**Real Time Ticket Booking data Sync to Google Big Query**

Google Pub Sub

* Read Pub-Sub
* Apply python UDF for transformation.
* Fault tolerance

IRCTC-DATA

BigQuery

Data Flow

Publish

Backend it uses Apache Beam

Python Mock Data

Register UDF

Generator

GCS Bucket

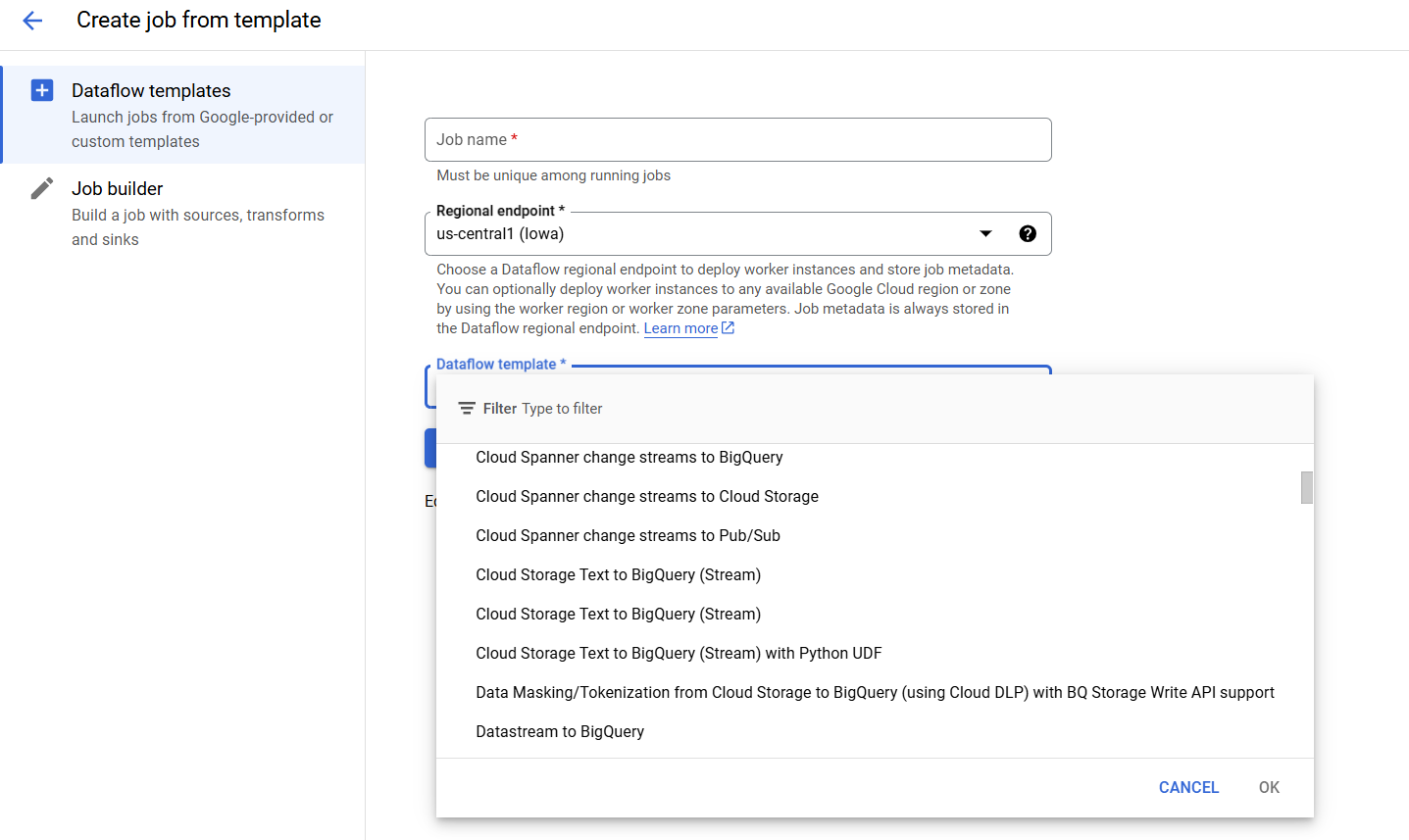
Transform-UDF.py

* In the pub-sub just create a topic, which has some check box which you can check-in if you want to enable it.

