**Evaluation Questions**

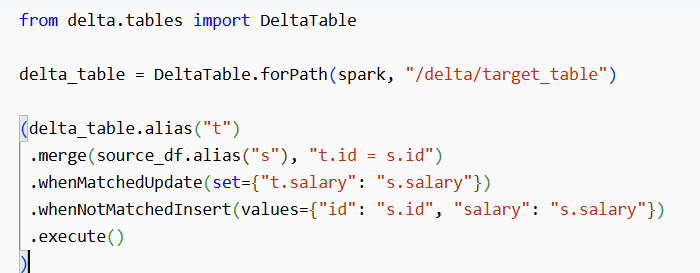
1. **Replacewhere – code**

The replaceWhere option in PySpark is used with Delta tables to dynamically overwrite specific partitions instead of the entire table.

**A white background with colorful text

AI-generated content may be incorrect.**

**2. Merge – code**



A close-up of a white background

AI-generated content may be incorrect.**3.Unpivot – code**

**4. GDAI process**

GDP is the centre for strong all the data from all domains. Now all the data are migrating to GDAI platform where unity catalog is enabled.

we are using 2 frameworks – MDMF and E- prep

**MDMF-Metadata-Driven Model Framework (source to bronze)**

MDMF consists of three main components

1. Data Ingestion
2. Data Standardization
3. Data Segregation

**Data Ingestion:**

This involves ingesting the data into the system. Getting the data from the source system

**Data Standardization:**

This step includes performing basic transformations, excluding business logic.

**Data Segregation:**

This involves controlling access to data

**Confidential:**

Highly sensitive data that cannot be accessed by everyone

Even though you (employee) are a part of the organisation, only limited persons (higher officials) with permitted access will be able to access the sensitive information.

**E-prep (bronze to gold):**

Bronze layerdata cleanse and transformed in this framework using python**.**

**5.Airflow**

**Airflow**

* Apache Airflow is a workflow automation and scheduling system.
* It is used to author, schedule, monitor ETL jobs and manage data pipelines.
* The rich user interface makes it easy to visualize pipelines running in production, monitor progress, and troubleshoot issues when needed.
* Apache Airflow is a workflow engine that will easily schedule and run your complex data pipelines. It will make sure that each task of your data pipeline will get executed in the correct order and each task gets the required resources

**DAG**

* DAG are a collection of tasks where all the tasks are connected via directed lines and DAG files containing Python code.
* A DAG consists of operators and operator defines an individual task that needs to be performed.
* The location of these DAG files is specified in the Airflow configuration file, but they need to be accessible by the Web Server, Scheduler, and Workers.

**Task**

* A task represents a single unit of work in a DAG. Tasks can be data transformations, API calls, or script executions.

**Scheduler**

* Airflow schedulerschedules the contained tasks, monitors the task execution and all DAGs, and then triggers the task instances
* It is responsible for scheduling and execution of DAGs. It retrieves and updates the status of the task in the database.
* It monitors and stays in sync (heartbeats) with a DAG folder for all DAG objects.
* The scheduler also has an internal component called Executor. The executor is responsible for spinning up workers and executing the task to completion.

**Operator**

* Operators determine what actually executes when your DAG runs.
* Operators allow for generation of certain types of tasks that become nodes in the DAG when instantiated.

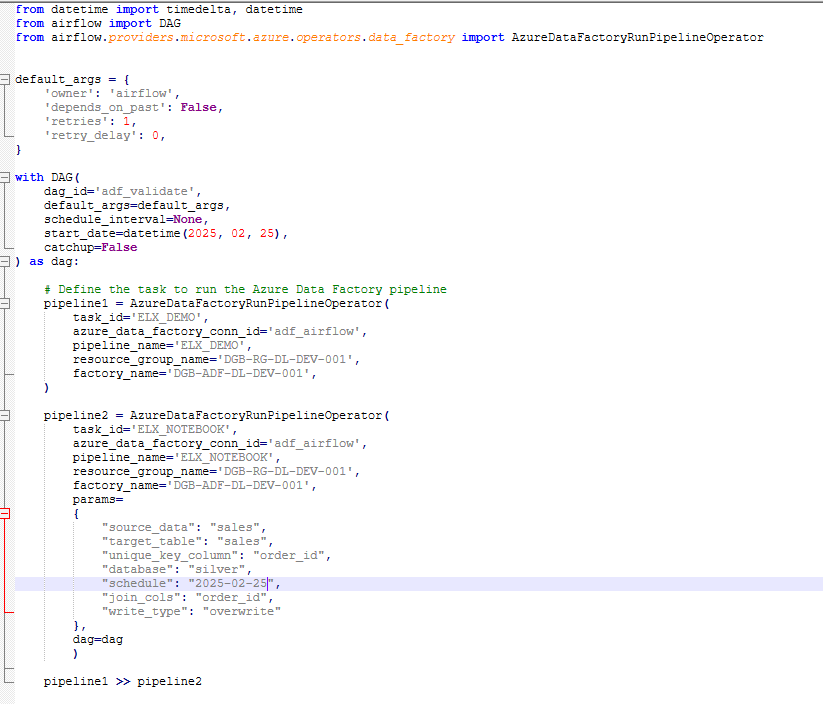
*Example: Bash Operator, Python Operator, MySql Operator, Postgres Operator, MsSql Operator, Oracle Operator*

**Start date and end date**

* DAGs need a start date and an optional end date to define execution periods.

**Catch up true and false**

* Catchup = False ensures that Airflow does not execute past DAG runs if it was paused or not running



**How to convert UTC to CET and CET to IST**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**Dag Dependency using external sensor**

* **DAG1:** Loads data from **source to bronze layer**
* **DAG2:** Moves data from **Bronze to Silver layer**
* **ExternalSensor:** Ensures **DAG2 waits for DAG1** to complete

A screen shot of a computer program

AI-generated content may be incorrect.

**6. Fact and Dimension table**

* Use 2 dim and 1 fact table, join and perform some transformation and KPI



**KPI**

A screenshot of a computer program

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

**7**. **Read excel file without using panda's**

In Databricks, you can read an Excel file without using Pandas by using Spark’s built-in method.

A screenshot of a computer code

AI-generated content may be incorrect.