Assignment Title: Data Engineering Case Study

Imagine you are a data engineer working for AdvertiseX, a digital advertising technology company. AdvertiseX specializes in programmatic advertising and manages multiple online advertising campaigns for its clients. The company handles vast amounts of data generated by ad impressions, clicks, conversions, and more. Your role as a data engineer is to address the following challenges:

Data Sources and Formats:

- Ad Impressions:
    - Data Source: AdvertiseX serves digital ads to various online platforms and websites.
    - Data Format: Ad impressions data is generated in JSON format, containing information such as ad creative ID, user ID, timestamp, and the website where the ad was displayed.
- Clicks and Conversions:
    - Data Source: AdvertiseX tracks user interactions with ads, including clicks and conversions (e.g., sign-ups, purchases).
    - Data Format: Click and conversion data is logged in CSV format and includes event timestamps, user IDs, ad campaign IDs, and conversion type.
- Bid Requests:
    - Data Source: AdvertiseX participates in real-time bidding (RTB) auctions to serve ads to users.
    - Data Format: Bid request data is received in a semi-structured format, mostly in Avro, and includes user information, auction details, and ad targeting criteria.

Case Study Requirements:

- Data Ingestion:
    - Implement a scalable data ingestion system capable of collecting and processing ad impressions (JSON), clicks/conversions (CSV), and bid requests (Avro) data.

    - Ensure that the ingestion system can handle high data volumes generated in real-time and batch modes.
- Data Processing:
    - Develop data transformation processes to standardize and enrich the data. Handle data validation, filtering, and deduplication.
    - Implement logic to correlate ad impressions with clicks and conversions to provide meaningful insights.
- Data Storage and Query Performance:
    - Select an appropriate data storage solution for storing processed data efficiently, enabling fast querying for campaign performance analysis.
    - Optimize the storage system for analytical queries and aggregations of ad campaign data.
- Error Handling and Monitoring:
    - Create an error handling and monitoring system to detect data anomalies, discrepancies, or delays.
    - Implement alerting mechanisms to address data quality issues in real-time, ensuring that discrepancies are resolved promptly to maintain ad campaign effectiveness.

This Ad Tech case study scenario focuses on the challenges and data formats commonly encountered in the digital advertising industry. Candidates can use this information to design a data engineering solution that addresses the specific data processing and analysis needs of AdvertiseX.