Data Engineer Tasks

Deadline: End of Tuesday (10 October 2023)

Task 1 Description: ETL Pipeline and Dockerization

Your task is to design and implement an ETL (Extract, Transform, Load) pipeline to handle the following data-related operations:

1. Import CSV data into a MySQL database.
2. Calculate aggregate data and store it in a PostgreSQL database, specifically:
    - Calculate the weekly average activity for each email domain.
    - Calculate the total activity for each email domain.
    - Filter out invalid sessions that have less than 2 users and a duration of less than 5 minutes.
3. Create a Dockerfile for the ETL codebase to containerize the application.
4. Create a Docker Compose configuration to automatically deploy both MySQL and PostgreSQL databases and run the ETL code within containers.

Nice to Have:

- Use pure Python as the programming language for the ETL code.
- Add Grafana to the Docker Compose setup with pre-loaded dashboards for monitoring purposes.

Submission: Please submit your completed task either by pushing the files to a Git repository and sharing the link or by sending the files as a zip file. Ensure all points are completed and provide a 1-2 sentence summary for each step, encompassing the encountered challenges, details about the implemented solutions, and, where relevant, explanations for not pursuing alternative approaches if they exist.

Task 2 Description: Data Mesh Implementation and Challenges Assessment

Background: You are tasked with implementing a data mesh architecture within a large, diverse organization that has multiple data sources, teams, and departments. The goal is to improve data accessibility, reliability, and scalability while fostering a culture of data ownership and collaboration.

Instructions:

Understanding Data Mesh (30%)
Provide a brief overview of what a data mesh is, its key principles, and the reasons why an organization might consider implementing it.Implementation Strategy (25%)

- Outline your approach for implementing a data mesh within the organization. Describe how you would identify data domains, define ownership, and establish data products.
- Explain how you would promote the concept of data ownership and a culture of data collaboration among different teams and departments.
- Discuss the role of data infrastructure and platform services in supporting the data mesh implementation.

Challenges Identification (25%)

- Identify and discuss at least three significant challenges you anticipate when implementing a data mesh in the organization. These challenges could be related to technology, culture, governance, or other aspects.
- For each identified challenge, propose potential solutions or mitigation strategies.

Tool Selection (20%)

- Recommend specific tools and technologies you would use to implement the data mesh architecture. Explain your choices and how they align with the principles of data mesh.
- Consider data integration, data cataloging, data quality, and data governance tools, among others.

Submission: Please provide a written document that addresses each of the points outlined above. You may use diagrams, charts, or any other visual aids to support your explanations where necessary.