# Dataflow: Qwik Start – Templates

## Objective

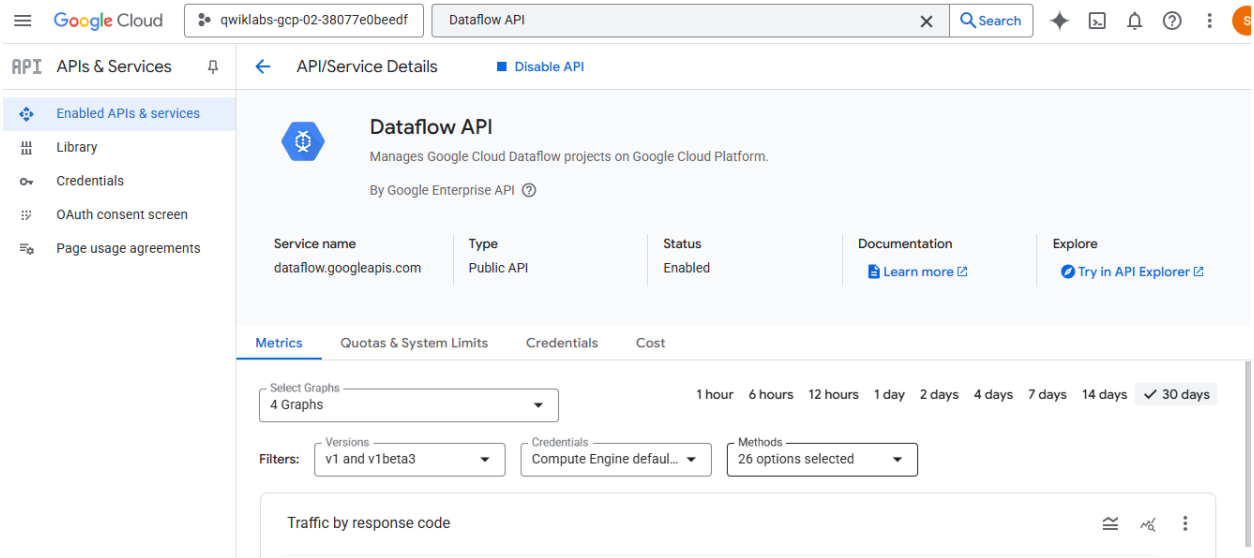
The goal of this project is to process real-time taxi ride data using Google Cloud's BigQuery, Cloud Storage, and Dataflow pipeline. The project involves enabling the Dataflow API, creating datasets and tables in BigQuery, configuring storage, running a Dataflow pipeline, and querying the processed data.

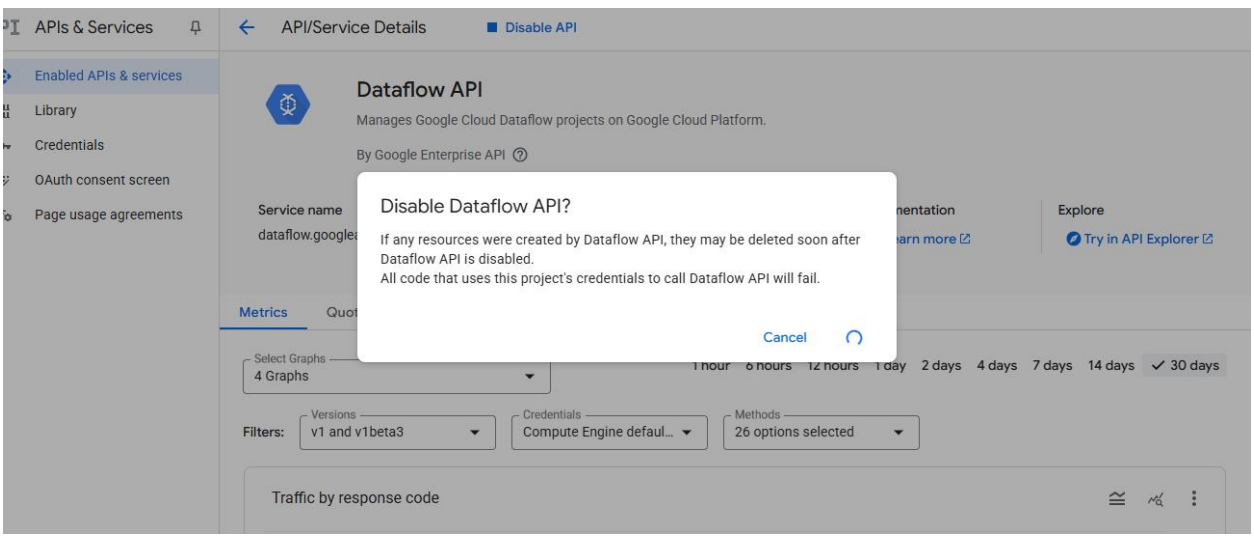
**Tools and Services Used:**

- **Google Cloud Console**
- **Cloud Shell**
- **BigQuery**
- **Cloud Storage**
- **Pub/Sub**
- **Dataflow**
- **bq command-line tool**
- **gsutil tool**

