Using BigQuery to do Analysis | Qwiklabs

Monday, November 2, 2020 2:55 PM

Clipped from:
https://googlecourses.qwiklabs.com/course_sessions/67951/labs/11672

Overview

In this lab you analyze 2 different public datasets, run queries on them, separately and then combined, to derive interesting insights.

What you'll learn

In this lab, you will:

- Carry out interactive queries on the BigQuery console.
- Combine and run analytics on multiple datasets.

Prerequisites

This is a **fundamental level** lab and assumes some experience with BigQuery and SQL.

Introduction

This lab uses two public datasets in BigQuery: weather data from the US National Oceanic and Atmospheric Administration (NOAA), and bicycle rental data from New York City.

You will encounter, for the first time, several aspects of Google Cloud Platform that are of great benefit to scientists:

1. **Serverless.** No need to download data to your machine in order to work with it - the dataset will remain on the cloud.
2. **Ease of use.** Run ad-hoc SQL queries on your dataset without having to prepare the data, like indexes, beforehand. This is invaluable for data exploration.
3. **Scale.** Carry out data exploration on extremely large datasets interactively. You don't need to sample the data in order to work with it in a timely manner.
4. **Shareability**. You will be able to run queries on data from different datasets without any issues. BigQuery is a convenient way to share datasets. Of course, you can also keep your data private, or share them only with specific persons -- not all data need to be public.

The end-result is that you will find if there are lesser bike rentals on rainy days.

Setup and requirements

Qwiklabs setup

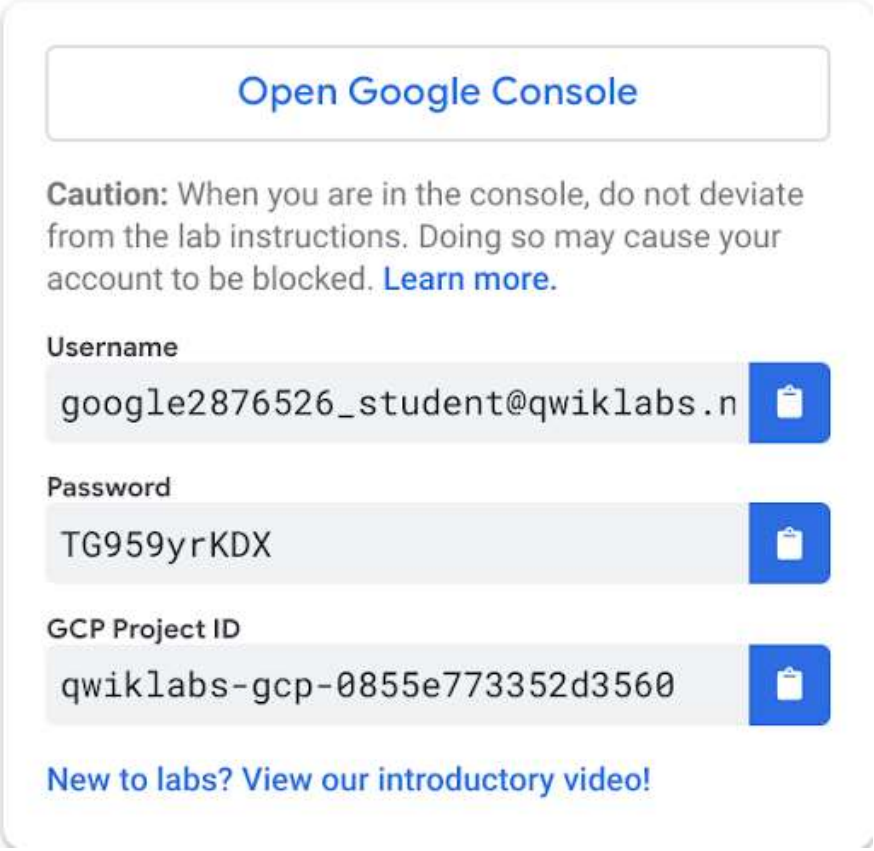
For each lab, you get a new GCP project and set of resources for a fixed time at no cost.

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 them to sign in to Cloud Platform Console.



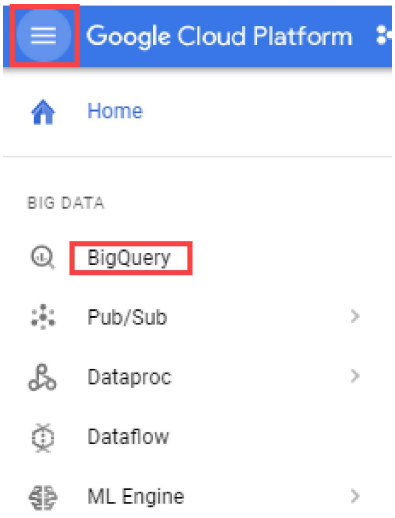
5. Click **Open Google Console**.
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.

