**GOOGLE CLOUD BATCH & STREAMING DATA PIPELINES - DOCUMENTATION**

**Overview**

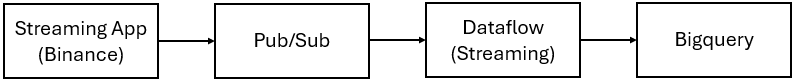
This document outlines end-to-end batch and streaming data pipelines implemented on Google Cloud Platform (GCP) using Apache Beam, Dataflow, BigQuery, Cloud Storage, Pub/Sub, and Cloud Composer (Airflow) for orchestration.

**STREAMING DATA PIPELINE:**

**Components Used**

1. **Binance API**: Real-time crypto data source.
2. **Pub/Sub**: Ingests real-time streaming data.
3. **Apache Beam**: Python SDK for writing streaming pipeline logic.
4. **Dataflow**: Executes the streaming Beam job.
5. **BigQuery**: Stores transformed real-time data.

**Architecture Diagram:**



**Streaming Flow Explanation**

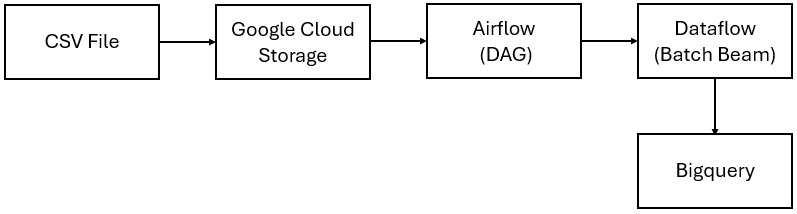
1. **Data Ingestion**:
   * A Cloud Function or custom script continuously fetches real-time crypto trade data from Binance API
   * The raw trade messages are pushed to a Pub/Sub topic.
2. **Apache Beam Pipeline**:
   * Subscribes to the Pub/Sub topic using ReadFromPubSub().
   * The message is a JSON payload containing event time, trade symbol, price, and quantity.
   * A DoFn class (ParseBinanceMessage) extracts and transforms this data.
3. **Google Dataflow**:
   * Runs the pipeline in streaming mode to ensure real-time processing and delivery to BigQuery.
4. **BigQuery**:
   * The transformed data is stored in a table for dashboards, alerts, or downstream analytics.

**BATCH DATA PIPELINE:**

**Components Used:**

1. **Cloud Storage (GCS)**: Stores the input CSV file.
2. **Apache Beam**: Python SDK for writing the batch data pipeline logic.
3. **Dataflow**: Managed service to execute Apache Beam pipelines.
4. **BigQuery**: Data warehouse for storing processed data.
5. **Cloud Composer (Airflow)**: Schedules and triggers the batch pipeline execution.

**ARCHITECTURE DIAGRAM:**



**Batch Flow Explanation**

1. **Data Ingestion**:
   * A CSV file is uploaded to a GCS bucket.
2. **Airflow (Cloud Composer)**:
   * A DAG is defined.
   * This DAG uses a BashOperator to run the Beam pipeline script using the DataflowRunner.
3. **Apache Beam Pipeline**:
   * Reads CSV from GCS.
   * Parses and formats the data.
   * Writes the output to a BigQuery table with schema enforcement.
4. **Google Dataflow**:
   * Executes the Beam job in a managed environment with autoscaling.
5. **BigQuery**:
   * Stores the processed data for analytics or downstream use cases.

**GIT Repo Link:**

<https://github.com/gajendra8892/Dataflow>