\*\***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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Store Methodlogy** | **DataStructure** | **Example** | **APIAccess(Authentication and Authorization)** | **APIAccess** | **DataTypes** |
| RDS | **Structured**, | Cell(Rows&  Columns) |  |  | Based on DB |
| Facebook Ads | **SemiStructured** | Simple or Nested Json(Key,Value pairs) | Oauth🡪Access token | [devlopment,basic,standard Graph explorer](https://developers.facebook.com/blog/post/2014/10/22/introducing-ads-api-access-levels/) | Strings,Int,  booleans |
| GoogleAds | **SemiStructured** | Simple or Nested Json(Key, Value pairs) | Oauth,Api Key,Service Account | [Devloper(liscnence),Access token(key) https://developers.google.com/google-ads/api/docs/oauth/overview](https://developers.google.com/google-ads/api/docs/oauth/overviewDevloper(liscnence),Access%20token(key)) | Strings,Int,  booleans |

\*\***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.

Pagination||API Limit calculation||Business need ||

🡪[Token expiry Code](https://colab.research.google.com/drive/1BjNHKwAEKsQkWkdvNAGQ3-DSKaaxi9q7?usp=sharing)

A screenshot of a computer

Description automatically generated

\*\***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.

[Airflow dag code](https://colab.research.google.com/drive/1RKYRzU8kNWDQNpd5yorRZG3HFmgf6pDm?usp=sharing)

DAG-Gives Priority/say what to execute in chronology

Open source python based framework which is used to orchestrate tasks developing, scheduling, and monitoring batch-oriented workflows and which provides lot of connectors and UI is pleasing .

Grid view, calendar view , Gantt view, Graph & Tree view

**Code in collab for scheduling**

\*\***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?

* Kubernetes is a Container orchestration platform
* Containers are called POD In Kubernetes.
* If containers are occupied. K8S can create another pod immediately and add the containers in multiple pods. Because it is Cluster by default.
* **Scalability:** Kubernetes has the Replica set. If application loads get increased it rolls out the Replica container. It’s a Manual process.
* You need to edit the yaml script and mention the replica numbers.
* HPA- Horizontal Pod Auto scaler. This will auto-scale the application.
* **Auto-Healing:** Kubernetes provides an Auto Healing feature. K8S Control & Fix the damages of containers. If a container goes down it will create a new container before it goes down are before it gets killed. With the help of API-Server.

Kubernetes does

* Automated Orchestration
* Load Balancing
* Horizontal and Vertical Scaling
* Self-Healing and Auto-Recovery

\***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.

[Collab Notebook](https://colab.research.google.com/drive/1DxWJzrYLKrNJgDW63DdzGAKdBJu_A4iq?usp=sharing)

🡪reponse status checks

🡪schema checks

🡪logging

🡪Semi Structured to Structured Data

\*\***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?

Monitoring :

🡪Cloud watch |SNS alerts (Pub/Sub)|Proper logging & Code travel |Avoid frequent retries|**Airflow UI**|**Delay Notification**

\*\***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.

1)Encryption at rest |AES algos| Frequent key rotation |Row and Column level security

2)Data Segregation |region|lob wise|Parittioning|

3)Audit logs , who is querying data

4) Creating different groups for access

\*\***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?

🡪 Resource insufficiency , initially it might be running for few mb of data later it might increasing and the job fails or particular batch size has increased.

🡪Using right loading strategy

🡪Breakage of transformation at join

🡪Predicate push down

🡪Joins optimization

🡪I/O , N/W Bandwidth issues

🡪 Native function in code & like avoiding udf’s|**Spark UI**

\*\***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.

Docs gives a

🡪kickstart in development/enhancement of the project

🡪Troubleshoot/re-fix| note down dev time issues 🡪Communication register

🡪Mapping document and solution diagram helps the team using data to understand the process much .lookup details.

To be noted …

🡪high level architecture diagram

🡪solution diagram

🡪Source to target matching along with explanation of transformation needs & logic.

🡪Monitoring SOP’s ,schedule,compute,upstream and downstream details.

🡪Careful details on Security implementation and Data Segeration logics if any.

🡪Test & Compliance results |backward comaprtability

ETL specifications, ETL diagrams, or ETL metadata, ETL monitoring procedures, ETL debugging, or ETL updating.

\*\***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.

1)the existing code is scalable mean the endpoint details, depth of the key and misc. ,should be maintained as Json data and any changes can be adhered

2)reject records can stored in some glue table so DA and dev team can be actively see the issue in case of Json structure changes .

3) **there can be two logics if the particular key not exits directly then recursively keys can be explored and depth are noted and values can be retrieved if still key could not be found then it can be Raised as an exception.**

DQ—schema level,null percentages ,match percentage ,anamoly detection