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

Explore bicycle rental data

Open BigQuery Console

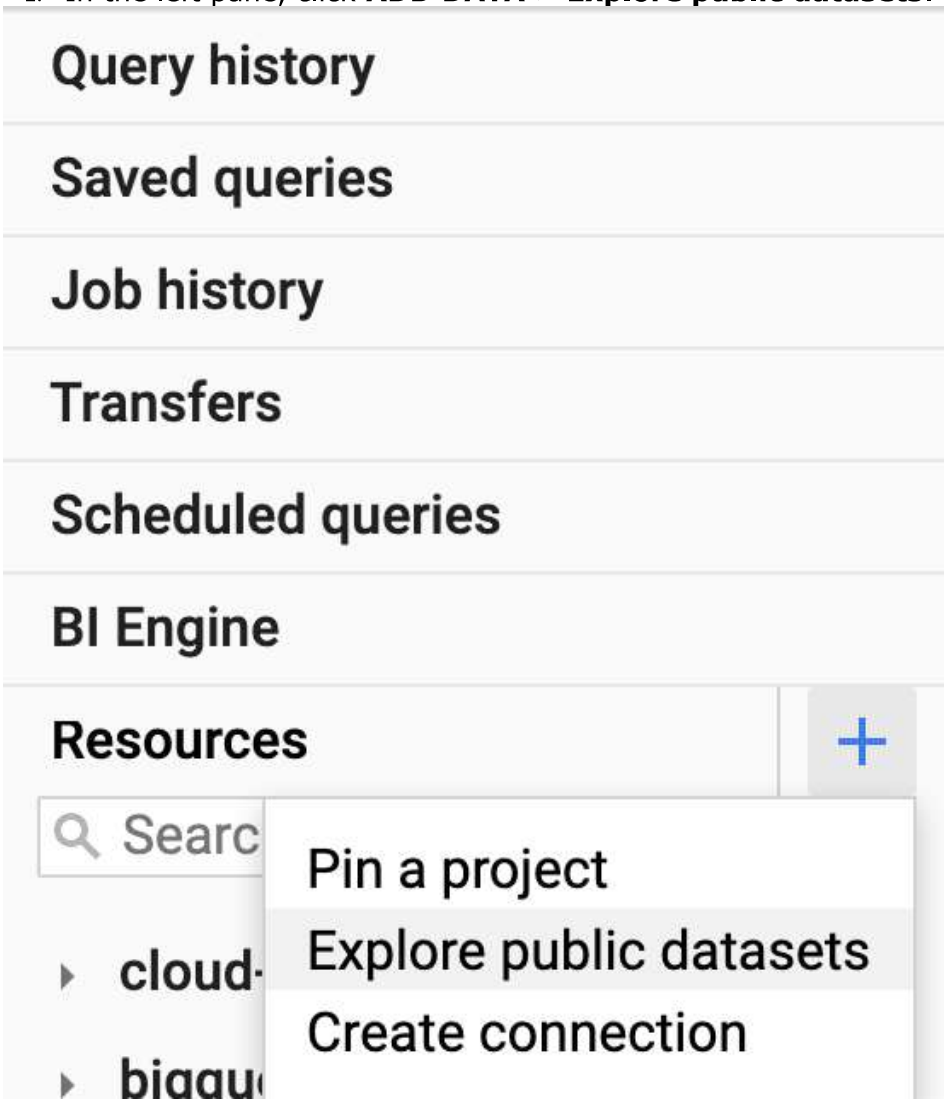
In the Google Cloud Console, select **Navigation menu** > **BigQuery**:



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**.

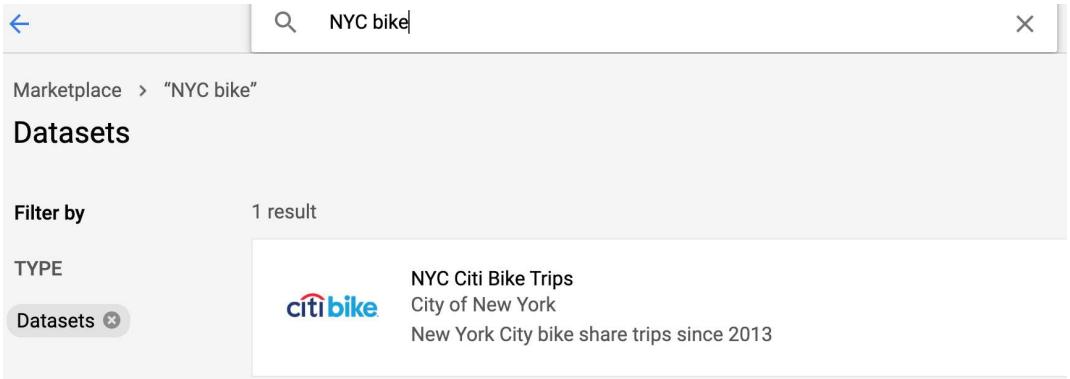
1. In the left pane, click **ADD DATA** > **Explore public datasets**.



