**My Roles and Responsibilities:**

I analyzed raw data files (CSV) from the backend Cloud Storage bucket for inconsistencies and null values.

I defined and implemented explicit schemas to avoid issues with automatic schema inference in PySpark.

I developed PySpark jobs on Dataproc to clean data, including removing irrelevant or duplicate records.

I handled datatype conversions and ensured null values were appropriately treated for accurate processing.

I transformed and wrote cleaned data into Delta formats in the curated/silver bucket.

I designed ETL pipelines to join customer and account datasets, computing total balances per customer.

I implemented aggregation logic and saved results to the refined/gold bucket in a structured format.

I created and managed BigQuery tables to store the final cleaned and aggregated data.

I scripted daily upsert operations using BigQuery MERGE statements to update customer account balances.

I ensured smooth integration of Cloud Scheduler and Dataproc jobs for end-to-end automation of ETL processes.