

Using BigQuery to do Analysis | Google Cloud Skills Boost

Qwiklabs : 8-10 minutes

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

Lab setup

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

1. Sign in to Qwiklabs using an **incognito window**.

2. Note the lab's access time (for example, 1:15:00), and make sure you can finish within that time.
There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click **Start lab**.
4. Note your lab credentials (**Username** and **Password**). You will use them to sign in to the Google Cloud Console.
5. Click **Open Google Console**.
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.
If you use other credentials, you'll receive errors or **incur charges**.
7. Accept the terms and skip the recovery resource page.

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.

2. Click **Done**.

Task 1. Explore bicycle rental data

1. In the left pane, click **+ ADD DATA** > **Public datasets** from **Additional sources**.

Add data

Popular sources



Local file

Upload a local file



Google Cloud Storage

Google object storage service



Connections to external data sources

Connection from BigQuery to an external data source

Additional sources

Viewing all 23 results.



Search for and pin a project

Search for a BigQuery project and add it to the Explorer



Pin a project by name

Add a BigQuery project to the Explorer by project name



Analytics Hub

Discover and subscribe to public, commercial or privately shared datasets



Google Drive

Google storage service



Amazon S3

Amazon object storage service



Azure Blob Storage

Microsoft object storage service



Public Datasets

BigQuery public datasets from the Google Cloud Public Dataset Program

The Datasets window opens.

2. In the **Search marketplace** bar, type `NYC bike` then press ENTER.

3. One result, **NYC Citi Bike Trips**, is returned. Click on the dataset name and then **View Dataset**.

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

Note: If the new project `bigquery-public-data` doesn't appear to the Explorer panel, then click on **+ ADD DATA > Star a project by name > Star a project** (`bigquery-public-data`) and **STAR**.

4. In the BigQuery console, you see two projects in the left pane, one named your Qwiklabs project ID, and one named **bigquery-public-data**.

5. In the left pane of the BigQuery console, select **bigquery-public-data > new_york_citibike > citibike_trips** table.

6. In the Table (citibike_trips) window, click the Preview tab.

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

Click **Compose New Query** and enter the following:

```
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
```

8. 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)

9. Next, run the below to find another interesting fact: total distance traveled 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_name = s.name AND
end_station_name = e.name) 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.

Task 2. Explore the weather dataset

1. In the left pane of the BigQuery Console, select the newly added bigquery-public-data project and select **ghcn_d > ghcnd_2015**.

2. Then click on the **Preview** tab. Your console should resemble the following:

ghcnd_2013

ghcnd_2014

ghcnd_2015

ghcnd_2016

ghcnd_2017

ghcnd_2018

ghcnd_2019

ghcnd_countries

ghcnd_inventory

ghcnd_states

ghcnd_stations

ghcn_m

ghcnd_2015

Schema

Details

Preview

Field name	Type	Mode	Description
id	STRING	REQUIRED	
date	DATE	NULLABLE	
element	STRING	NULLABLE	
value	FLOAT	NULLABLE	
mflag	STRING	NULLABLE	
qflag	STRING	NULLABLE	
sflag	STRING	NULLABLE	
time	STRING	NULLABLE	

Examine the columns and some of the data values.

3. 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
```

4. 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).

Task 3. 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?

1. Click **Compose New Query** and enter the following:

```
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
```

2. 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!

End your lab

When you have completed your lab, click **End Lab**. Google Cloud Skills Boost 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.

Copyright 2022 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.