Dataflow: Qwik Start - Templates

GSP192

Overview

In this lab, you will learn how to create a streaming pipeline using one of Google's Cloud Dataflow templates. More specifically, you will use the Cloud Pub/Sub to BigQuery template, which reads messages written in JSON from a Pub/Sub topic and pushes them to a BigQuery table. You can find the documentation for this template here.
You'll be given the option to use the Cloud Shell command line or the Cloud Console to create the BigQuery dataset and table. **Pick one method to use**, then continue with that method for the rest of the lab. If you want experience using both methods, run through this lab a second time.

Setup

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This Qwiklabs hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

What you need

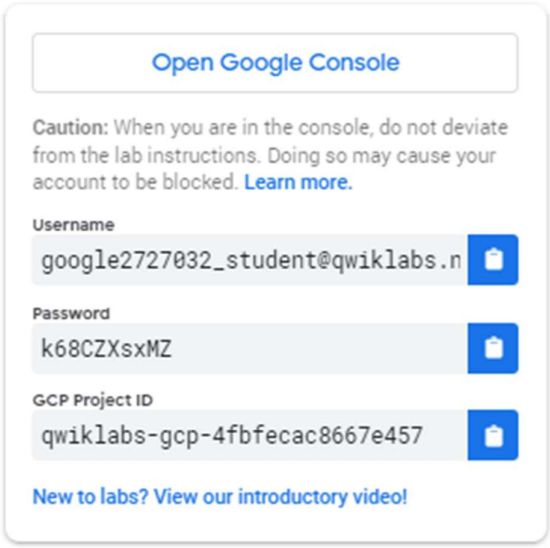
To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- Time to complete the lab.
  Note: If you already have your own personal Google Cloud account or project, do not use it for this lab.

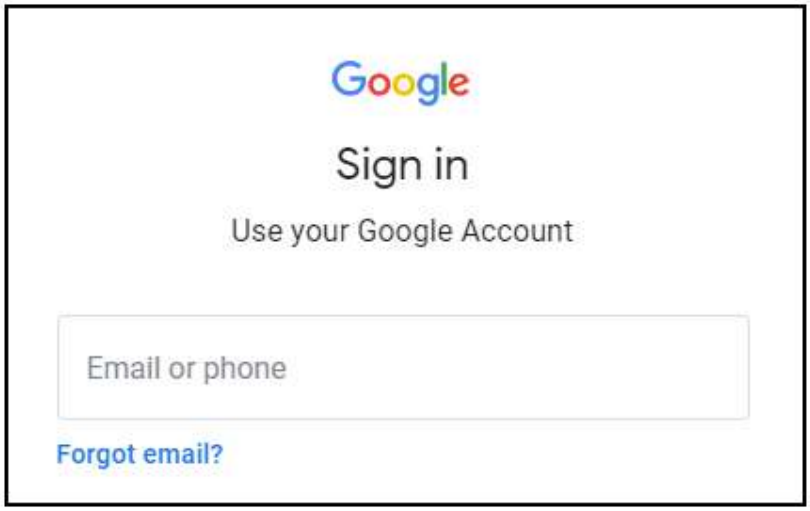
Note: If you are using a Pixelbook, open an Incognito window to run this lab.

How to start your lab and sign in to the Google Cloud Console

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is a panel populated with the temporary credentials that you must use for this lab.

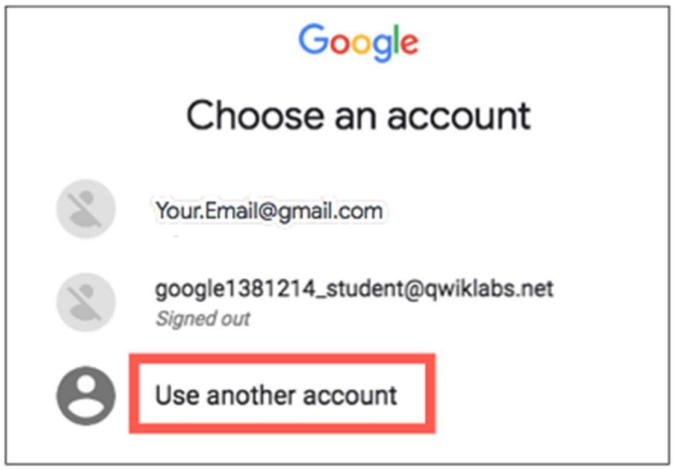


2. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.



Tip: Open the tabs in separate windows, side-by-side.

If you see the **Choose an account** page, click **Use Another Account**.



3. In the **Sign in** page, paste the username that you copied from the Connection Details panel. Then copy and paste the password.

