Weather Data in BigQuery

GSP009

Overview

In this lab you will analyze historical weather observations using BigQuery and use weather data in conjunction with other datasets.

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. The following lab can get you up to speed with these Google Cloud services:

- BigQuery: Qwik Start - Console
  Take these labs first if you have never worked with BigQuery or MySQL, then come back to this one.

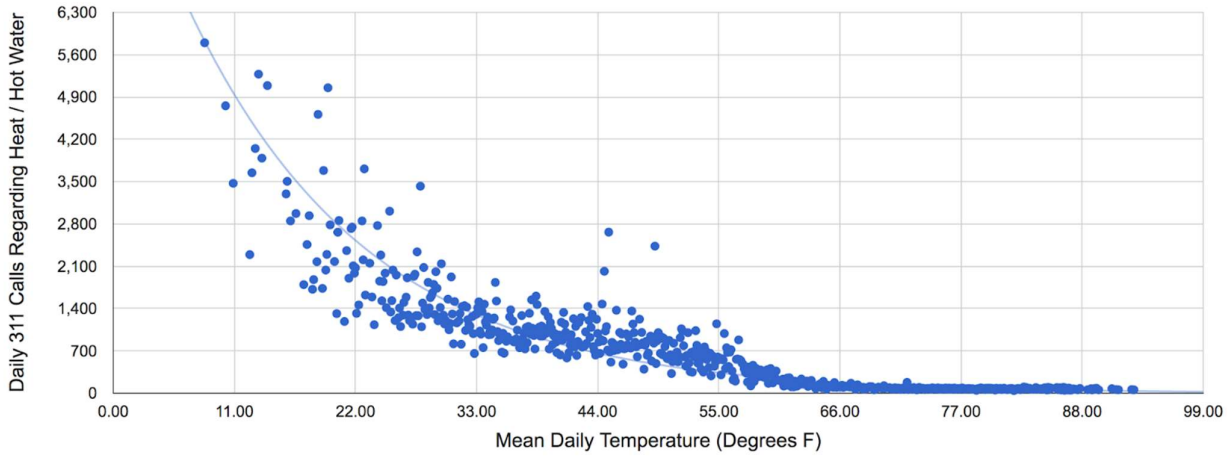
Introduction

This lab uses two public datasets in BigQuery: weather data from NOAA and citizen complaints data from New York City.

You will encounter, for the first time, several aspects of Google Cloud 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 what types of municipal complaints are correlated with weather. For example, you will find (not surprisingly) that complaints about residential furnaces are most common when it is cold outside:



Setup and requirements

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 Qwiklabs 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.

What you need

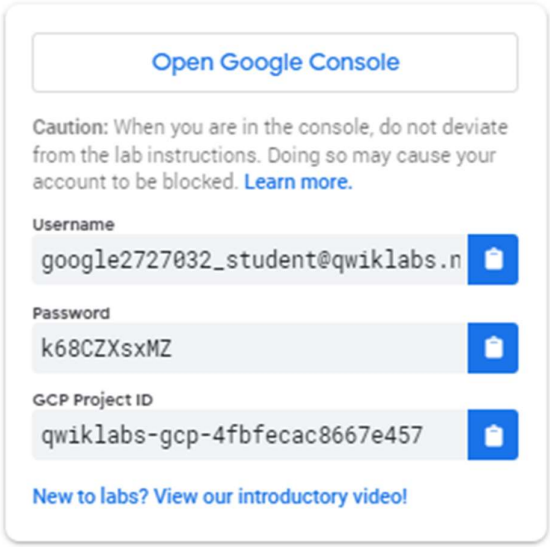
To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- Time to complete the lab.
  Note: If you already have your own personal Google Cloud account or project, do not use it for this lab.

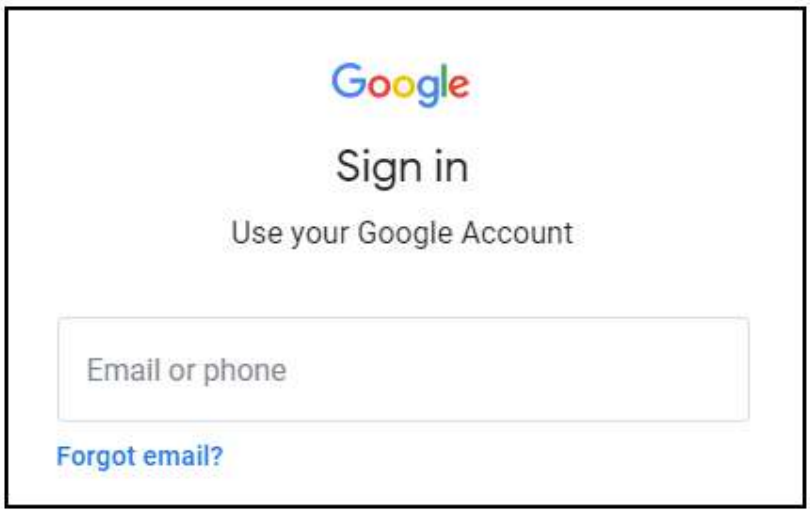
Note: If you are using a Pixelbook, open an Incognito window to run this lab.