**Task 1: Re-enable Dataflow API**

- **Steps Taken**:
    - Located the **Dataflow API** in Cloud Console.
    - Disabled the API, then re-enabled it to ensure connectivity.
    - Verified the API status after re-enabling.




**Task 2: Create BigQuery Dataset, Table, and Cloud Storage Bucket (via Cloud Shell)**

- **Dataset Creation**:
  - Used bq mk taxirides to create the dataset.
- **Table Creation**:
  - Executed bq mk with appropriate schema definitions to instantiate taxirides.realtime.
- **Storage Bucket Configuration**:
  - Created a Cloud Storage bucket using gsutil mb gs://$BUCKET_NAME/.

```
CLOUD SHELL
Terminal (qwiklabs-gcp-02-38077e0beedf)
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to qwiklabs-gcp-02-38077e0beedf.
Use `gcloud config set project [PROJECT_ID]` to change to a different project.
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$ bq mk taxirides
Dataset 'qwiklabs-gcp-02-38077e0beedf:taxirides' successfully created.
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$
```

```
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$ bq mk \
--time_partitioning_field timestamp \
--schema ride_id:string,point_idx:integer,latitude:float,longitude:float,\
timestamp:timestamp,meter_reading:float,meter_increment:float,ride_status:string,\
passenger_count:integer -t taxirides.realtime
Table 'qwiklabs-gcp-02-38077e0beedf:taxirides.realtime' successfully created.
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$
```

```
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$ export BUCKET_NAME=qwiklabs-gcp-02-38077e0beedf
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$ gsutil mb gs://$BUCKET_NAME/
Creating gs://qwiklabs-gcp-02-38077e0beedf/...
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$
```

### Task 3: Create BigQuery Dataset, Table, and Cloud Storage Bucket (via Console)

- **Dataset Creation**:
  - Used the BigQuery UI to create the dataset taxirides.
- **Table Creation**:
  - Created the table realtime manually, defining schema fields.
- **Storage Bucket Configuration**:
  - Created a Cloud Storage bucket via the Cloud Console.

Create a Cloud Storage bucket using the Cloud console

1. Go back to the Cloud Console and navigate to **Cloud Storage** > **Buckets** > **Create bucket**.
2. Use the Project ID as the bucket name to ensure a globally unique name: <Bucket Name>
3. Leave all other default settings, then click **Create**.

## Create dataset

Project ID *
qwiklabs-gcp-02-38077e0beedf

Change

Dataset ID *
taxirides

Must contain only letters, numbers, or underscores.

Location type

- ○ Region
  Specify a region to colocate your datasets with other Google Cloud services.
- ◉ Multi-region
  Allow BigQuery to select a region within a group to achieve higher quota limits.

Some locations have been restricted due to a policy set by your organization. Learn more about restricting locations.

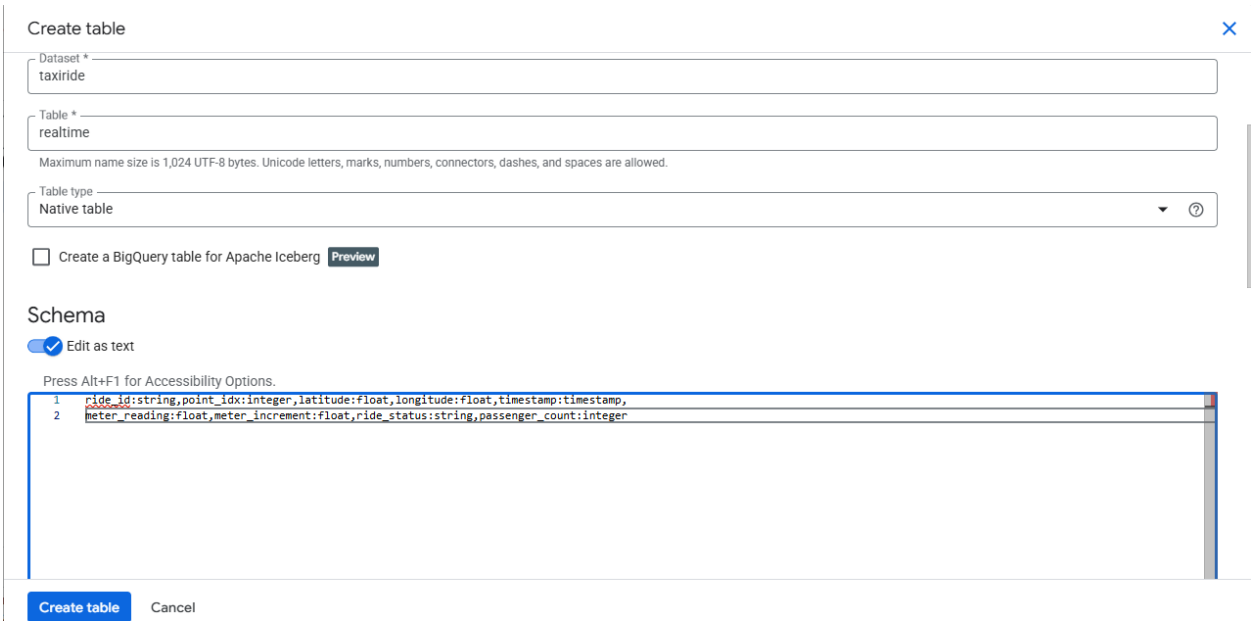
Multi-region *
US (multiple regions in United States)