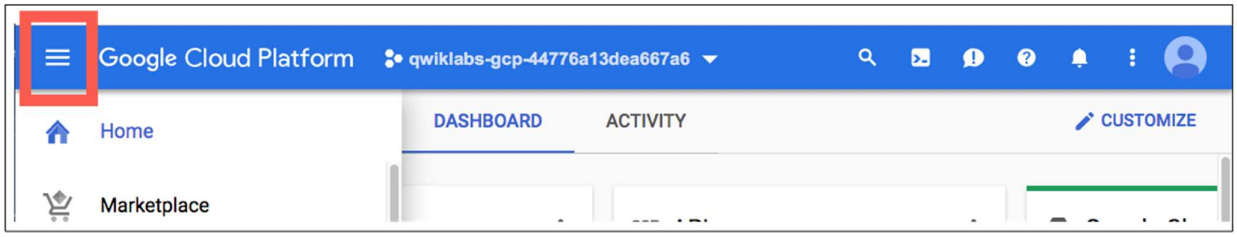
   Important: You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

4. Click through the subsequent pages:

   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

After a few moments, the Cloud Console opens in this tab.

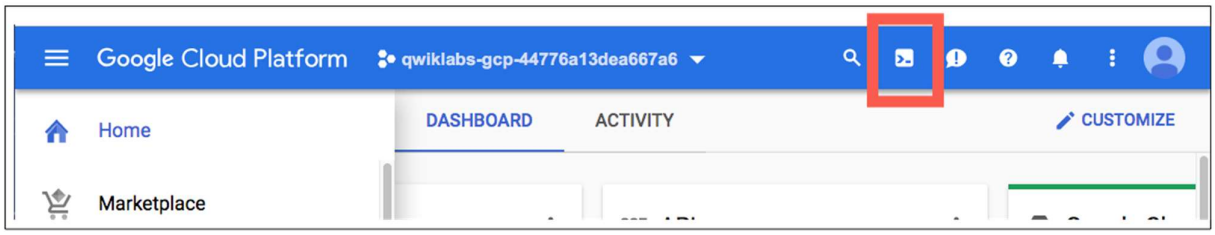
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.



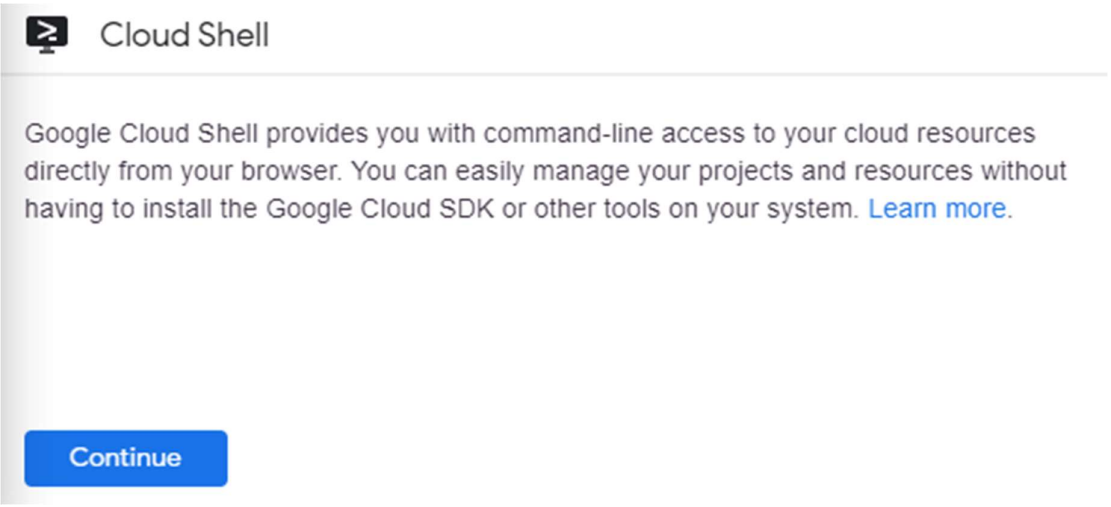
Activate Cloud Shell

Cloud Shell is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud. Cloud Shell provides command-line access to your Google Cloud resources.

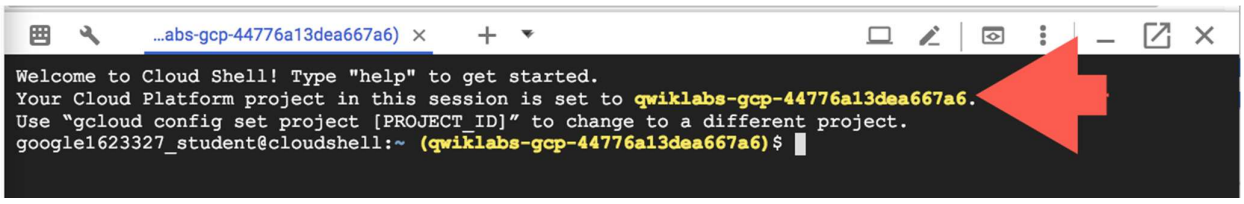
In the Cloud Console, in the top right toolbar, click the **Activate Cloud Shell** button.



