Machine Learning Platform Engineer Take-Home Assignment

Introduction

We are excited to invite you to participate in our take-home assignment for the **Machine Learning Platform Engineer** position at **Engine By MoneyLion**. This assignment is designed to assess your ability to work with data and database design, implement ETL processes, develop APIs, and incorporate monitoring, logging, and testing into your solutions.

Objective

Build a data processing pipeline that ingests raw data, processes it, stores it in a database, and exposes it via a RESTful API. Your solution should include monitoring, logging, and testing to ensure reliability and maintainability. **Please complete this assessment in Python.**

Estimated Time to Complete

Approximately 4-6 hours. We understand that your time is valuable, **so please focus on the core tasks (Tasks 1-3)**. If you have extra time, feel free to tackle the additional tasks.

What We're Evaluating

We aim to assess your skills in the following areas:

1. **Data Handling and Database Design**: Efficiently designing a relational database schema and managing data ingestion, including handling data quality issues.
2. **ETL Processes**: Building robust data pipelines to extract, transform, and load data, performing meaningful data transformations and aggregations.
3. **API Development**: Developing RESTful APIs following best practices, with proper error handling and input validation.
4. **Monitoring and Logging**: Implementing logging for debugging and auditing, and setting up basic monitoring to track application performance.
5. **Testing**: Writing tests to ensure code reliability and correctness.
6. **Software Engineering Principles**: Applying best practices in code organization, readability, and modularity.
7. **Documentation and Communication**: Providing clear setup instructions and API documentation, and explaining design decisions and assumptions.

Please focus on these areas while completing the assignment, as they are critical to the role and will form the basis of our evaluation.

Dataset

You will work with a synthetic dataset related to user interactions on a fintech platform. The dataset includes two CSV files:

1. users.csv – Contains user information.
    a. user_id: Unique identifier for each user.
    b. signup_date: Date when the user signed up (YYYY-MM-DD).
    c. country: Country of the user.
2. transactions.csv – Contains transaction records.
    a. transaction_id: Unique identifier for each transaction.
    b. user_id: ID of the user who made the transaction.
    c. transaction_date: Date of the transaction (YYYY-MM-DD).
    d. amount: Transaction amount in USD.
    e. transaction_type: Type of transaction (deposit, withdrawal, purchase).

Tasks

Data Ingestion and Database Design

1. **Design a relational database schema** to store the data efficiently (Use OpenSource libraries like SQLite).
2. **Implement a script** to ingest the data from the CSV files into your database.

ETL Pipeline

1. **Extract** data from the database
2. **Transform** the data:
    - Calculate the total transaction amount per user.
    - Identify the top 10 users by transaction volume.
    - Aggregate daily transaction totals across all users.
3. **Loads** the processed data back into the database or prepares it for use by the API**.**

API Development

Develop a RESTful API to expose the processed data.

1. Get User Transaction Summary
2. Get Top Users
3. Get Daily Transactions

Monitoring and Logging

1. Describe the key events and errors you would track in your application
2. Describe how you would monitor application performance and health

Testing

If time permits, write tests for your application. Otherwise, describe what are the key testing methodologies you would need to implement in order to **verify functionality and reliability.**

Containerization

Containerize your application to simplify setup and serving.

Submission Guidelines

1. **Code Repository**:
    - Provide a link to a Git repository (GitHub, GitLab, etc.) containing your code.
2. **README File**:
    - Include a README.md with setup instructions and any notes.
    - Include notes on how you would approach incomplete tasks.

Please provide precise installation instructions for your application - e.g. a ‘requirements.txt’, or a poetry ‘pyproject.toml’ file so we can run the application locally.

Additional Notes

1. **Use of Libraries and Tools**:
    - You are free to use any open-source libraries or tools that you deem appropriate.
    - Keep dependencies reasonable and avoid unnecessary complexity.
2. **Assumptions**:
    - Document any assumptions you make during the development process.
3. **Questions**:
    - If you have any questions or need clarifications, feel free to reach out to us.

Good Luck!

We look forward to reviewing your submission and discussing it with you during the interview process. If you need any further assistance or have additional questions, please don’t hesitate to contact us.