Project Goal

The project aims to create a streaming pipeline in the Google Cloud Platform from Twitter API and create an interactive dashboard in Data Studio with live data.

Data Source

https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview

Objectives

To build a data pipeline to stream the tweets in real-time that:

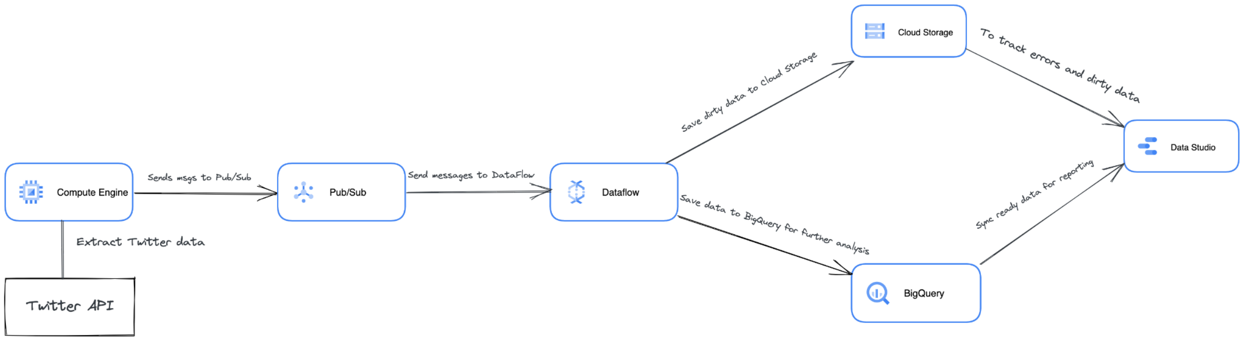
1. Stores raw tweets in the data warehouse (e.g., Google BigQuery) that can be used for analytics purposes.
2. Run some analytics and get insights from tweets data by creating a dashboard in Google Data Studio.

Tools

1. **Compute Engine:** IaaS that offers high-performance virtual machines.
2. **Cloud Pub/Sub:** messaging bugger that allows you to send and receive messages between different applications.
3. **Google Dataflow:** A fully managed service that allows you to transform steaming and batch data by using Apache Beam.
4. **Google BigQuery:** A fully managed data warehouse for analytics.
5. **Google Data Studio:** reporting tool used for creating interactive dashboards.

Architecture

This is a high-level diagram that shows how the pipeline should work:


Cloud Storage
To track errors and dirty data
Save dirty data to cloud Storage
Data Studio
Compute Engine
Sends msgs to Pub/Sub
Pub/Sub
Send messages to DataFlow
Dataflow
Save data to BigQuery for further analysis
Sync ready data for reporting
Extract Twitter data
Twitter API
BigQuery

Flow

1. The data from Twitter API will be extracted by using Compute Engine virtual machine.
2. Then, the received message will be sent to Cloud Pub/Sub.
3. Cloud Pub/Sub sends a message to Cloud Dataflow to run some transformations on the data and make sure the data is clean and appropriate.
4. Dirty data will be sent to Cloud Storage for further analysis.
5. Relevant data will be saved in Google BigQuery.
6. Finally, we create a dashboard to show some statistics based on the data.