Data Engineer Assessment :
Building ETL Pipelines for Multi-source Data Ingestion

This assessment is designed to evaluate the skills and knowledge of a data engineer specializing in designing and building ETL pipelines for pulling data from various sources, including Facebook Ads, Google Ads, RDS, CleverTap, etc. The assessment will also cover the use of relevant tools such as Apache Airflow, Kubernetes, and other associated technologies.

Section 1: Data Source Understanding

1. Explain the differences between Facebook Ads, Google Ads, RDS (Relational Database Service), and CleverTap in terms of data structure, API access, and data types.

Section 2: ETL Pipeline Design

2. Design a high-level ETL pipeline architecture for extracting data from Facebook Ads and Google Ads, transforming it, and loading it into an RDS database. Consider data extraction frequency, data transformations, error handling, and scalability.

Section 3: Apache Airflow

3. What is Apache Airflow, and how does it facilitate ETL pipeline orchestration? Provide an example of an Airflow DAG (Directed Acyclic Graph) for scheduling and orchestrating the ETL process described in Section 2.

Section 4: Kubernetes Integration

4. Explain the role of Kubernetes in deploying and managing ETL pipelines. How can Kubernetes ensure scalability, fault tolerance, and resource optimization for ETL tasks?

Section 5: Data Transformation

5. Given a JSON data sample from Facebook Ads containing ad performance metrics, write a Python function to transform this data into a structured format suitable for storage in an AWS Redshift database.

Section 6: Error Handling and Monitoring

6. Describe strategies for handling errors that may occur during the ETL process. How would you set up monitoring and alerting mechanisms to ensure the health and performance of the ETL pipelines?

Section 7: Security and Compliance

7. Data security is crucial when dealing with sensitive user information. Describe the measures you would take to ensure data security and compliance with relevant regulations while pulling and storing data from different sources.

Section 8: Performance Optimization

8. Discuss potential performance bottlenecks that might arise in the ETL process, particularly when dealing with large volumes of data. How would you optimize the ETL pipeline to ensure efficient data processing?

Section 9: Documentation and Collaboration

9. How important is documentation in the context of ETL pipeline development? Describe the components you would include in documentation to ensure seamless collaboration with other team members and future maintainers of the pipeline.

Section 10: Real-world Scenario

10. You have been given a scenario where CleverTap's API structure has changed, affecting your ETL pipeline. Explain the steps you would take to adapt your existing pipeline to accommodate this change while minimizing disruptions.

Evaluation Criteria:
- Accuracy and depth of explanations.
- Clarity and coherence of architectural and technical designs.
- Demonstrated understanding of Apache Airflow and Kubernetes.
- Practicality of data transformation and error handling strategies.
- Thoroughness of security measures and compliance considerations.
- Creativity and effectiveness in performance optimization approaches.
- Importance of documentation and collaboration in the provided context.
- Real-world adaptability and problem-solving skills.

Please note that this assessment is designed to gauge a candidate's expertise in data engineering, ETL pipeline design, and associated technologies. The candidate's ability to effectively communicate their solutions, strategies, and thought processes will also be evaluated.

Data Engineer - ETL Pipeline Assessment

Instructions: This assessment is designed to evaluate the technical knowledge and skills of a data engineer specialising in building ETL pipelines for extracting data from various sources such as Facebook Ads, Google Ads, RDS (Relational Database Service), CleverTap, etc. using tools like Apache Airflow, Kubernetes, and other relevant technologies. Please provide detailed explanations wherever required.

Part 1: Data Source Integration
1. Explain the steps you would take to extract data from Facebook Ads and Google Ads APIs. Highlight the authentication process and any specific considerations for handling API rate limits.

2. How would you design an ETL process to pull data from a relational database (RDS) like MySQL or PostgreSQL? Discuss the factors you would consider to ensure efficient and reliable data extraction.

3. CleverTap provides event-based user data. How would you approach extracting and transforming this event data into a usable format for further analysis? Outline the key components of your ETL pipeline.

Part 2: ETL Pipeline Development
4. What is Apache Airflow? Describe its role in building and managing ETL pipelines. Provide an example of an Airflow DAG (Directed Acyclic Graph) that orchestrates the ETL process you described in question 3.

5. Kubernetes is a container orchestration platform. How can it enhance ETL pipeline deployment and management? Explain the concept of containerization and its benefits in this context.

6. In the context of ETL pipelines, what are some common data transformation challenges you might encounter? Provide examples of transformations you might need to perform on the extracted data before loading it into the destination.

Part 3: Scalability and Monitoring
7. Scalability is crucial for handling large volumes of data. How would you design your ETL pipeline to handle an increasing amount of data over time? Discuss any relevant techniques or tools you would employ.

8. Monitoring and logging are essential for maintaining the health and performance of ETL pipelines. What strategies would you implement to monitor the various components of your pipeline, and how would you handle error scenarios?

Part 4: Best Practices and Security

9. Data security is paramount when dealing with sensitive user information. Describe the security measures you would implement to ensure the privacy and integrity of the data flowing through your ETL pipeline.

10. What are some best practices for documenting your ETL pipeline? Explain the importance of clear documentation and how it can benefit both your team and other stakeholders.

Part 5: Practical Scenario
11. Imagine you are tasked with building an ETL pipeline to consolidate data from Facebook Ads, Google Ads, RDS, and CleverTap into a data warehouse. Outline the high-level architecture of your solution, including the role of Airflow and Kubernetes, the data flow, and the tools you would use for data transformation and loading.

Submission Guidelines:
- Provide detailed explanations for each question to showcase your understanding.
- Feel free to use diagrams, code snippets, or pseudocode to illustrate your points.
- Your solutions will be assessed based on technical accuracy, depth of understanding, and practical feasibility.

Note: This assessment is designed to gauge your expertise in building ETL pipelines using specific tools and technologies. It is recommended to take your time and provide thoughtful responses to demonstrate your skills effectively.