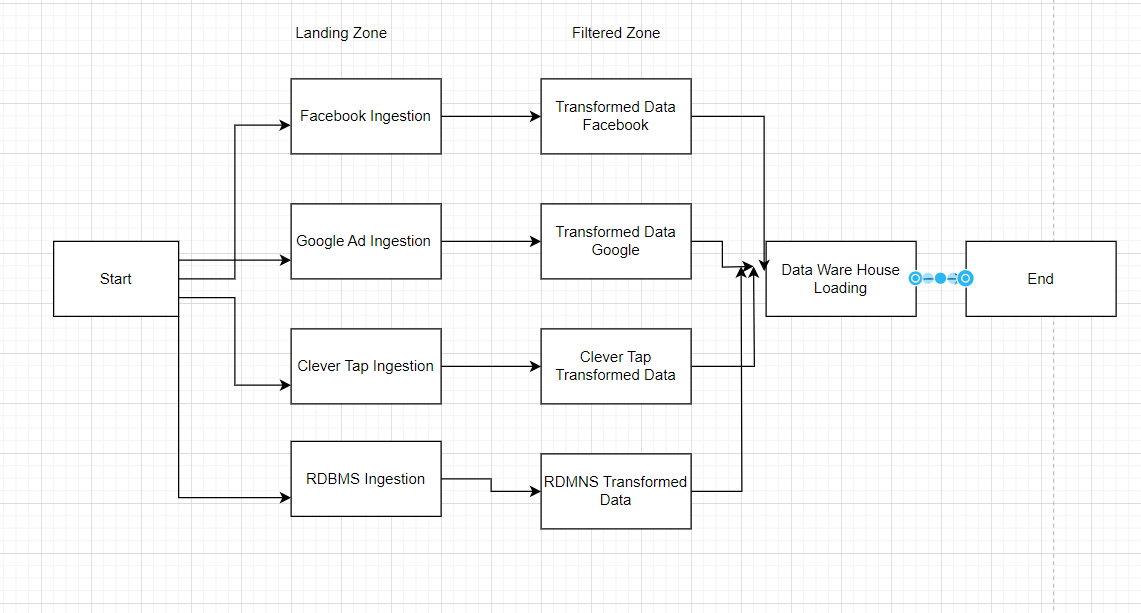
11. Imagine you are tasked with building an ETL pipeline to consolidate data from Facebook

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.

Airflow Architecture



Airflow: For orchestration and scheduling ETL tasks

Kubernetes:

Containerize the data transformation logic using Docker containers.

Manage these containers with Kubernetes for scalability and easier deployment across environments.

Airflow can interact with Kubernetes to schedule and manage containerized transformation tasks

Processing Framework:Apache Spark

Storage Layer -HDFS

Landing Zone

Separate Task for Facebook,Google Ad, Clever Tap,RDMS- To Store the data in fixed format like Orc or json with Append mode partitioned on Date to keep track of the Historical Data with the Retention of 5 months

Filtered Zone

Loosely Coupled Task for the fault tolerant with the Savemode as Overwrite to store the Current Filtered data of all the data in Unified Schema if it is representing the single Target table or Multiple Schema if the target table is different

Data is cleaned and filtered and has derived values as to uniquely identified the record

Data Warehouse

Final task is to load the data filtered data (techniques to handle the incremental load) into Datawarehouse and log the metrics to keep track of the DAG Status incase of failures