Better Collective

Data Engineering Assignment

Overview

Congratulations on advancing to the next stage! We are excited to present you with a set of tasks designed to assess your data engineering skills and problem-solving abilities.

Structure

The assignment consists of two distinct tasks:

- The first task focuses on ingestion and analyzing a large dataset, while keeping memory limitations in mind, testing your ability to handle data efficiently and script effectively.
- The second task involves designing a Data architecture, assessing your understanding of data integration and tool selection in constructing effective data pipelines.

Task 1 - Dataset Ingestion and Processing:

You will be analyzing a dataset, focusing on extracting insights.

- **Dataset Access**: Please find your dataset on this link.

Important Notes:

- **Resource Constraints**: Ensure that the dataset ingestion process is optimized for deployment in an ECS environment with a strict memory limit of 4GB.
- **Approach and Assumptions**:Explain your approach clearly—why you chose certain methods and how they help with memory efficiency. List any assumptions you made!

Time Limit and Focus:

- Aim to complete this task efficiently, focusing on clarity of thought and execution within 120 minutes.

Task 2 - ETL Architecture Design

Create a data architecture to seamlessly integrate data from various sources—APIs, databases, and Kafka. It's also possible to design multiple architectures and discuss the trade-offs between them.

Data Sources:

- APIs: Access dynamic and potentially real-time data.
- Databases: Connect to structured data stored in various databases.
- Kafka: Handle high-throughput message streams.

Architecture Design:

Develop a thorough diagram that illustrates your ETL process, showcasing how data flows efficiently from source to target systems.

Tool Selection and Justification:
Explain your choice of tools for extraction, transformation, and loading, highlighting scalability, reliability, and modernity

Important Notes:

- **Visual Representation**: Leverage diagramming tools of your choice to visually communicate your ETL design.
- **Thought Process:** We value your reasoning and understanding of the workflow over the specific technologies chosen.

Time Limit and Focus:

- Please aim to complete the following task efficiently, focusing on clarity of thought and completion within 60 minutes.

We look forward to seeing your innovative approach to these tasks and understanding your thought process at each stage. Good luck!