**Question 4:**

In general terms, how would you design a system where data from three different sources is combined into one large SQL table? Assume that the source data changes at different rates, and the destination table should update more than once per day.

**Answer 4:**

The process to update data into a target table from three different data sources involves several key steps:

1. **Data Ingestion into Source OLTP Table**:
   * First, data is collected from various external sources. These could be databases, APIs, or other systems where transactions happen. This raw data is stored in a **source table** (which is like a temporary holding area) for further processing.
2. **Define Target Table schema based on the business requirement**:
   * Based on what the business needs, we create a **target table** that will store the final data in a structured way. This table’s structure is designed to be efficient and easy to query, even though the data is coming from multiple sources that may have very different formats. We define the fields, data types, and relationships so it all makes sense for the business.
3. **Processing the Data While Ensuring No Duplicate Work**:
   * For each data source, we write a **procedure** that extracts the data from the source, transforms it as needed, and loads it into the target table.
   * To avoid processing the same data more than once (for example, if the process runs again), we use a **processed flag**. Once a piece of data is successfully moved to the target table, the flag is set to "True," so it’s not processed again. This way, the process is **idempotent** meaning we can run it multiple times without affecting data that’s already been processed.
4. **Handling Errors During Data Processing**:
   * Sometimes, things can go wrong during the process. Maybe the data is in the wrong format, or some fields are missing. To manage this, we have a special **error-handling procedure**.
   * If something goes wrong, this procedure catches the error, logs it in a separate error table, and allows us to review and fix issues later. This way, we don’t lose track of any problems.
5. **Orchestrating the Whole Process**:
   * Since data from each source might update at different rates, we use an orchestration tool like **Airflow** to manage the scheduling. Airflow allows us to set up a workflow that calls the different procedures at the right times, ensuring that data from all sources gets processed regularly and efficiently.
6. **Using the Final Data:**
   * Once the target table is populated with all the necessary data, it’s ready to be used. The business can now access this cleaned and processed data for reporting, analytics, or whatever purpose is required.

Let’s understand using example in detail:

**Architecture**

A diagram of a company

Description automatically generated

This system is designed to take data from three different sources — **Sales**, **Inventory**, and **Product** — and combine them into a central table in **Snowflake**, called **product\_performance**. The goal is to ensure that the central table is always up to date with the latest data from each source, even though each source updates at different times.

**Step 1: Data Ingestion**

The process begins by **ingesting data** from Excel (or a similar data source) using **Python with Pandas**. This data ingestion pulls data from the three different sources: Sales, Inventory, and Product. The data is then loaded into separate staging tables in the OLTP system, which eventually feeds into Snowflake for further processing.

**Step 2: Source Tables in Snowflake**

Once the data is ingested, it is stored in three separate **source tables** in Snowflake:

* **Sales Table**: This stores all sales-related data, such as transaction details.
* **Inventory Table**: This holds data related to stock levels and inventory movements.
* **Product Table**: This contains information about product pricing and details.

Each of these tables is updated at different intervals because the rate of change for each type of data is different. Sales might need to be updated more frequently, while inventory and product data may change less often.

**Step 3: Stored Procedures to Process Data**

Three **stored procedures** are created in Snowflake to process the data from each of the source tables and push the updates to the **product\_performance** table:

* **process\_sales\_data**: This procedure is responsible for processing and updating the sales data.

A screenshot of a computer program

Description automatically generated

* **process\_inventory\_data**: This one handles updates related to inventory data.

A screenshot of a computer program

Description automatically generated

* **process\_product\_data**: This procedure deals with product updates like pricing changes.

A screenshot of a computer

Description automatically generated

**Step 4: Scheduling with Airflow (Future Scope)**

To make sure that the data is updated regularly, the system uses **Airflow** to schedule and automate the execution of these stored procedures:

* The **process\_sales\_data** procedure is scheduled to run **every hour**, ensuring that the latest sales data is always up to date in the product\_performance table.
* The **process\_inventory\_data** procedure runs **every 4 hours** since inventory data doesn't need to be updated as frequently as sales data.
* The **process\_product\_data** procedure is scheduled to run **every 8 hours**, as product information, like pricing, typically changes less often.

**Step 5: Updating the Target Table**

* The **product\_performance** table is the final destination for all the processed data. Each time a procedure runs, the data from the corresponding source table (Sales, Inventory, or Product) is processed and the central **product\_performance** table is updated. This table combines all the data into one place, providing a single view of how products are performing based on sales, inventory, and pricing information.

**A screenshot of a computer

Description automatically generated**