Click **Continue**.



It takes a few moments to provision and connect to the environment. When you are connected, you are already authenticated, and the project is set to your *PROJECT_ID*. For example:



`gcloud` is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.
You can list the active account name with this command:

```
gcloud auth list
```

(Output)

```
Credentialed accounts:
 - <myaccount>@<mydomain>.com (active)
```

(Example output)

```
Credentialed accounts:
 - google1623327_student@qwiklabs.net
```

You can list the project ID with this command:

```
gcloud config list project
```

(Output)

```
[core]
project = <project_ID>
```

(Example output)

```
[core]
project = qwiklabs-gcp-44776a13dea667a6
```

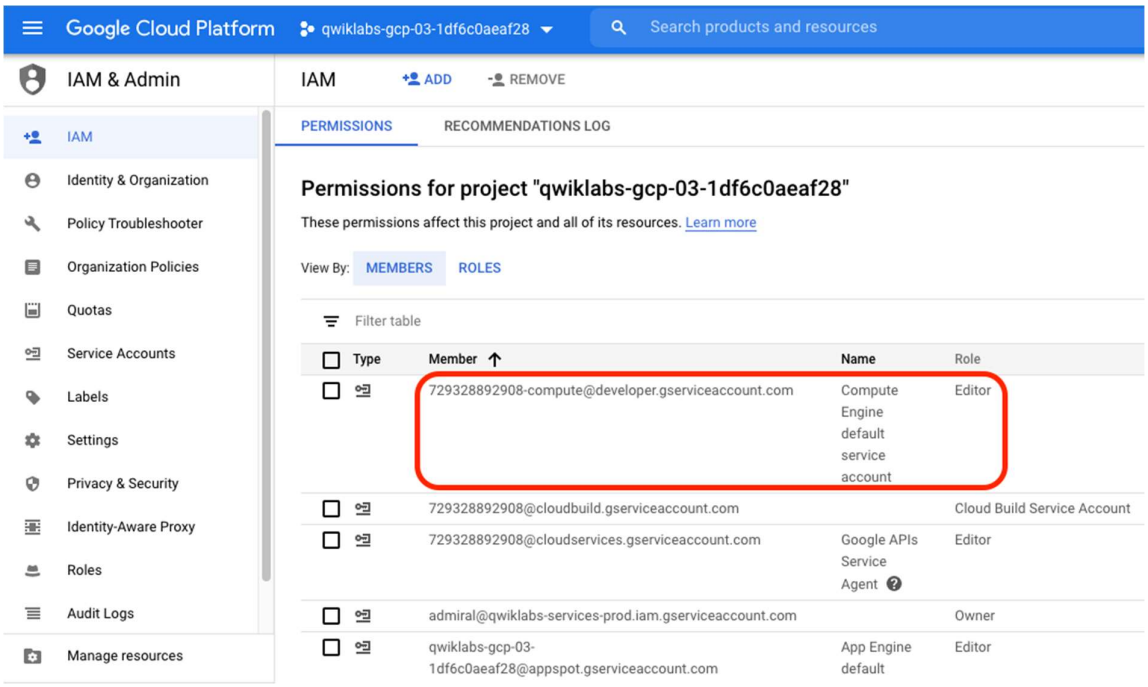
For full documentation of `gcloud` see the gcloud command-line tool overview.

Check project permissions

Before you begin your work on Google Cloud, you need to ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the **Navigation menu** ( ), click **IAM & Admin** > **IAM**.

2. Confirm that the default compute Service Account `{project-number}-compute@developer.gserviceaccount.com` is present and has the `editor` role assigned. The account prefix is the project number, which you can find on **Navigation menu** > **Home**.



