Data Engineer Technical Test

Create an ETL Pipeline that is able to handle the following scenarios:

Scenario 1 - ingest, process and load raw data from a csv file to a SQL database

Scenario 2 - ingest, process and load raw data from a json file to a SQL database

The developed ETL Pipeline should have the following features:

- clear data transformation and cleaning functions
- extendable to allow ingestion of new raw data sources
- extendable to allow loading into new data stores
  e.g. storing the output in an object store instead of a SQL database
- consist of modular and reusable components
- incorporates unit tests
- logging

The expected submission is a working demonstration of the created Pipeline that can be independently run by us. The submission must consist of the following:

1. source code of your ETL pipeline hosted on a public GitHub repo

2. a runnable docker container that hosts your developed ETL pipeline / Provide a Jupyter Notebook to show the code execution with desired output

3. a runnable docker container that hosts a SQL database that acts as the target to load the processed data from the ETL pipeline / Provide a Jupyter Notebook to show the code execution output which shows the processed data is loaded in a SQL database

4. instructions on how to run your demonstration and explanation of the expected outputs

You will be provided with the following resources:

1. A csv file called test.csv
2. A json file called test.json

We would like your pipeline to ingest the provided files SQL table requirements:

● Create a table called test from the provided csv file. The table must conform (i.e., comply with rules, standard) to the following requirements:

○ Each row must have a unique identifier
○ Columns created_at and last_login must be of type timestamp
○ Column is_claimed must be of type boolean
○ Column paid_amount must be of type numeric with 2 decimal places

● Create 3 tables called users, telephone_numbers and jobs_history from the provided json file.

The created tables are to have the following requirements:

○ Tables user, telephone_numbers and jobs_history must be linked to each other via the appropriate foreign key relationship
○ Columns created_at, updated_at, and logged_at must be of type timestamp
○ Columns dob, start, and end must be of type date

Additional requirement:
Personally Identifiable Information (PII) or sensitive data should be masked where applicable