**Project Assignment: Data Pipeline for Customer Account Analysis**

Resource Group: Resource23

Storage Account: microdata23

Containers: mycontainer

bronze

Azure Data Factory: microdfactory

Data Bricks: microdatabricks

Cluster: Ramoju

Notebook 1: Incremental

Notebook 2: ETL

**Step 1: Data Ingestion (Backend Storage to Raw (Bronze) Container)**

1. Create Containers and Upload Files:
   1. Go to the Azure Storage Account (**microdata23**) and create two containers: **mycontainer** and **bronze**.
   2. Upload the 5 CSV files into **mycontainer** and leave the **bronze** container empty.
2. Configure Source in Azure Data Factory:
   1. In Azure Data Factory (**microdfactory**), select the Copy Data activity.
   2. Create a new Azure Data Lake Storage Gen2 linked service **( ls\_source**) with a Wildcard File Path. Set the file name pattern to **\*.csv**.
3. Configure Sink in Azure Data Factory:
   1. In the Sink tab, create a new Azure Data Lake Storage Gen2 linked service (**ls\_sink**) and set the file path to the **bronze** container for storing the data.
4. Copy Activity: Transfer data from the source to the sink and verify the files in the **bronze** container.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 2: Databricks Activity (Incremental/Delta Processing)**

1. Databricks Notebook: Create a Databricks notebook as **Incremental**
2. Read Data: Read data from bronze container.
3. Data Cleaning: Implement logic to identify and remove any hanging or irrelevant data from the sources.
4. Data Transformation: Apply necessary transformations to prepare the data for further processing. This might involve schema changes, data type conversions, or handling missing values.
5. Save the cleaned data in another container as Silver.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 3: Databricks Activity (ETL Processing)**

1. Databricks Notebook: Create a Databricks notebook as **ETL**
2. Data Source: Read the data from the **accounts** and **customers** files in the **silver** container.
3. Transformation: Implement transformations to clean the data, removing irrelevant or missing entries.

* Write a SQL query to calculate the total balance across all accounts for each customer, ensuring all columns from both accounts and customers tables are included.

1. Reference:   
   <https://github.com/midha-abhishek/CustomerAccountLoanDB/blob/main/aggregate_and_insights.sql#L3>

* Data Loading: Store the transformed data in the **gold** container after processing.

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 4: Azure Synapse Analytics**

1. External Tables: Create an external table in Azure Synapse Analytics to map to data stored in the silver and gold containers of your data lake.