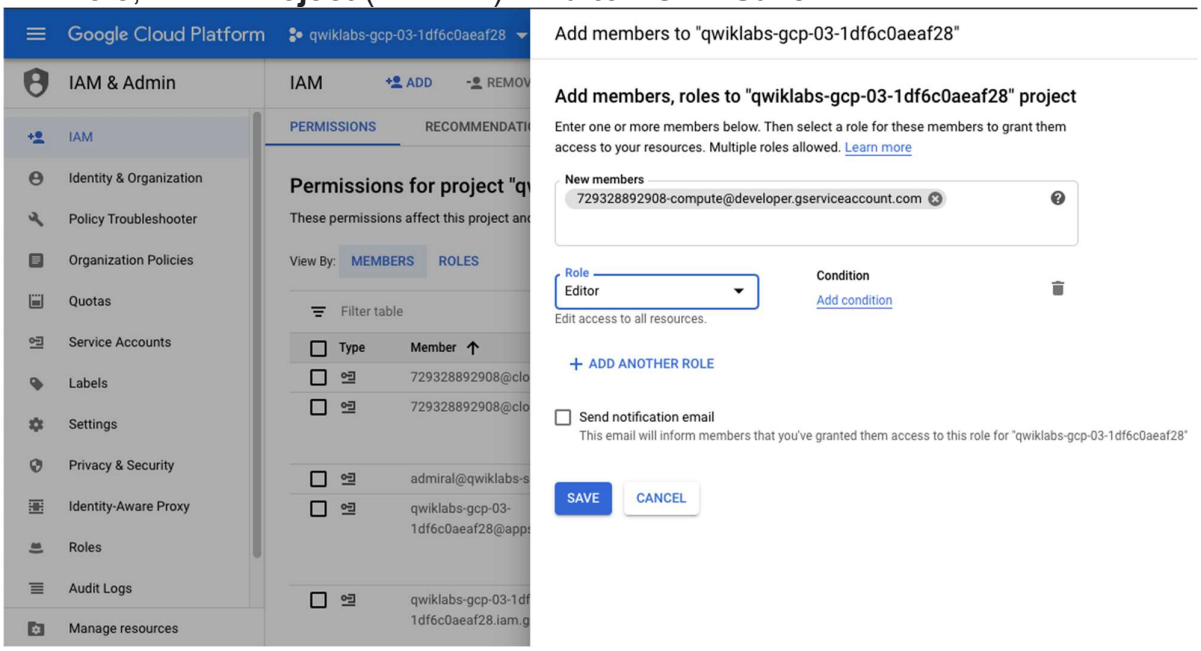
If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

- In the Google Cloud console, on the **Navigation menu**, click **Home**.
- Copy the project number (e.g. `729328892908`).
- On the **Navigation menu**, click **IAM & Admin** > **IAM**.
- At the top of the **IAM** page, click **Add**.
- For **New members**, type:

```
{project-number}-compute@developer.gserviceaccount.com
```

Replace `{project-number}` with your project number.

- For **Role**, select **Project** (or Basic) > **Editor**. Click **Save**.



Create a Cloud BigQuery Dataset and Table Using Cloud Shell

Let's first create a BigQuery dataset and table.

Note: This section uses the `bq` command-line tool. **Skip down** if you want to run through this lab using the console.
Run the following command to create a dataset called `taxirides`:

```
bq mk taxirides
```

Your output should look similar to:

```
Dataset '<myprojectid:taxirides>' successfully created
```

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created BigQuery dataset, you will see an assessment score.

Now that you have your dataset created, you'll use it in the following step to instantiate a BigQuery table. Run the following command to do so:

```
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
```

Your output should look similar to:

```
Table 'myprojectid:taxirides.realtime' successfully created
```

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created table in BigQuery dataset, you will see an assessment score.

On it's face, the `bq mk` command looks a bit complicated. However, with some assistance from the BigQuery command-line documentation, we can break down what's going on here. For example, the documentation tells us a little bit more about **schema**:

- Either the path to a local JSON schema file or a comma-separated list of column definitions in the form `[FIELD]:[DATA_TYPE], [FIELD]:[DATA_TYPE]`.
In this case, we are using the latter—a comma-separated list.

Create a storage bucket

Now that we have our table instantiated, let's create a bucket. Run the following commands to do so:

```
export BUCKET_NAME=<your-unique-name>

gsutil mb gs://$BUCKET_NAME/
```

