Loading data into BigQuery | Qwiklabs

Wednesday, November 4, 2020 2:00 PM

Clipped from:
https://googlecourses.qwiklabs.com/course_sessions/67951/labs/11690

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.

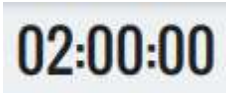
In this lab you will ingest subsets of the NYC taxi trips data into tables inside of BigQuery.

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

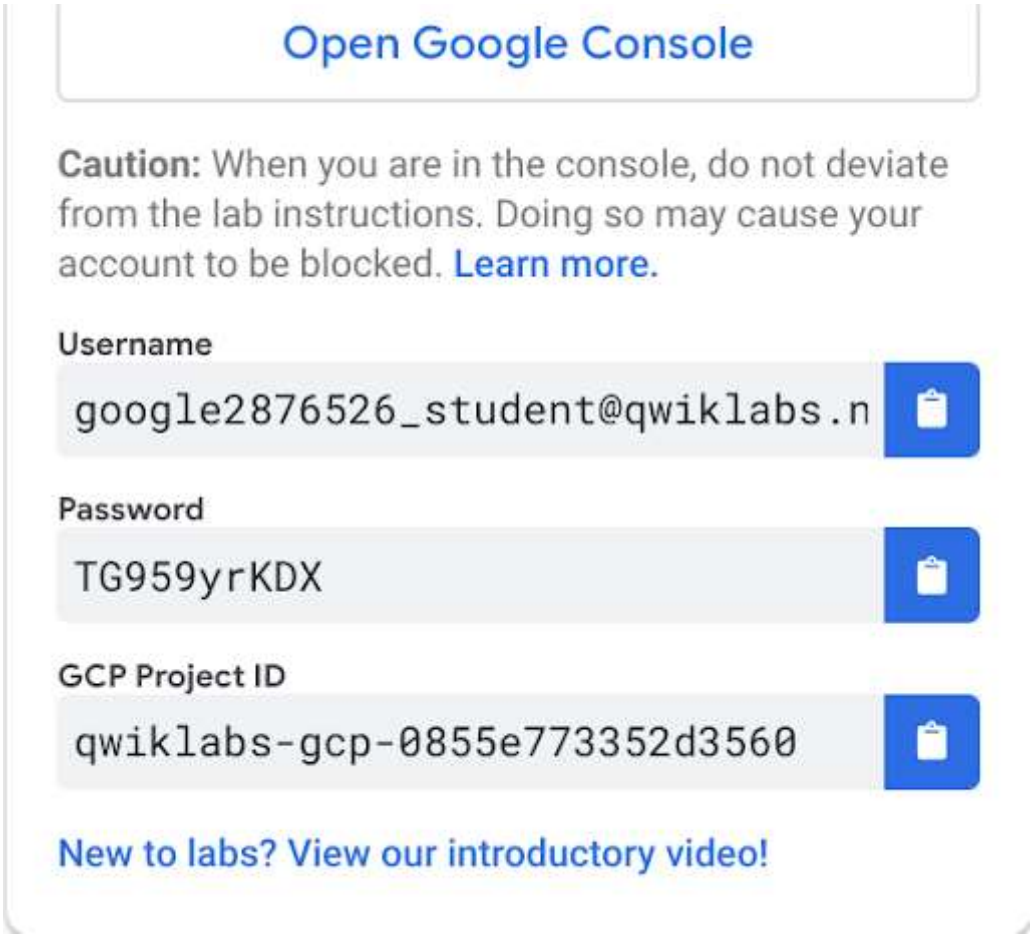
Setup

For each lab, you get a new GCP project and set of resources for a fixed time at no cost.

1. Make sure you signed into Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example,


and make sure you can finish in that time block.
3. When ready, click


.
4. Note your lab credentials. You will use sign in to Cloud Platform Console.

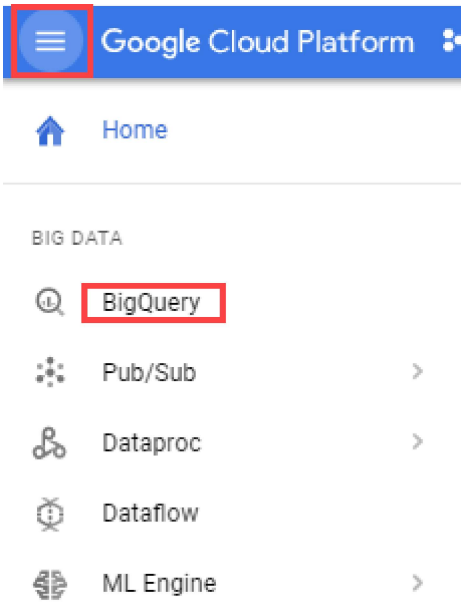


5. Click **Open Google Console**.
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.

1. Accept the terms and skip the recovery resource page.

Open BigQuery Console

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

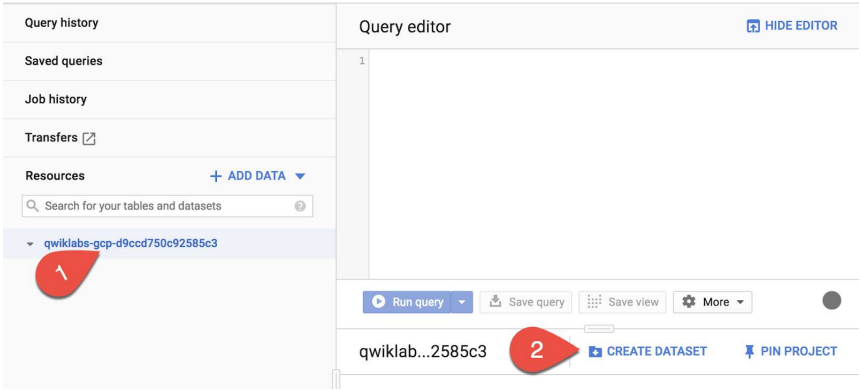


The **Welcome to BigQuery in the Cloud Console** message box opens. This message box provides a link to the quickstart guide and lists UI updates.

Click **Done**.

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.

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 othe
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**.

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

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

1. 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
```

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_subse
```

```
t_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.

1. When the load job is complete, you will get a confirmation on the screen.
2. Back on your BigQuery console, select the 2018trips table and view **details**. Confirm that the row count has now almost doubled.
3. You may want to run the same query like earlier to see if the top 5 most expensive trips have changed.

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;
```

1. 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
```

Congratulations!

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

End your lab

When you have completed your lab, click **End Lab**. Qwiklabs removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.

Manual Last Updated: Sep 27, 2019

Lab Last Tested: Sep 27, 2019