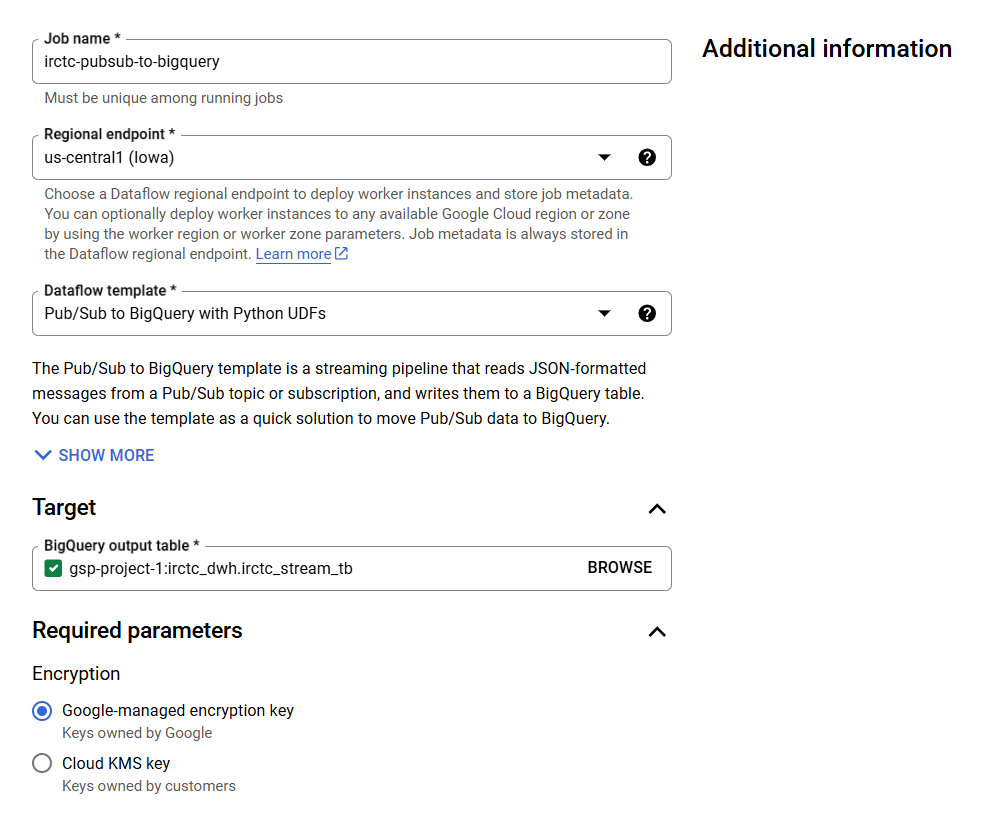
A screenshot of a computer

Description automatically generated

* Data Flow to move the data from pub-sub to Big Query.
* By default, Data Flow has a lot of pre-defined templates.



* We are going to use pub-sub to big query with python UDF.
* Create Dataset which is more of a database in BigQuery, then create the table.
* Choose that table as target.



* Since we are using pub/sub to bigquery with Python UDF’s we need to provide the transformation function in a bucket.

A screenshot of a computer

Description automatically generated

* After pointing to the py file.
* Then basic ram, workers all these configurations needs to be made.
* After all the configurations, click on create run.
* The workers responsibility is to make sure, if there are failed records, that needs to be stored in the bigquery table.
* You have to give the necessary permission for doing it.

A screenshot of a computer

Description automatically generated

**Python Library Required**

pip install google-cloud-pubsub google-cloud-bigquery

**BigQuery Output**

A screenshot of a computer

Description automatically generated

When you kill the data flow job, these are the 3 options it shows.

A screenshot of a computer

Description automatically generated

Cost Details in Pub/Sub

In **Google Cloud Pub/Sub**, a topic itself does **not incur costs** just by existing. Costs are incurred based on **usage**—when data is published, retained, or delivered to subscribers. Below is a breakdown of how Pub/Sub incurs costs:

**1. Pub/Sub Topic Itself**

* **Free**: A Pub/Sub topic sitting idle (with no data published to it) does not cost anything.
* Simply having a topic or subscription created does not spawn any machines or resources.

**2. Costs in Pub/Sub**

Costs are based on the following factors:

**a. Data Published to the Topic**

* When messages are published to the topic, you are charged based on the **amount of data sent** (measured in MB).
* **Cost for data publishing**: $40 per TiB ($0.04 per GB).

**b. Message Retention**

* If you configure message retention for a topic or subscription (beyond default delivery), there is a cost for storing retained messages.
* Default message retention:
  + **7 days** for acknowledged messages.
  + **7 days** for unacknowledged messages.
* **Cost for storage**: $0.27 per GiB per month for retained messages.

**c. Data Delivered to Subscribers**

* You are charged for delivering messages to subscribers.
* **Cost for data delivery**: $40 per TiB ($0.04 per GB).
* This applies whether subscribers use **Push** (HTTP) or **Pull** subscriptions.

**d. Cross-Region Data**

* If messages are sent across Google Cloud regions, cross-region network egress costs may apply.

**3. Cost-Free Scenarios**

If you:

* Have created a topic but are not publishing any messages.
* Have no active subscriptions or deliveries.
* Have not enabled any advanced features (like message retention or dead-letter topics).

Then, you incur **zero cost**.

**4. Advanced Features That Can Increase Costs**

* **Message Storage Retention**: Retaining messages for more than the default duration incurs storage costs.
* **Dead Letter Topics**: Retaining undelivered messages in a dead-letter topic incurs storage costs.
* **Snapshots**: Retained snapshots of subscriptions consume storage and incur costs.

**5. How to Monitor Costs**

Use Google Cloud's **Cost Management Tools**:

* Go to Cloud Billing Reports to see Pub/Sub-related charges.
* Use **Stackdriver Monitoring** or **Pub/Sub Metrics** to track message publishing, delivery, and storage.

**6. Summary**

* A topic by itself incurs **no costs**.
* Costs are based on **data published**, **messages retained**, and **data delivered to subscribers**.
* If the topic is idle, there is **no charge**.

Let me know if you'd like to calculate estimated costs for your specific Pub/Sub usage!