Test Completed Task

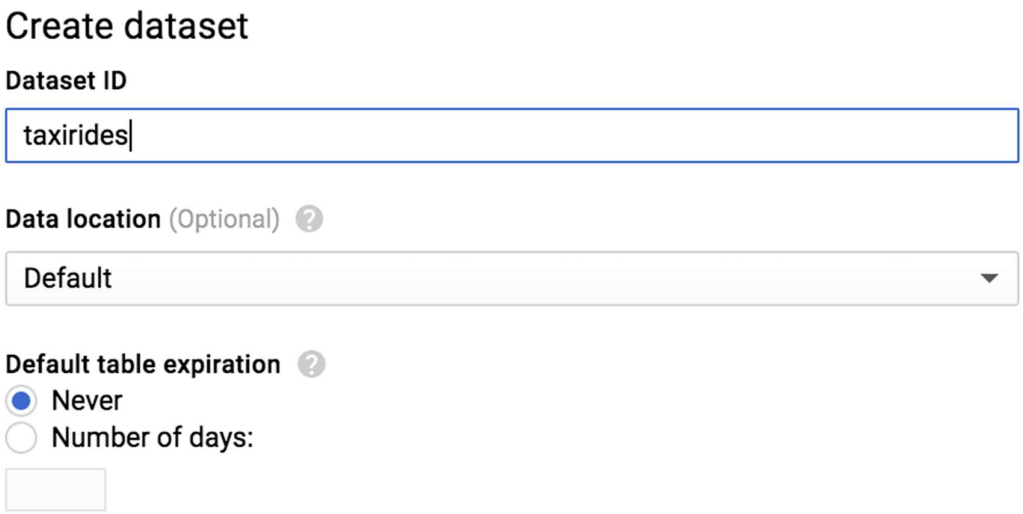
Click **Check my progress** to verify your performed task. If you have successfully created Cloud Storage bucket, you will see an assessment score.

Once you've made your bucket, **scroll down to the "Run the Pipeline" section** .

Create a Cloud BigQuery Dataset and Table Using the Cloud Console

Note: Don't go through this section if you've done the command-line setup!
From the left-hand menu, in the Big Data section, click on **BigQuery**. Then click **Done**.

Click on your project name in the left-hand navigation, then click **CREATE DATASET** on the right-hand side of the console. Input `taxirides` as your dataset ID:



Leave all of the other default settings in place and click **Create dataset**.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created BigQuery dataset, you will see an assessment score.

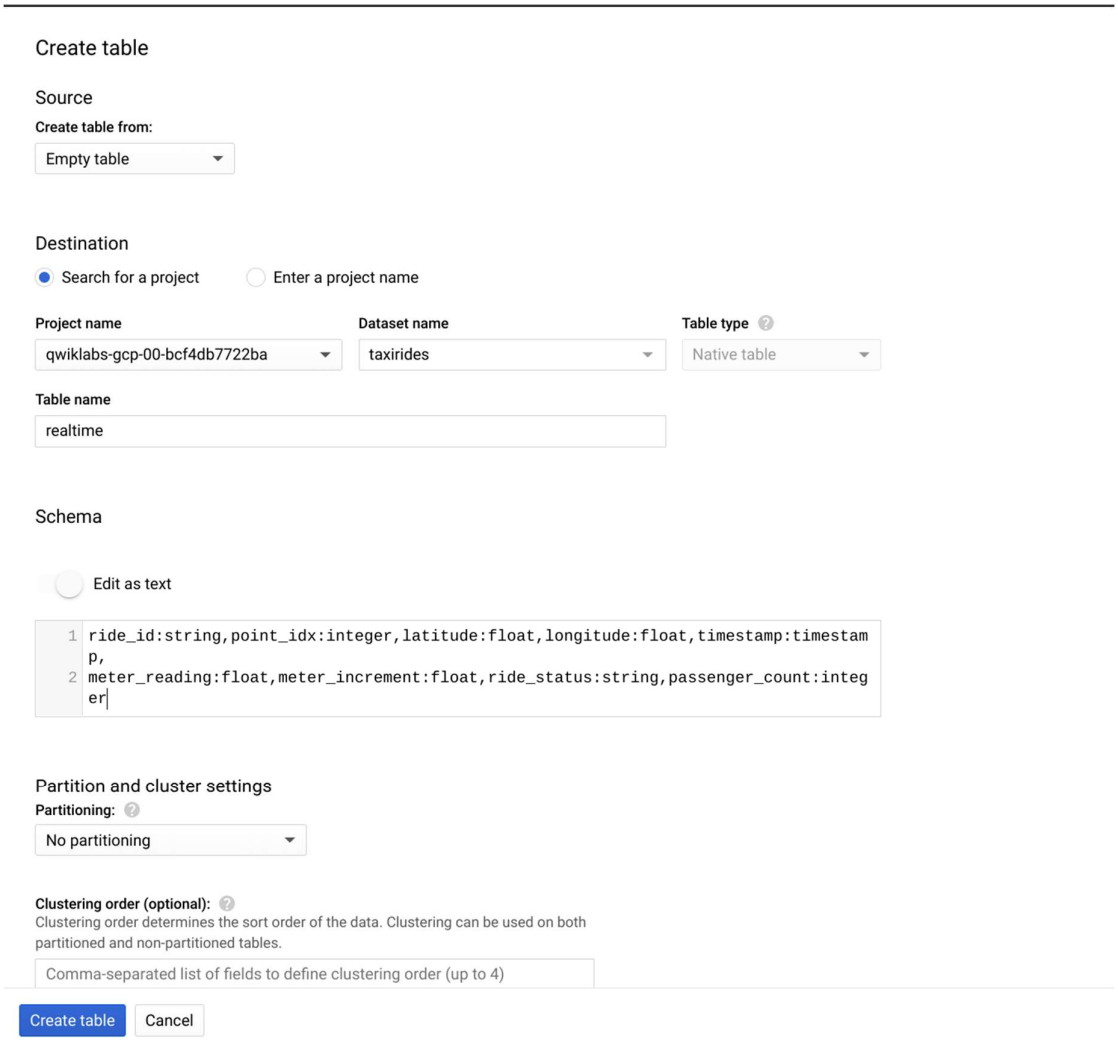
You should now see the taxirides dataset underneath your project ID in the left-hand console—click on it and then select **CREATE TABLE** in the right-hand side of the console.

