Loading Your Own Data into BigQuery

cloudskillsboost.google/games/2854/labs/17206

GSP865

Overview

BigQuery is Google's fully managed, NoOps, low cost analytics database. With BigQuery you can query terabytes and terabytes of data without having any infrastructure to manage or needing a database administrator. BigQuery uses SQL and can take advantage of the pay-as-you-go model. BigQuery allows you to focus on analyzing data to find meaningful insights.

In this lab you will ingest subsets of the NYC taxi trips data into tables inside of BigQuery.

What you'll learn

- Loading data into BigQuery from various sources
- Loading data into BigQuery using the CLI and Console
- Using DDL to create tables

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 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.

To complete this lab, you need:

Access to a standard internet browser (Chrome browser recommended).

Note: Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges incurred to your personal account.

Time to complete the lab---remember, once you start, you cannot pause a lab.

Note: If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.

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 the **Lab Details** panel with the following:
    - The **Open Google Console** button
    - Time remaining
    - The temporary credentials that you must use for this lab
    - Other information, if needed, to step through this lab
2. Click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.

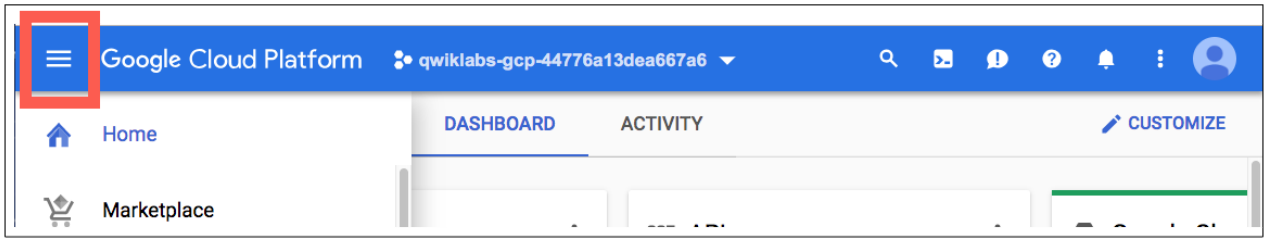
    Tip: Arrange the tabs in separate windows, side-by-side.

    Note: If you see the **Choose an account** dialog, click **Use Another Account**.
3. If necessary, copy the **Username** from the **Lab Details** panel and paste it into the **Sign in** dialog. Click **Next**.
4. Copy the **Password** from the **Lab Details** panel and paste it into the **Welcome** dialog. Click **Next**.

    Important: You must use the credentials from the left panel. Do not use your Google Cloud Skills Boost credentials. **Note:** Using your own Google Cloud account for this lab may incur extra charges.
5. 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.



Open the BigQuery console

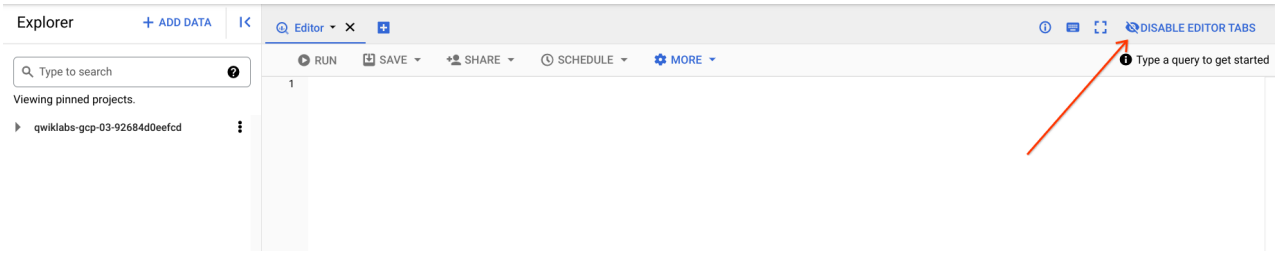
1. In the Google Cloud Console, select **Navigation menu** > **BigQuery**.

The **Welcome to BigQuery in the Cloud Console** message box opens. This message box provides a link to the quickstart guide and the release notes.

