Link to the GitHub repo: https://github.com/andrianhevalo/Capstone-Project

To estimate the project's total cost, let's first list all GCP tools that I've used:

1. *Google Cloud Storage*
2. *Google Pub/Sub*
3. *Google Dataflow*
4. *Google BigQuery*

Google Cloud Storage

I used the standard storage type located in Iowa (central1), which costs **0.02$** per GB per month. Since I use GCS only as data storage, I don't pay for any data processing (e.g., data retrieval).

Google Pub/Sub

According to Pub/Sub official pricing policy (link: https://cloud.google.com/pubsub/pricing):

The cost of Pub/Sub has three components:

- Throughput costs for message publishing and delivery
- Egress costs associated with throughput that crosses a Google Cloud zone or region boundary
- Storage costs for snapshots, messages retained by topics, and acknowledged messages retained by subscriptions

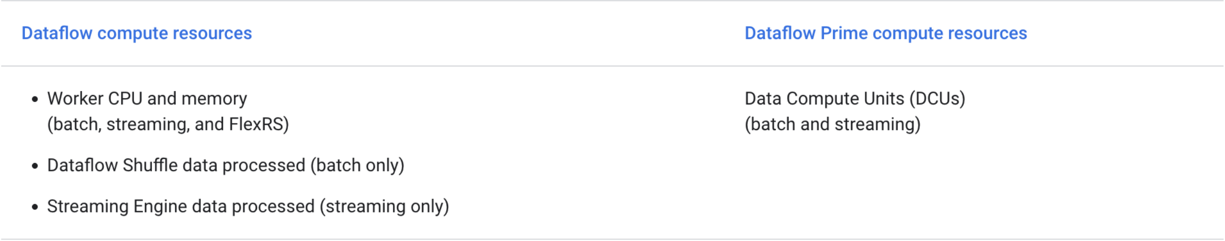
Let's investigate each one of them:

1. Throughput costs: one message generated by our script is 1 MB in size on average. Documentations say that Every calendar month, the first 10 GiB of throughput identified as the **Message Delivery Basic** SKU for a billing account is free. After that, the price is **$40 per TiB** in all Google Cloud regions. However, if you are using BigQuery subscriptions, read the next section. Our API produces approx 1440 messages per day. The daily throughput is 1440 * 1 MB = 1440 MB. Hence, the total throughput cost is approx (40$/1000000) * 1440 = **0.05$.**
2. Egress costs: we don't fall into that cost since the project is located only in one region.
3. Storage costs: the same as previous (we don't use any snapshots to store data or similar).

Cloud Dataflow

From official pricing: https://cloud.google.com/dataflow/pricing

Dataflow usage is billed for resources that your jobs use. Resources are measured and billed differently depending on whether you're using Dataflow or Dataflow Prime.



Dataflow compute resources	Dataflow Prime compute resources
• Worker CPU and memory (batch, streaming, and FlexRS) • Dataflow Shuffle data processed (batch only) • Streaming Engine data processed (streaming only)	Data Compute Units (DCUs) (batch and streaming)

Let's use GCP pricing:
https://cloud.google.com/products/calculator#id=786a882d-8442-4e4f-802b-868757daab0a

Based on our parameters, the total cost per month is **64$.**

Google BigQuery

1. Storage: we are using the Active storage option, which costs **$0.02** per GB. Hence, the monthly cost for BigQuery will be (0.02/1000) * 1440 MB = **$0.028**.
2. Data Querying: we don't apply any queries as of now, so the total cost is **0$**.

Conclusion

Let's add up all expenses: Cloud Storage + Cloud Pub/Sub + Cloud Dataflow + Google BigQuery = 0.02 + 0.05 + 64 + 0.028 = **$64.1.**

What is the most expensive part of the project?

Google Dataflow. Considering how much resources this tool consumes (e.f. Google Compute Engine), I think it's obvious. Also, we are using the Streaming mode, which increases the cost.

Is there a way to reduce a bill?

Yes, there are plenty of ways to reduce the bill:

1. Use BigQuery subscription to Pub/Sub (without Dataflow as a bridge). This aims to reduce the cost significantly since we wouldn't be using Dataflow for ETL. Link: https://cloud.google.com/blog/products/data-analytics/pub-sub-launches-direct-path-to-bigquery-for-streaming-analytics
2. Use Batch mode in Dataflow instead of Streaming. Batch mode is way cheaper than streaming.

If the price changes linearly/logarithmically/exponentially, what causes this change?

1. Messages started to come more often to Pub/Sub, leading to increasing storage and throughput costs.
2. Users are running queries on data (each query is billed - $5 per TB).