In the **Destination** > **Table Name** input, enter `realtime`.

Under Schema, toggle the **Edit as text** slider and enter the following:

```
ride_id:string,point_idx:integer,latitude:float,longitude:float,timestamp:timestamp,
meter_reading:float,meter_increment:float,ride_status:string,passenger_count:integer
```

Your console should look like the following:



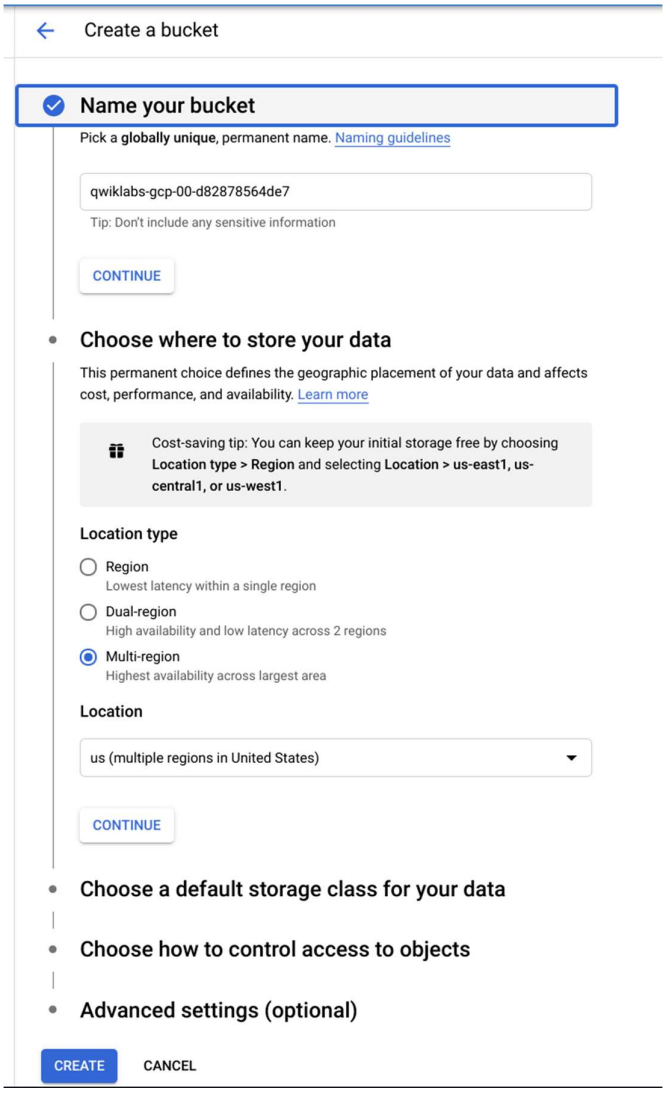
Now, click **Create table**.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created table in BigQuery dataset, you will see an assessment score.

Create a storage bucket

Go back to the Cloud Console and navigate to **Storage** > **Browser** > **Create bucket**:



Give your bucket a unique name. Leave all other default settings, then click **Create**.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created Cloud Storage bucket, you will see an assessment score.

Run the Pipeline

From the **Navigation menu** find the Big Data section and click on **Dataflow**.

Click on **+ Create job from template** at the top of the screen.

Enter a **Job name** for your Cloud Dataflow job.

Under **Dataflow Template**, select the *Pub/Sub Topic to BigQuery* template.

