Data Engineering Assessment

Tags	Data Engineer
Status	Not started
Type Of Work	Assessment

Data Engineering Assessment Task

Instructions:

- Please complete the following tasks within the stipulated time frame (typically 3-5 days).
- Feel free to reach out if you have any clarifications or questions.

Task 1: Data Ingestion and Cleansing

You are provided with a set of raw marketing data files in various formats, including CSV, JSON, and XML, containing data such as customer interactions, campaign details, and user profiles. Your task is to create a data pipeline that ingests this data, cleanses it, and stores it in a structured format, such as a relational database. Perform the following steps:

Sample Data Structure:

For the purpose of this assessment, consider the following sample data structures for the raw marketing data:

1. **Customer Interactions (CSV):**

   ```
   InteractionID,UserID,CampaignID,InteractionType,InteractionDate
   1,101,1,Click,2023-10-01
   2,102,2,View,2023-10-01
   3,103,1,Click,2023-10-02
   ```

2. **Campaign Details (JSON):**

   ```
   {
       "CampaignID": 1,
       "CampaignName": "Summer Sale",
       "StartDate": "2023-09-15",
       "EndDate": "2023-10-15"
   }
   ```

3. **User Profiles (XML):**

   ```
   <Users>
       <User>
           <UserID>101</UserID>
           <UserName>Alice</UserName>
           <Email>alice@example.com</Email>
       </User>
       <User>
           <UserID>102</UserID>
           <UserName>Bob</UserName>
           <Email>bob@example.com</Email>
       </User>
       <User>
           <UserID>103</UserID>
           <UserName>Charlie</UserName>
           <Email>charlie@example.com</Email>
       </User>
   </Users>
   ```

Tasks:

1. Write a Python script to ingest the raw data files from a specified directory using appropriate data reading libraries for CSV, JSON, and XML.
2. Perform data cleansing, including handling missing values, data type conversions, and removing duplicates based on the unique identifiers (e.g., InteractionID for interactions, UserID for users, CampaignID for campaigns).
3. Design a schema for a relational database that can accommodate this marketing data. You can provide the SQL schema creation script.
4. Load the cleansed data into the database tables.

Task 2: Data Transformation and Aggregation

Now that the data is in a structured format, you need to perform some transformations and aggregations for analysis purposes. Write a Python script that accomplishes the following:

1. Calculate the total number of interactions per campaign and store the results.
2. Calculate the average time spent on the website for each user and store the results.
3. Aggregate customer data to create a summary table with key customer metrics.

Task 3: Data Warehousing and Data Streaming

Imagine you work for an e-commerce company, and they need a data warehouse solution to store and manage their sales data for reporting and analysis. Additionally, the company wants to incorporate real-time data streaming for immediate insights. Design a comprehensive solution that includes both batch processing (for historical data) and data streaming (for real-time data).

Batch Processing:

1. Design a data warehouse schema using a data modeling tool (e.g., dbdiagram.io or draw.io). Your schema should include tables for Orders, Products, Customers, and Sales.
    - Explain your design choices, including primary keys, foreign keys, and data types, in a document.
2. Describe the batch processing pipeline for populating and updating the data warehouse with historical data. This should include steps for data extraction, transformation, and loading (ETL).
3. Implement error handling mechanisms for batch processing, such as logging and retry mechanisms, to ensure data consistency and reliability.

Data Streaming:

1. Design a real-time data streaming solution for capturing and processing sales data as it happens. Consider using a technology like Apache Kafka, Apache Flink, or Apache Spark Streaming for this purpose.

2. Explain the schema design for real-time data processing, including the structure of the streaming data topics or streams.
3. Describe how you would handle potential errors and data quality issues in the streaming pipeline. Include mechanisms for monitoring and alerting.

Task 4: Prepare a Non-Technical Presentation

Imagine you need to present your data engineering approach to a non-technical audience, such as company stakeholders or managers. Your goal is to explain how you would tackle the following three key aspects of data engineering:

Presentation Guidelines:

1. **Data Ingestion and Cleansing (Task 1)**: Describe your approach to ingesting and cleansing raw marketing data. Highlight the importance of data quality and how you would ensure that the data is accurate and reliable for analysis.
2. **Data Transformation and Aggregation (Task 2)**: Explain how you would transform and aggregate the marketing data to derive meaningful insights. Discuss the types of analyses or reports that could be generated from the transformed data.
3. **Data Warehousing and Data Streaming (Task 3)**: Outline your strategy for creating a data warehouse to store and manage the marketing data efficiently. Discuss the benefits of a data warehouse in the context of marketing data analysis.