### External Dataset

The selected region supports the following external dataset types: Cloud Spanner

☐ Link to an external dataset

Create dataset  Cancel




## Task 4. Run the pipeline

Deploy the Dataflow Template:

```
gcloud dataflow jobs run iotflow \
  --gcs-location gs://dataflow-templates-"Region"/latest/PubSub_to_BigQuery \
  --region "Region" \
  --worker-machine-type e2-medium \
  --staging-location gs://"Bucket Name"/temp \
  --parameters inputTopic=projects/pubsub-public-data/topics/taxirides-realtime,outputTableSpec="Table Name":taxirides.realtime
```

In the **Google Cloud Console**, on the **Navigation menu**, click **Dataflow > Jobs**, and you will see your dataflow job.

```
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$ gcloud dataflow jobs run iotflow \
    --gcs-location gs://dataflow-templates-us-central1/latest/PubSub_to_BigQuery \
    --region us-central1 \
    --worker-machine-type e2-medium \
    --staging-location gs://qwiklabs-gcp-02-38077e0beedf/temp \
    --parameters inputTopic=projects/pubsub-public-data/topics/taxirides-realtime,outputTableSpec=qwiklabs-gcp-02-38077e0beedf:taxirides.realtime
createTime: '2025-06-10T12:12:02.628041Z'
currentStateTime: '1970-01-01T00:00:00Z'
id: 2025-06-10_05_12_01-17326645662049072591
location: us-central1
name: iotflow
projectId: qwiklabs-gcp-02-38077e0beedf
startTime: '2025-06-10T12:12:02.628041Z'
type: JOB_TYPE_STREAMING
student_01_dada926e8de2@cloudshell:~ (qwiklabs-gcp-02-38077e0beedf)$
```

## Task 5. Submit a query

You can submit queries using standard SQL.

1. In the BigQuery **Editor**, add the following to query the data in your project:

SELECT * FROM `"Bucket Name".taxirides.realtime` LIMIT 1000

2. Now click **RUN**.

3. When the query runs successfully, you'll see the output in the **Query Results** panel as shown below:

Query results SAVE AS EXPLORE IN DATA STUDIO

Query complete (2.116 sec elapsed, 0 B processed)

Job information Results JSON Execution details

| Row | ride_id | point_idx | latitude | longitude | timestamp |
|---|---|---|---|---|---|
| 1 | b0810fbd-78a8-4159-b9ff-963695e2a23d | 225 | 40.753550000000004 | -73.98504000000001 | 2018-07-25 23:28:20.870530 UTC |
| 2 | 1a10dc8b-3623-41bf-938a-9fca26c2ae10 | 311 | 40.752930000000006 | -73.96584 | 2018-07-25 23:24:10.608380 UTC |
| 3 | 5253c100-1a30-4a3e-89ee-6c0c861cf44f | 224 | 40.74331 | -73.99172 | 2018-07-25 23:26:34.636480 UTC |
| 4 | 3efa96c2-4695-4c0b-96b6-da33a4b74ccf | 8 | 40.7533 | -73.97832000000001 | 2018-07-25 23:24:06.823150 UTC |
| 5 | d6d37615-ccba-4416-9932-e956e0f0ba65 | 747 | 40.682140000000004 | -74.00594000000001 | 2018-07-25 23:24:10.103770 UTC |