Junior Data Engineer Task

Your task is to design a relational data model and ingest schemaless JSON into a SQL database. The input data represents Kafka events recorded by the order service, which processes orders that come into a digital ordering system from various website stores

We would like you to spend no more than 3 hours on the task. You may find that this is not enough time to complete all parts of the test, but try your best and get as far as you can in the allotted time. You can submit your answers via zipped files or a link to a Github repository

Input

The input is a zipped standard JSON file consisting of an array of nested orders with various attributes. The majority of our integrations are written in Python 3, so a script written in Python 3 would be preferred, however other languages can be used. You are free to use any libraries needed to complete the task.

Note: the names of fields in the JSON are not always semantic, and may not mean the same thing as they do in the real world and this should not affect the extraction of the data. The data will also be heavily anonymised with dummy data

Data Modelling

Create a normalised Entity Relationship Diagram (ERD) showing how you would model the data from the JSON file. You are free to submit this diagram however you please, this can be a screenshot of a hand drawing, or created via a free online tool. It just must be readable

Processing

The JSON file contains a nested structure and it can be assumed that the destination of the data is a SQL based database. This will require the file to be normalised for performance reasons. It is up to you the level of normalisation you choose. Your project should also show some level of testing and documentation.

For this task you do not have to create a connection to a database, simply produce the Python & SQL code that when run would load data into a database/data warehouse. Should you have the time and feel comfortable, you can attempt to use SQLite, PostgreSQL, or MySQL in a docker container - note, you will not be marked differently by using docker