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 a panel populated with the temporary credentials that you must use for this lab.

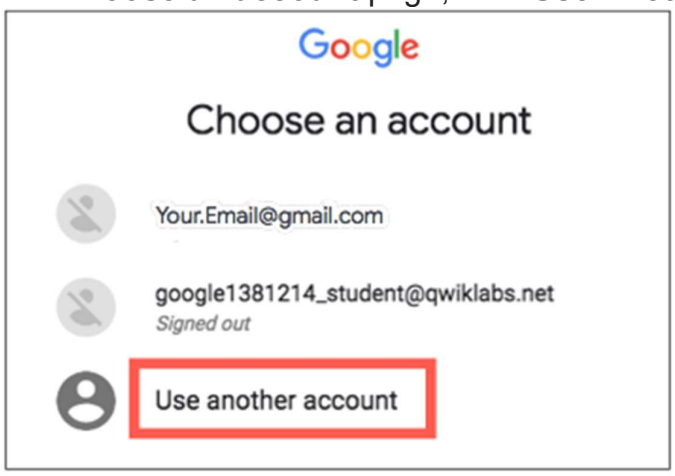


2. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.



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

If you see the **Choose an account** page, click **Use Another Account**.



3. In the **Sign in** page, paste the username that you copied from the Connection Details panel. Then copy and paste the password.

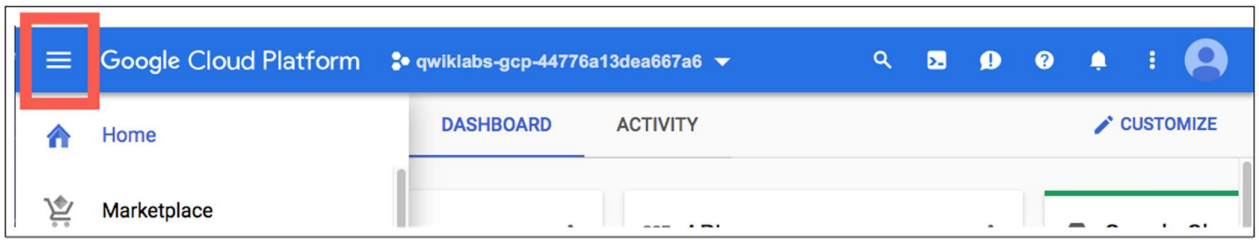
   Important: You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

4. 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.

