Interview - Technical Test

Background

You are a Data Engineer working as part of a JLR Analytics team where you are enriching a dataset for the purpose of analysing the profitability of 'options' on cars.

You have downloaded a base dataset from the accounting system. The base dataset includes the sales price for each Option on a vehicle which has been sold in the last few years, the unique identifier of a vehicle is the VIN (Vehicle Identification Number).

Your goal this project sprint is to enrich the base dataset with the production cost for each option so that the overall profit can be calculated. Profit = sales price - production costs

You have been emailed a dataset from another team which contains option costs. You have done some initial analysis on this and found that the dataset does not contain a 100% match for the option costs.
An analyst in your team has come up with a methodology to create a production cost if it is not in the dataset.
This is the logic that the analyst has come up with, the logic should be implemented in this order;

1. If sales price is Zero or negative then production cost should be zero
2. If there is a matching record in the options dataset per option code and model, then use the material cost from the option dataset
3. If there is no exact match in the options dataset, then use the average across all models for that option code; i.e. material cost = avg("material cost")
4. If there is no matching option code then assume that the production cost is ≈45% of the sales price

It is your job to implement this calculation, enrich the base dataset with the production cost for each record to calculate the profit per option on each vehicle.
If you find any poor quality data or errors with the dataset please make a note and bring them up at the interview.

Technical assessment

* We would prefer to see Python or SQL code used for this task if you have the skills and experience to do so. Otherwise please use an environment/application that you are most comfortable with. *

Write code to do the following. You will be asked to show the code you have written and run it during the interview.

1. Create automated process to ingest and store the data on an on-demand basis.
2. Write an ETL query to show that you can implement the joins and logic to enrich the dataset.

Consider the following questions that you will be asked during the interview. You will only be asked to explain what you would do, not to have implemented it in your code.

- How would you test and validate that your work is correct?
- Explain what you would do make your code ready to run in production.
- What tools would you use to deploy your code in a production environment?
- What aspects/design choices of your proposed solution has the most effect on the performance?
- Discuss what is needed to future-proof the solution (e.g. business or cloud provider's technical specification changes)?

You will be asked to show your ways of working on screen during the interview, explain the methods you have used and how you have implemented the solution. Feel free to bring supporting diagrams or notes with you to help explain your solution.