Customer360 Data Pipeline: From Raw to Insights

Project Overview

The **Customer360 Data Pipeline** is an end-to-end ETL solution built on **PySpark** that processes customer-related data from raw ingestion to business-ready insights. The pipeline extracts data from multiple sources, applies necessary transformations in three distinct layers (Bronze, Silver, and Gold), and stores the transformed data into parquet for analysis.

This solution is designed to:

- **Ingest raw data** from CSV files.
- **Clean, enrich, and validate data** in the Silver layer.
- **Aggregate and curate data** for business insights in the Gold layer.

Architecture

The pipeline follows the **Lakehouse** architecture pattern, which consists of three layers:

1. **Bronze Layer (Raw Data)**:
    - Ingest raw customer transaction and profile data into **Google Cloud Storage (GCS)**.
2. **Silver Layer (Transformed Data)**:
    - Use **Apache Beam** and **Google Cloud Dataflow** to clean, enrich, and validate the data. This layer processes and prepares data for analytics.
3. **Gold Layer (Aggregated Insights)**:
    - Transform data into analytical views in **BigQuery**, providing aggregated and business-ready datasets.

Input Datasets

1. **Customer Transactions (customer_transactions.csv)**:
    - Fields: transaction_id, customer_id, transaction_date, amount, category

Dataset link:
https://docs.google.com/spreadsheets/d/1U5esDcKDWzbEothjo9RXPmK5dcaYTgyc6DOPceSjLGk/edit?usp=sharing

2. **Customer Profiles (customer_profiles.csv)**:
    - Fields: customer_id, name, email, signup_date

Dataset link:

https://docs.google.com/spreadsheets/d/1Yk8YLIf87tHvKaz5D5O9N6JoUZZPZqM_wnjih_3EP_0/edit?usp=sharing

---

ETL Pipeline Stages

1. Bronze Layer (Raw Data Storage)

- **Purpose**: Ingest raw CSV files into **Google Cloud Storage**.
- Use below location:
    - Bucket name: customer360
    - File_path: bronze_layer/<date>/<dataset_name>/<file_nm>

2. Silver Layer (Data Transformation)

- **Purpose**: Clean, validate, and enrich data using PySpark.
- **Transformations**:
    - Remove duplicate - [for each date we should have a unique id row]
    - Clean invalid or missing values.
    - Enrich with additional calculated fields such as:
        - is_large_transaction for transactions > $1000.
        - year, month, day for easier aggregation.
        - Standardize category and email values.
            - Category remove special characters.
            - While validating email, consider email domain ends with ".com" and ".net".
    - Write clean data to **BigQuery**.
- **Python Script:[Pyspark|Dataflow]**
  The script performs the following:
    - Reads raw data from GCS.
    - Applies transformations: validation, cleaning, and enrichment.
    - Removes duplicates.
    - Writes transformed data into BigQuery.

- Key transformations include:
    - amount validation (e.g., values exceeding $1M or less than 0 are discarded).
    - Standardization of category and email.
    - Derivation of fields like year, month, and is_large_transaction.

3. Gold Layer (Aggregated Data)

- **Purpose**: Perform business-level aggregations and generate curated datasets for analysis.
- **Aggregations**:
    - **Total Spending Per Customer**: Aggregates transaction amounts by customer.
    - **Top Categories by Spending**: Identifies the most popular categories based on spending.
    - **Customer Profile with Spending**: Merges profile and transaction data to summarize customer spending.

Orchestration with Cloud Composer

The entire ETL pipeline is orchestrated using **Cloud Composer**, which manages task dependencies and ensures the smooth execution of the pipeline. The DAG (Directed Acyclic Graph) defined in Cloud Composer triggers the various tasks in sequence:

1. **Ingest Data to Bronze Layer**: Upload CSV files to GCS.
2. **Run Dataflow Job**: Execute the Dataflow job to process and clean data into the Silver layer.
3. **Execute BigQuery SQL**: Perform aggregation queries to produce Gold layer datasets.

Technology Stack

- **Google Cloud Platform (GCP)**
    - **Google Cloud Storage (GCS)**: Storage for raw CSV files (Bronze layer).

    - **Google Cloud Dataflow (Apache Beam)**: Data transformation and processing (Silver layer).
    - **BigQuery**: Data storage and querying for transformed and aggregated datasets (Silver and Gold layers).
    - **Cloud Composer**: Orchestration of ETL jobs and dependencies.

Key Benefits

- **Scalability**: The solution scales with growing data, leveraging GCP's serverless services.
- **Data Quality**: Multiple transformations ensure data is clean, enriched, and ready for analysis.
- **Flexibility**: The pipeline supports different types of input data and transformation rules.
- **Real-time**: The solution can be modified to support real-time streaming data as well.

Conclusion

The **Customer360 Data Pipeline** is a robust, scalable, and flexible ETL solution that provides clean, aggregated, and business-ready insights. This architecture ensures that raw data from multiple sources is transformed into a valuable resource for analytics, enabling data-driven decision-making across the organization.

Let me know if you'd like to add more details or any changes to this documentation!