Panaseer – Data Ingest Engineer - Technical Exercise
Transforming Cocktail Data into a Common Information Model

Overview

As a Data Ingest Engineer, you will work with various data sources to ingest and transform data into a unified format for our platform. This exercise simulates a typical task you might encounter in the role, using TheCocktailDB API to retrieve and manipulate data. While ingesting the data is important to the role, understanding trends in the data and explaining decisions you've made will be critical.

Objective

- Ingest data from a publicly available API.
- Transform raw data into a common information model.
- Handle edge cases and data inconsistencies.
- Make informed decisions when faced with different options.
- Consider backward compatibility in an ETL (Extract, Transform, Load) process.
- Demonstrate your problem-solving approach and communication skills.

Instructions

1. Data Ingestion

- **Task**: Use TheCocktailDB API to fetch cocktail data.
  - API Documentation: https://www.thecocktaildb.com/api.php
- **Goal**: Retrieve a dataset of cocktails that can be used for transformation.
- **Note**: No authentication is required to use this API.

2. Data Transformation

- **Task:** Map the raw cocktail data from the API to a common information model of your design.
- **Requirements:**
  - Define a schema that includes essential fields (e.g., `CocktailID`, `Name`, `Category`, `Alcoholic`, `GlassType`, `Ingredients`, `Instructions`).
  - Justify why if you included certain fields and excluded others.
  - Handle data types appropriately (e.g., strings, lists, boolean values).

3. Edge Case Handling

- **Task:** Identify and address potential edge cases in the data.
- **Requirements:**
  - Implement solutions to handle one or two edge cases as examples.
  - Document the edge cases you identified and how you handled them.
- **Examples of Edge Cases:**
  - Cocktails with missing ingredient information.
  - Variations in measurement units (e.g., "1 oz", "30 ml").
  - Cocktails with special characters or formatting issues in their names or instructions.

4. Options and Choices

- **Scenario:**
  - You have two options for data transformation:
    - **Option A:** Transform data during ingestion (real-time processing).
    - **Option B:** Store raw data first and transform it afterward (batch processing).
- **Task:**
  - Choose one of the two options. Write a brief statement of your choice and reasoning. Bullet points are fine.
- **Requirements:**

    - Be prepared to explain your choice in the interview, outlining the advantages and disadvantages of both options.
    - Describe any risk mitigation strategies you would employ.
    - Discuss how your choice impacts scalability, performance, and maintenance.

Deliverables

- **Code Repository or Written Report:**
    - **Option 1:** Submit your code via a Git repository (e.g., GitHub, GitLab) with appropriate documentation.
    - **Option 2:** Provide a detailed written report explaining your approach and solutions.
    - **Include:**
        - Clear instructions on how to run your code (if applicable).
        - Comments and documentation within your code.
        - Explanations for your design decisions and thought processes.
- **Be prepared to talk through your work:**
    - Your overall approach and workflow.
    - How you handled data transformation and edge cases.
    - Your decision-making process for the ambiguous challenge.
    - Your strategy for ensuring backward compatibility.
    - Be prepared to answer questions and engage in discussion.
    - You may prepare some simple diagrams to help explain your points, but please do not prepare slides.
    - Think of this more like pair-programming or a collaborative code-review as you'd encounter it working with the team in the role.

Time Commitment

- The exercise is designed to be completed in **2 to 4 hours**. Please do not take more than 4 hours or you risk over-cooking it.
- If you run out of time to complete all the tasks, don't worry. Rather than coding until the last-minute set yourself an end time and use 30 minutes at the end to reflect how you can tell us about your experience and how you approached the work. If you didn't get through all the scenarios perhaps think about how you would've approached the next one and tell us about it. If you did get through all of them then tell us what you would do next with more time.

Support and Questions

If you have any questions or need clarification, feel free to reach out to us at any time.