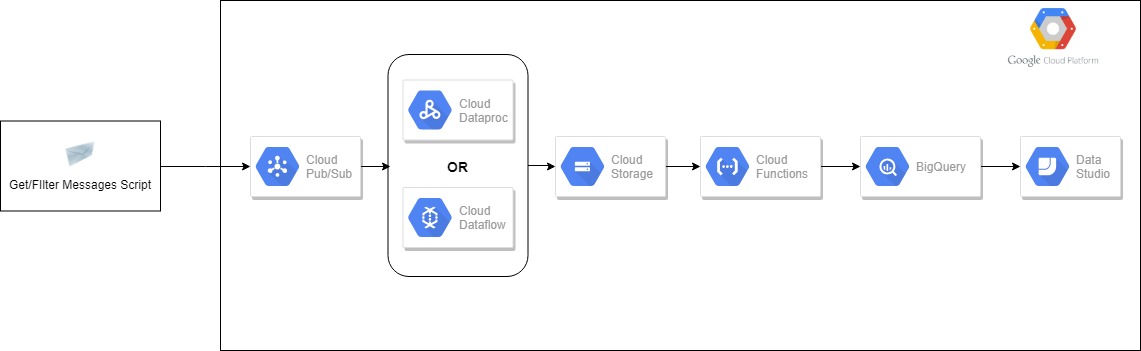
Task 1 - Documentation

The pipeline was design on GCP, because its ecosystem makes easier to manipulate data. From ingestion to Visualization there are options to get a data project done. This pipeline also was built seeking been efficient with time and cost. Once it been said, follow below the pipeline diagram:



Task 2 Extract

For this task I have used a Python script to consume from Wikipedia API(SCE) and to send this data to Pub/Sub. Once the events are in Pub/Sub I have chose the simplest solution - a predefined Dataflow job, once the scope of the problem doesn’t mention any necessary transformation. This ready-to-use Dataflow pipeline, reads from an topic on Pub/Sub and writes files in a 5 minutes window. Also it is possible to consume from Pub/Sub and writing to Storage with Dataproc.

I could use Dataflow and Dataproc also to streaming data directly to Big Query, but BQ has an extra fee about streaming data in its tables. Once there is not an short SLA, I have found this extra cost not worthy at this moment.

Task 3 - Storage

I have used Big Query as my storage database and to insert data that comes from Storage, I chose Cloud Function. Once the file gets on Storage it sends a request to Big Query to load this file. In order to achieve one of the requirements of this task, that is about to keep the size down, I have proposed to have daily tables that expires after a configurable number of days.

Task 4 - Visualization

IN order to accomplish this task I have used Data Studio.

Task 5 - Query