2. Click **Done**.

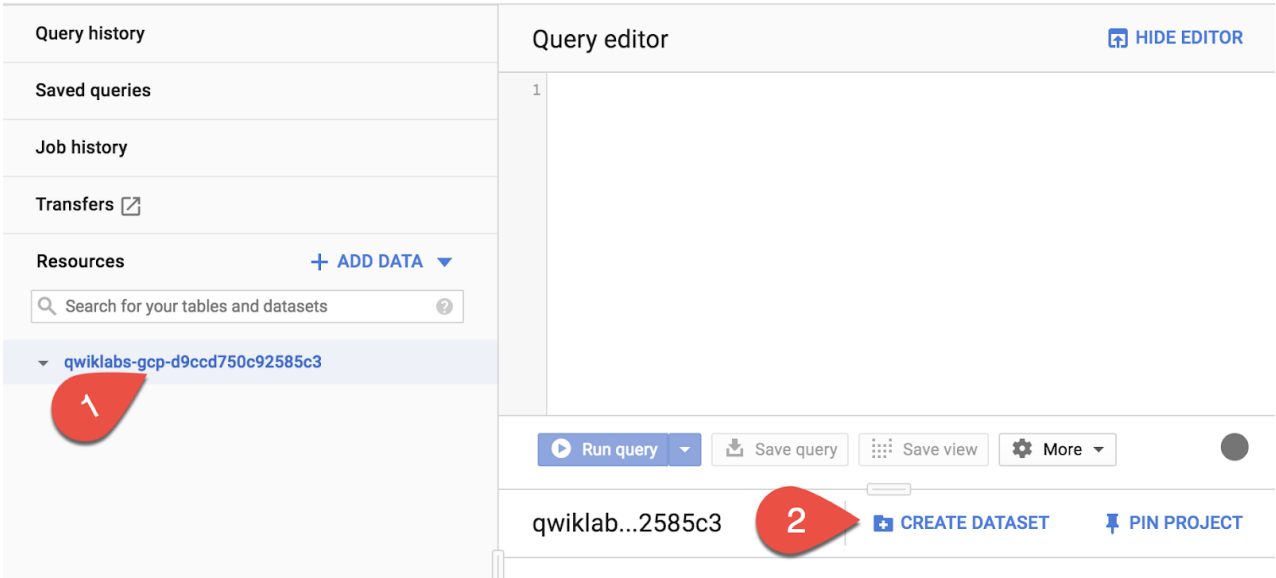
The BigQuery console opens.

On the top of the page, click **DISABLE EDITOR TABS**. This adjusts the BigQuery user interface to non-preview mode.



Create a new dataset to store tables

In the BigQuery console, click on the name of your project, then click **Create Dataset**.



Set the *Dataset ID* to **nyctaxi**. Leave the other fields at their default values.

Click **Create dataset**.

You'll now see the nyctaxi dataset under your project name.

To check your progress in this lab, click **Check my progress** below. A checkmark means you're successful.

Creating a dataset to store new tables

Ingest a new Dataset from a CSV

In this section, you will load a local CSV into a BigQuery table.

1. Download a subset of the NYC taxi 2018 trips data locally onto your computer from here :

2. In the BigQuery Console, Select the **nyctaxi** dataset then click **Create Table**

Specify the below table options:

Source:

- Create table from: **Upload**
- Choose File: **select the file you downloaded locally earlier**
- File format: **CSV**

Destination:

Table name: **2018trips** Leave all other setting at default.

Schema:

Check **Auto Detect** (**tip**: Not seeing the checkbox? Ensure the file format is CSV and not Avro)

Advanced Options

Leave at default values

Click **Create Table**.

3. You should now see the **2018trips** table below the nyctaxi dataset.

Select the 2018trips table and view **details**:

4. Select **Preview** and confirm all columns have been loaded (sampled below):

You have successfully loaded in a CSV file into a new BigQuery table.

Running SQL Queries

Next, practice with a basic query on the 2018trips table.

1. In the Query Editor, write a query to list the top 5 most expensive trips of the year:

```
#standardSQL SELECT * FROM nyctaxi.2018trips ORDER BY fare_amount DESC LIMIT 5
```

To check your progress in this lab, click **Check my progress** below. A checkmark means you're successful.

Ingest a new Dataset from a CSV

Ingest a new Dataset from Google Cloud Storage

Now, lets try load another subset of the same 2018 trip data that is available on Cloud Storage. And this time, let's use the CLI tool to do it.

1. In your Cloud Shell, run the following command :

bq load \ --source_format=CSV \ --autodetect \ --noreplace \ nyctaxi.2018trips \ gs://cloud-training/OCBL013/nyc_tlc_yellow_trips_2018_subset_2.csv
Note: With the above load job, you are specifying that this subset is to be appended to the existing 2018trips table that you created above.

2. When the load job is complete, you will get a confirmation on the screen.

3. Back on your BigQuery console, select the 2018trips table and view **details**. Confirm that the row count has now almost doubled.

4. You may want to run the same query like earlier to see if the top 5 most expensive trips have changed.

To check your progress in this lab, click **Check my progress** below. A checkmark means you're successful.

Ingest a dataset from google cloud storage

Create tables from other tables with DDL

The 2018trips table now has trips from throughout the year. What if you were only interested in January trips? For the purpose of this lab, we will keep it simple and focus only on pickup date and time. Let's use DDL to extract this data and store it in another table

1. In the Query Editor, run the following CREATE TABLE command :

#standardSQL CREATE TABLE nyctaxi.january_trips AS SELECT * FROM nyctaxi.2018trips WHERE EXTRACT(Month FROM pickup_datetime)=1;

2. Now run the below query in your Query Editor find the longest distance traveled in the month of January:

#standardSQL SELECT * FROM nyctaxi.january_trips ORDER BY trip_distance DESC LIMIT 1
To check your progress in this lab, click **Check my progress** below. A checkmark means you're successful.

Create tables from other tables with DDL

Congratulations!

You've successfully created a new dataset and ingested data into BigQuery from CSV, Google Cloud Storage, and other BigQuery tables.

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 June 23, 2022

Lab Last Tested June 23, 2022