Loading data into BigQuery

## Overview

[BigQuery](http://bigquery.cloud.google.com/) 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

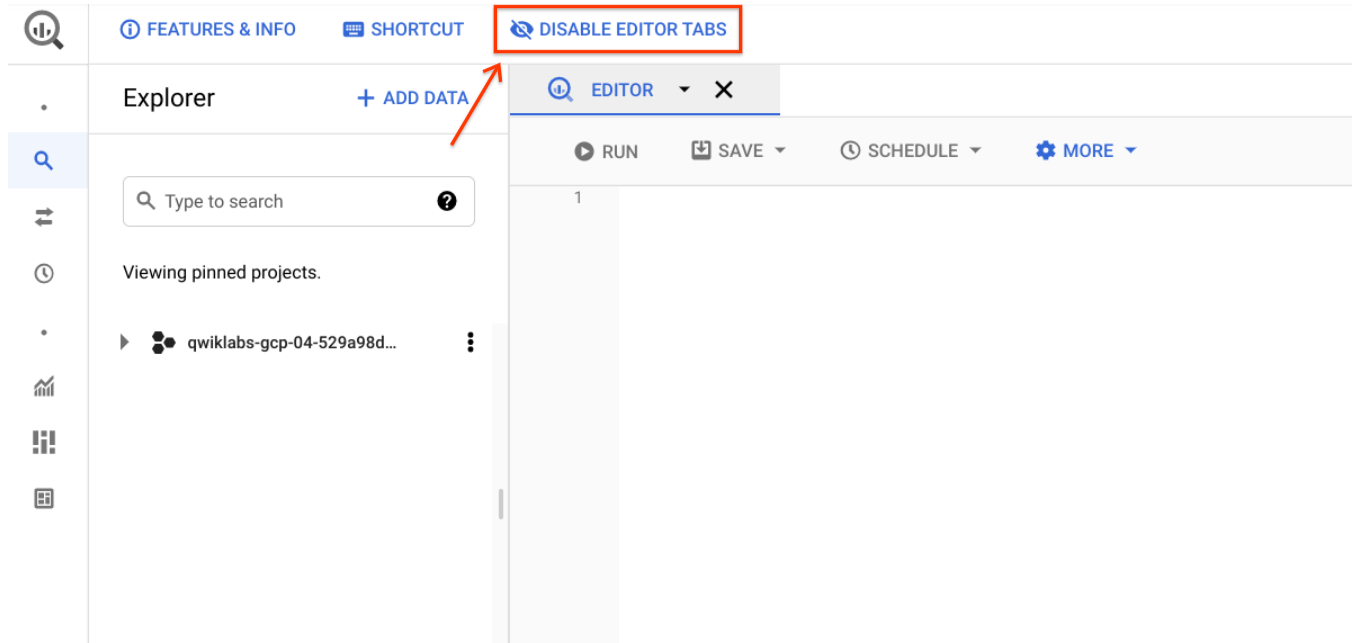
### **Open BigQuery Console**

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 lists UI updates.

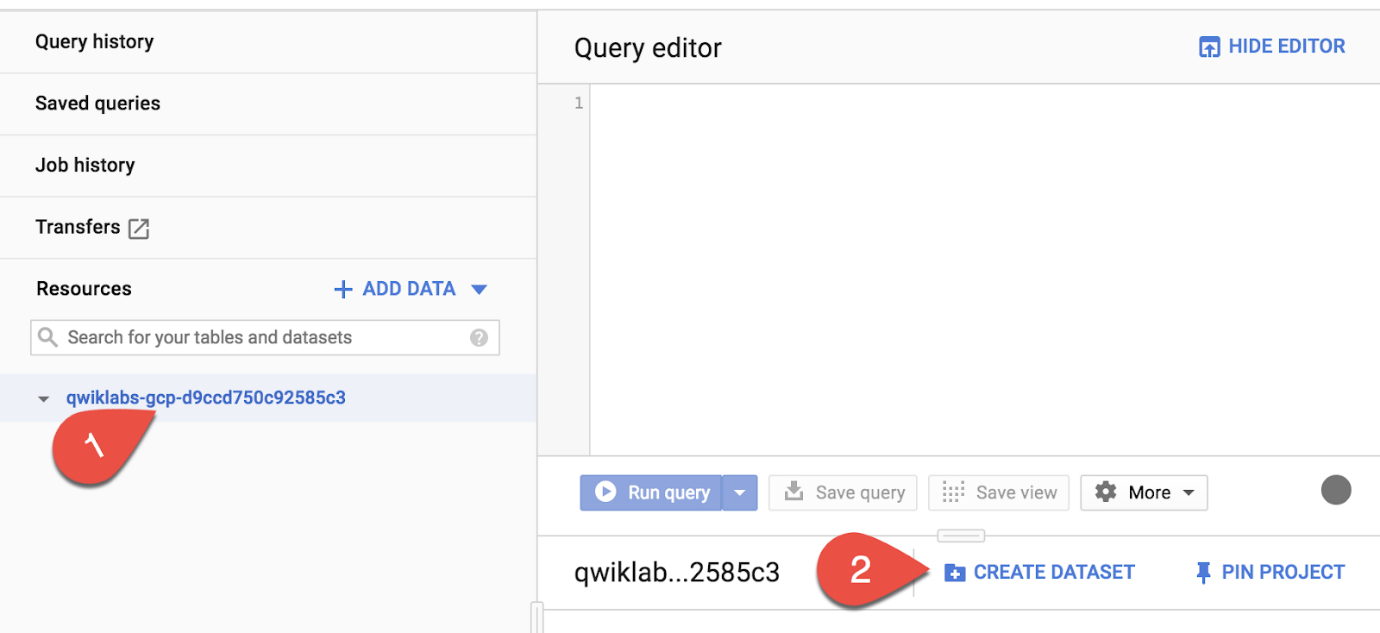
1. Click **Done**.

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.

Click Check my progress to verify the objective.

Creating a dataset to store new tables

Check my progress

## 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](https://storage.googleapis.com/cloud-training/OCBL013/nyc_tlc_yellow_trips_2018_subset_1.csv) :
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**.

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

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

How many rows are in the table?



1,090



900



10,018



1,200

Submit

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

Copied!

content\_copy

What was the highest fare amount in the year?



339



300



250

Submit

Click Check my progress to verify the objective.

Ingest a new Dataset from a CSV

Check my progress

## 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

Copied!

content\_copy

**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.

Click Check my progress to verify the objective.

Ingest a dataset from google cloud storage

Check my progress

## 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;

Copied!

content\_copy

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

1

Copied!

content\_copy

Click Check my progress to verify the objective.

Create tables from other tables with DDL

Check my progress

## Congratulations!

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