The Datasets window opens.

1. In the **Search** bar, type "NYC bike" then press **Enter**.

1 result, NYC Citi Bike Trips, displays.

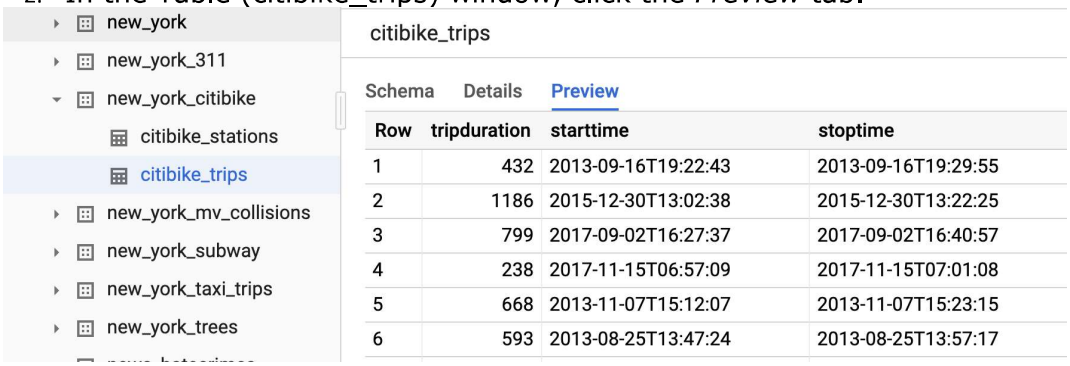


1. Click the NYC Citi Bike Trips dataset and then click **VIEW DATASET**.

Your The Google BigQuery console opens in a new browser tab.To keep your workspace organized, close this new browser tab and refresh the first tab.

1. In the BigQuery console (in the first browser tab) you see two projects in the left pane, one named your Qwiklabs project ID, and one named **bigquery-public-data**.

1. In the left pane of the BigQuery console, select **bigquery-public-data** > **new_york_citibike** > **citibike_trips** table.
2. In the Table (citibike_trips) window, click the *Preview* tab.



Row	tripduration	starttime	stoptime
1	432	2013-09-16T19:22:43	2013-09-16T19:29:55
2	1186	2015-12-30T13:02:38	2015-12-30T13:22:25
3	799	2017-09-02T16:27:37	2017-09-02T16:40:57
4	238	2017-11-15T06:57:09	2017-11-15T07:01:08
5	668	2013-11-07T15:12:07	2013-11-07T15:23:15
6	593	2013-08-25T13:47:24	2013-08-25T13:57:17

1. Examine the columns and some of the data values.

Paste the following in the Query editor textbox:

```
SELECT
  MIN(start_station_name) AS start_station_name,
  MIN(end_station_name) AS end_station_name,
  APPROX_QUANTILES(tripduration, 10)[OFFSET (5)] AS typical_duration,
  COUNT(tripduration) AS num_trips
FROM
  `bigquery-public-data.new_york_citibike.citibike_trips`
WHERE
  start_station_id != end_station_id
GROUP BY
  start_station_id,
  end_station_id
ORDER BY
  num_trips DESC
LIMIT
  10
```

Click **Run**. Look at the result and try to determine what this query does ? (Hint: typical duration for the 10 most common one-way rentals)

1. Next, run the below to find another interesting fact: total distance travelled by each bicycle in the dataset. Note that the query limits the results to only top 5.

