Instructions

Read carefully all the requirements and create a product that fits the expectations. You have to create a public GitHub repository to upload the exercise and share the link in the answer field.

Requirements

Create a data processing pipeline with the requirements below. The repository must include a README.md file containing:

- An architecture diagram illustrating how the solution would be implemented in a cloud environment of your choice (AWS, GCP, Azure, etc.).
- A brief explanation of the design choices and the technologies used.
- The Code can be performed in either a Jupyter Notebook or *.py* file.
- Run instructions providing clear steps to set up and execute the project.
- Answer the questions in *section 2* showing actual code.

1. Data Ingestion and Processing with PySpark

- Download all Cafe Rewards Offer datasets from Kaggle: [Café Rewards Offer Dataset.]
- Ingest the raw data from the offers, customers, and events datasets using a tool of your choice (using PySpark is a plus).
- Perform necessary transformations to clean and structure the data.
- Move the data from the raw layer to the trusted layer (e.g., removing unwanted columns,joining datasets, enriching data).
- Further process the data to create a refined layer (e.g., aggregations, metrics, or summary tables).ps: *Hint: check the analytical questions in step 2.*

2. Analytical Questions

Based on the processed data, answer the following questions:

- Which marketing channel is the most effective in terms of offer completion rate?
- How is the age distribution of customers who completed offers compared to those who did not?
- What is the average time taken by customers to complete an offer after receiving it?

Presentation and Code Review

Once you have completed the exercise, you will be required to present your project to the technical interview panel. During the presentation, you should explain your user story, design choices, technical architecture, and demonstrate the functionality of the application. This will be done over Google Meet, and you will screen share either your GitHub repository or IDE.
After the presentation, the interview panel will conduct a code review of your project. You will be asked to explain your coding decisions and answer any questions related to the code. The interview panel will evaluate your project based on the following criteria:

- Clean Architecture:
    - Your architecture should adhere to Clean Architecture principles, including separation of concerns and independence of components.
    - Ensure clear separation between the raw, trusted, and refined layers.
- Functionality
    - Your application should perform as expected according to the requirements, correctly ingesting, processing, and storing data.
    - Show that the processed data is accurate and ready for analysis.
- Data Processing:
    - Explain the logic behind the data processing part, including how you joined and transformed the datasets.
- Analytical Insights:
    - Present the answers to the analytical questions based on the processed data.

Scoring

Each criterion will be weighted equally, and the maximum score for the exercise is 100 points. We hope this exercise challenges your technical abilities and helps us assess your suitability for the Data Engineer position. Good luck!