Under **Input Pub/Sub topic**, enter:

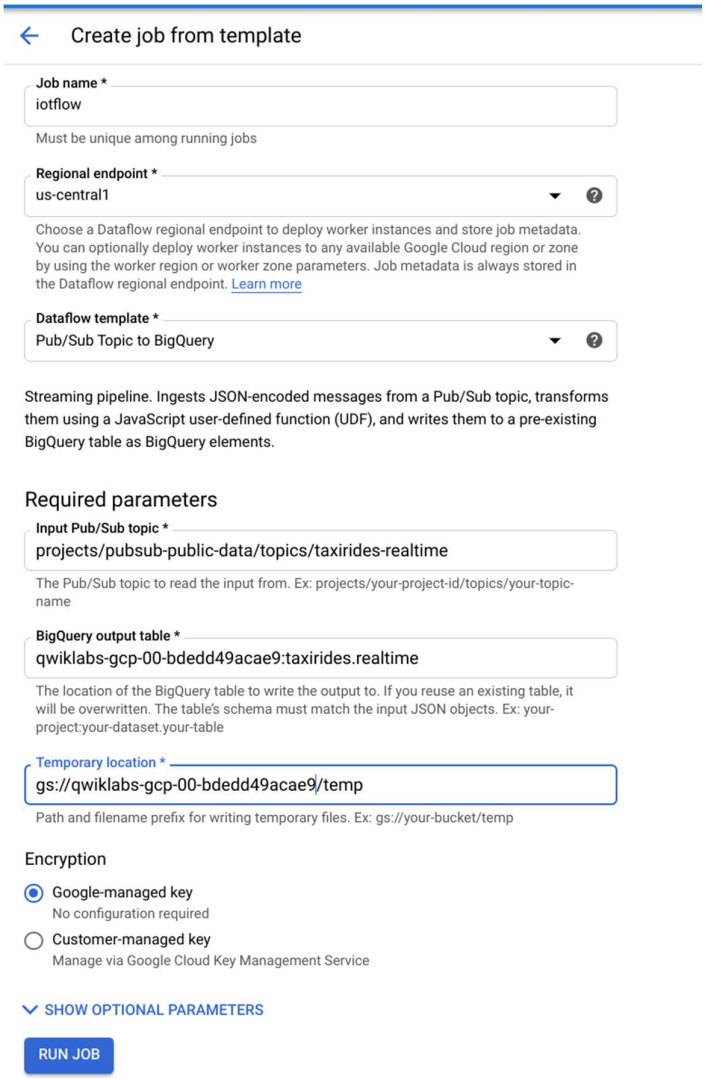
```
projects/pubsub-public-data/topics/taxirides-realtime
```

Under **BigQuery output table**, enter the name of the table that was created:

```
<myprojectid>:taxirides.realtime
```

Add your bucket as **Temporary Location**:

```
gs://Your_Bucket_Name/temp
```



Click the **Run job** button.

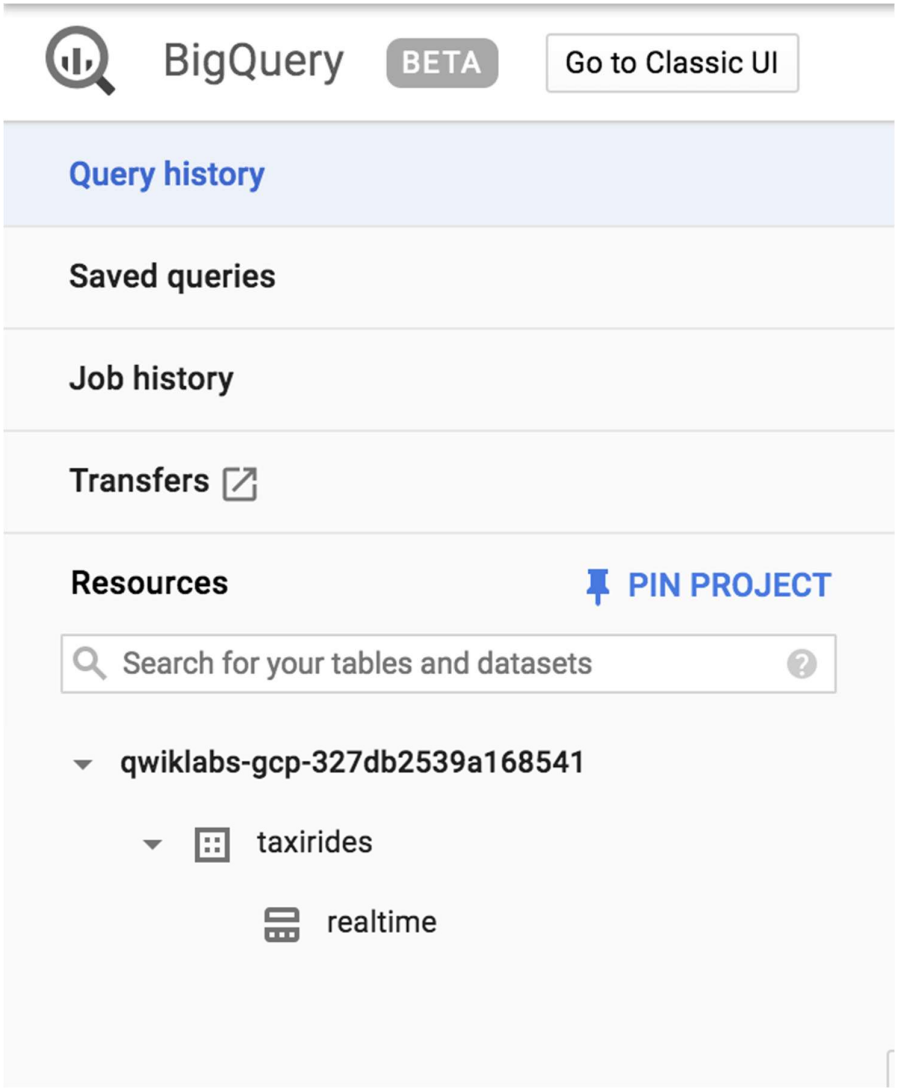
Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully run the Dataflow pipeline, you will see an assessment score.

