\*\*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.

Apache Airflow is an open-source platform designed specifically for programmatically authoring, scheduling, and monitoring data pipelines. It allows you to define workflows, automate tasks, and manage dependencies between those tasks using uspstream and downstream

1. Airflow uses Directed Acyclic Graphs (DAGs) to define workflows. Each DAG represents a complete ETL process, consisting of tasks with defined dependencies
2. Airflow allows you to schedule tasks within a DAG to run at specific intervals
3. Airflow provides a user interface to monitor the execution status of tasks within a DAG. It also logs execution details for debugging and troubleshooting purposes.

Example of the DAG for the Clever Tap API

START >> DATA INGESTION(RAW -Storing Data As it is)>>DATA TRANSFORMATION(Cleaning the Data based on the Events validating the business Logic) >>DATA LOADING(into Target System full refreshed or incremental) >>END

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

**Containerization:**

Containerization is a technology for packaging software applications along with their dependencies into standardized units called containers. These containers share the operating system kernel of the host machine but run in isolation from each other, ensuring consistent execution environments regardless of the underlying infrastructure

**Features**

**Isolation:** Containers isolate each stage of your ETL pipeline, preventing conflicts between dependencies and libraries used by different tasks.

**Portability:** Containerized ETL pipelines can be easily deployed across different environments (development, testing, production) without modification.

**Scalability:** You can easily scale individual tasks within the pipeline by adding or removing containers as needed.

Kubernetes is a container orchestration platform that automates the deployment, scaling, and management of containerized applications

**1.Simplified Deployment:** Define your ETL tasks as containers and deploy them as Kubernetes pods. Kubernetes handles scheduling and resource allocation

**2.Scalability:** Easily scale individual tasks or the entire pipeline based on processing needs

**3.Self-healing:** If a container crashes, Kubernetes automatically restarts it, ensuring pipeline resilience

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 onthe extracted data before loading it into the destination

**Data Type Length** –(Consider in the case of RDMS -Data Type length is of fixed inconsistent length of the targeted data)

-Define the Fixed length in retrieving the data

**Inconsistent Formatting:** Data might have inconsistencies in formatting (e.g., dates in different formats, addresses with varying levels of detail)

- Define standard formats for dates, addresses, and other relevant data points and convert all extracted data to match those formats

Outliers-Total Bill Value and the Purchase bill value is of different value

-Investigate the data to find the Error

**Schema Incompatibility:** Data from different sources might have different structures and field names

Define the Common schema for all data points

**Data Aggregation:** Summarize data points into higher-level summaries for analysis.

* Aggregate sales data by product category or user activity by day/week/month

**Data Quality Checks:** Ensure data adheres to defined rules and formats before loading it into the target system.

* Implement data validation rules to check for invalid email formats, address structures, or out-of-range values

**Enhance Data with Additional Context:** Combine extracted data with data from other sources for a more holistic view.

Join customer data with product information from a product catalog to enrich user profiles with product preferences