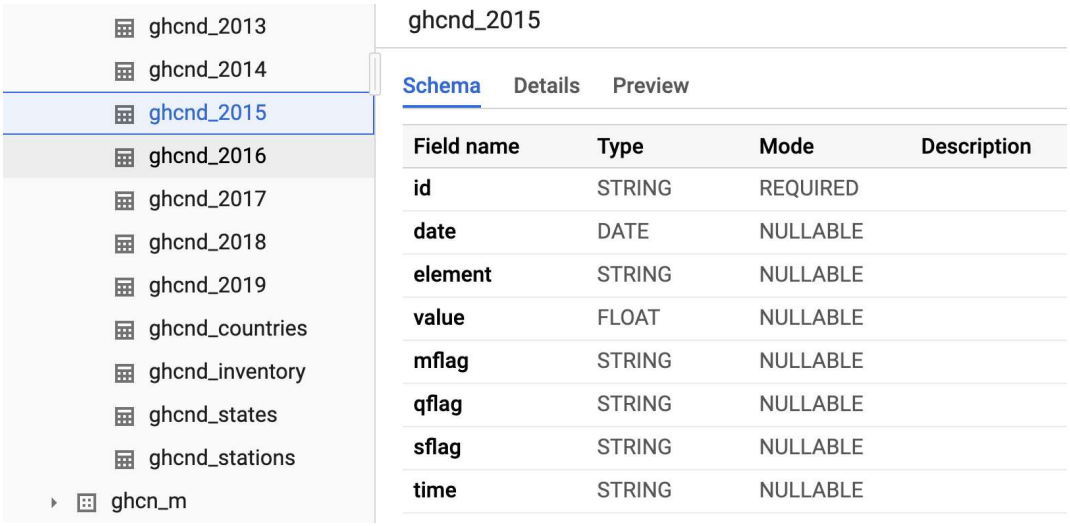
```
WITH
  trip_distance AS (
SELECT
  bikeid,
  ST_Distance(ST_GeogPoint(s.longitude,
      s.latitude),
    ST_GeogPoint(e.longitude,
      e.latitude)) AS distance
FROM
  `bigquery-public-data.new_york_citibike.citibike_trips`,
  `bigquery-public-data.new_york_citibike.citibike_stations` as s,
  `bigquery-public-data.new_york_citibike.citibike_stations` as e
WHERE
  start_station_id = s.station_id
  AND end_station_id = e.station_id )
SELECT
  bikeid,
  SUM(distance)/1000 AS total_distance
FROM
  trip_distance
GROUP BY
  bikeid
ORDER BY
  total_distance DESC
LIMIT
  5
```

Note: For this query, we also used the other table in the dataset called **citibike_stations** to get bicycle station information.

1. Click **HIDE EDITOR** on top right to close the Query editor.

Explore the weather dataset

In the left pane of the BigQuery Console, select the newly added `bigquery-public-data` project and select **ghcn_d** > **ghcnd_2015**. Then click on the **Preview** tab. Your console should resemble the following:



Examine the columns and some of the data values.

Click **COMPOSE NEW QUERY** and enter the following:

```
SELECT
  wx.date,
  wx.value/10.0 AS prcp
FROM
  `bigquery-public-data.ghcn_d.ghcnd_2015` AS wx
WHERE
  id = 'USW00094728'
  AND qflag IS NULL
  AND element = 'PRCP'
ORDER BY
  wx.date
```

Click **Run.**

This query will return rainfall (in mm) for all days in 2015 from a weather station in New York whose id is provided in the query (the station corresponds to NEW YORK CNTRL PK TWR)

Find correlation between rain and bicycle rentals

How about joining the bicycle rentals data against weather data to learn whether there are fewer bicycle rentals on rainy days?

Click **COMPOSE NEW QUERY** and run the following query :

```
WITH bicycle_rentals AS (
  SELECT
    COUNT(starttime) as num_trips,
    EXTRACT(DATE from starttime) as trip_date
  FROM `bigquery-public-data.new_york_citibike.citibike_trips`
  GROUP BY trip_date
),

rainy_days AS
(
SELECT
  date,
  (MAX(prcp) > 5) AS rainy
FROM (
  SELECT
    wx.date AS date,
    IF (wx.element = 'PRCP', wx.value/10, NULL) AS prcp
  FROM
    `bigquery-public-data.ghcn_d.ghcnd_2015` AS wx
  WHERE
    wx.id = 'USW00094728'
)
GROUP BY
  date
)

SELECT
  ROUND(AVG(bk.num_trips)) AS num_trips,
  wx.rainy
FROM bicycle_rentals AS bk
JOIN rainy_days AS wx
ON wx.date = bk.trip_date
GROUP BY wx.rainy
```

Click **Run**.

Now you can see the results of joining the bicycle rental dataset with a weather dataset that comes from a completely different source.

Row	num_trips	rainy
1	28598.0	false
2	19503.0	true

Running the query yields that, yes, New Yorkers ride the bicycle 47% fewer times when it rains.

Summary

In this lab you did ad-hoc queries on two datasets. You were able to query the data without setting up any clusters, creating any indexes, etc. You were also able to mash up the two datasets and get some interesting insights. All without ever leaving your browser!

Congratulations!

You learned how to run some very interesting queries on BigQuery!

Manual Last Updated Sep 25, 2019

Lab Last Tested Sep 27, 2019