You'll watch your resources build and become ready for use.

Now, let's go view the data written to BigQuery by clicking on **BigQuery** found in the Navigation menu.

When the BigQuery UI opens, you'll see the **taxirides** table added under your project name and **realtime** underneath that:


BigQuery
BETA
Go to Classic UI
Query history
Saved queries
Job history
Transfers
Resources
PIN PROJECT
Search for your tables and datasets
qwiklabs-gcp-327db2539a168541
taxirides
realtime

Submit a query

You can submit queries using standard SQL.

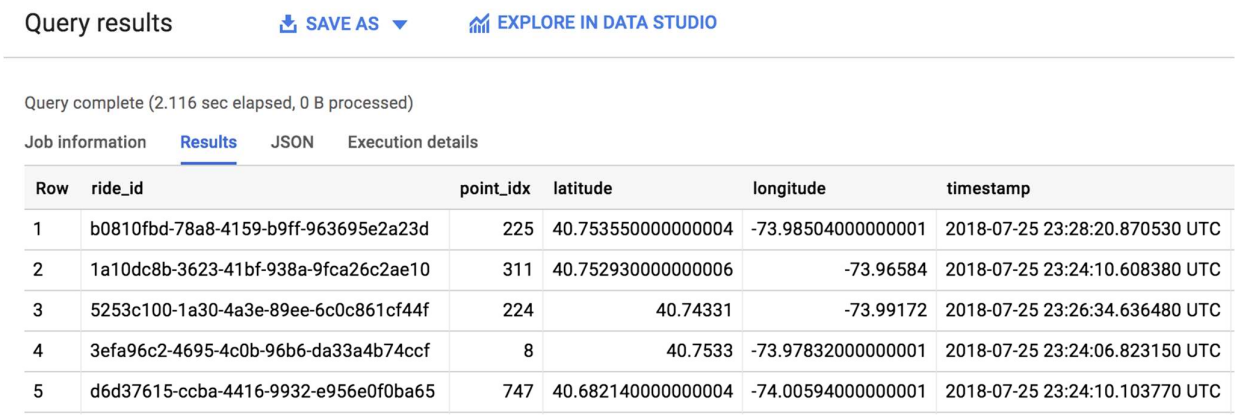
In the "Query editor" field add the following, replacing *myprojectid* with the Project ID from the Qwiklabs page:

```
SELECT * FROM `myprojectid.taxirides.realtime` LIMIT 1000
```

Now click **Run Query**.

If you run into any issues or errors, run the query again (the pipeline takes a minute to start up.)

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



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

Great work! You just pulled 1000 taxi rides from a Pub/Sub topic and pushed them to a BigQuery table. As you saw firsthand, templates are a practical, easy-to-use way to run Dataflow jobs. Be sure to check out some other Google Templates here.

Test your Understanding

Below are a multiple choice questions to reinforce your understanding of this lab's concepts. Answer them to the best of your abilities.

Google Cloud Dataflow supports batch processing.
True

Which Dataflow Template used in the lab to run the pipeline?
Pub/Sub to BigQuery

Congratulations!

Finish Your Quest

Continue your Quest with Baseline: Data, ML, AI. A Quest is a series of related labs that form a learning path. Completing this Quest earns you the badge above, to recognize your achievement. You can make your badge (or badges) public and link to them in your online resume or social media account. Enroll in this Quest and get immediate completion credit if you've taken this lab. See other available Qwiklabs Quests.

Next Steps / Learn More

This lab is part of a series of labs called Qwik Starts. These labs are designed to give you a little taste of the many features available with Google Cloud. Search for "Qwik Starts" in the lab catalog to find the next lab you'd like to take!

Google Cloud Training & Certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.

Manual Last Updated October 12, 2020
Lab Last Tested October 12, 2020