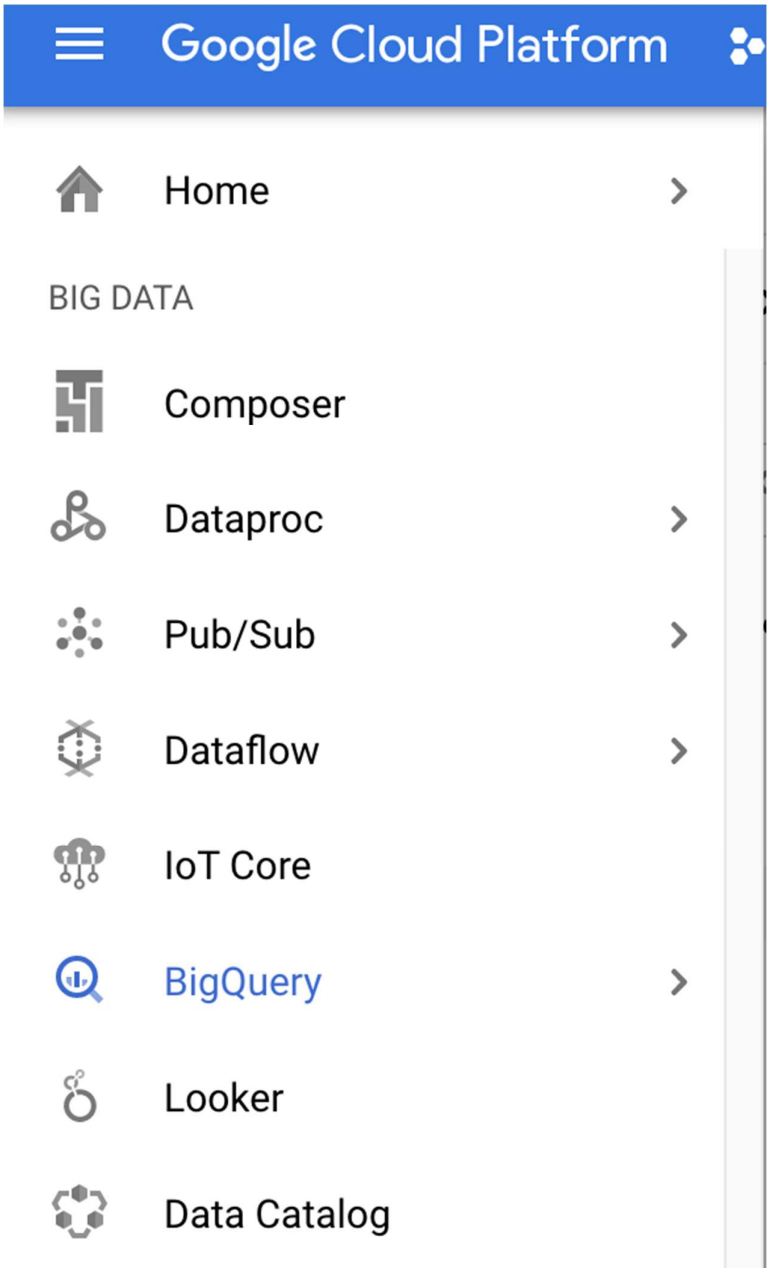
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.



Explore weather data

Open BigQuery Console

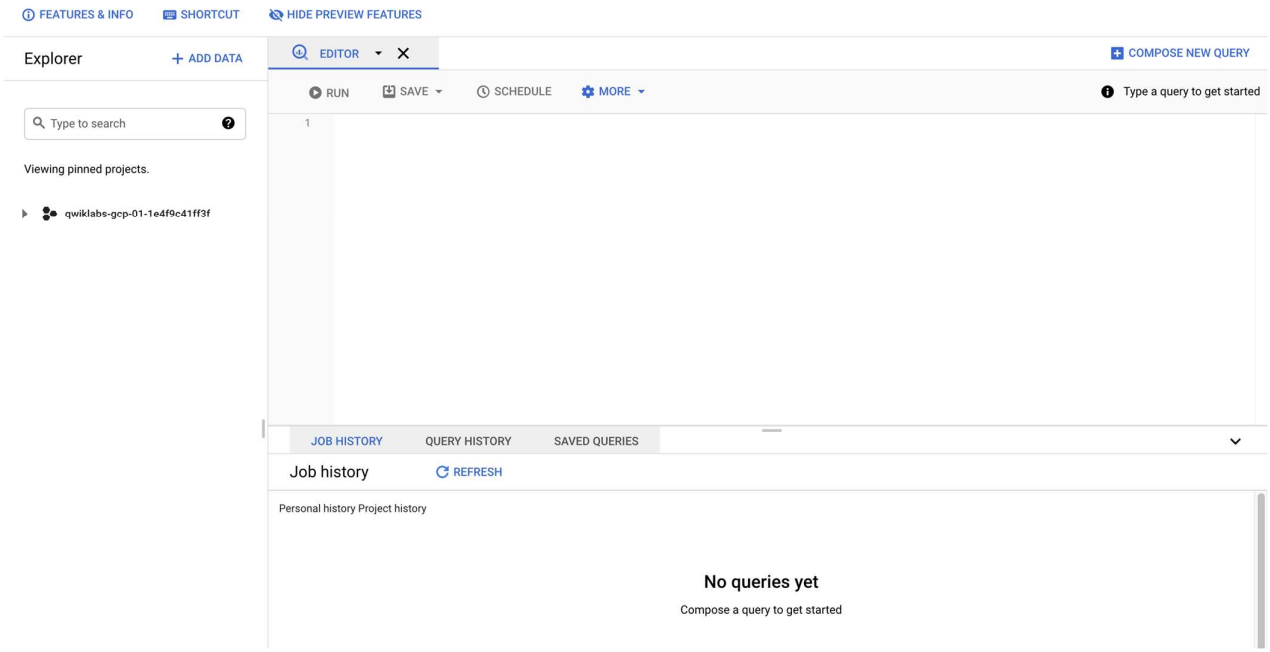
In the Google Cloud Console, select **Navigation menu** > **BigQuery**:



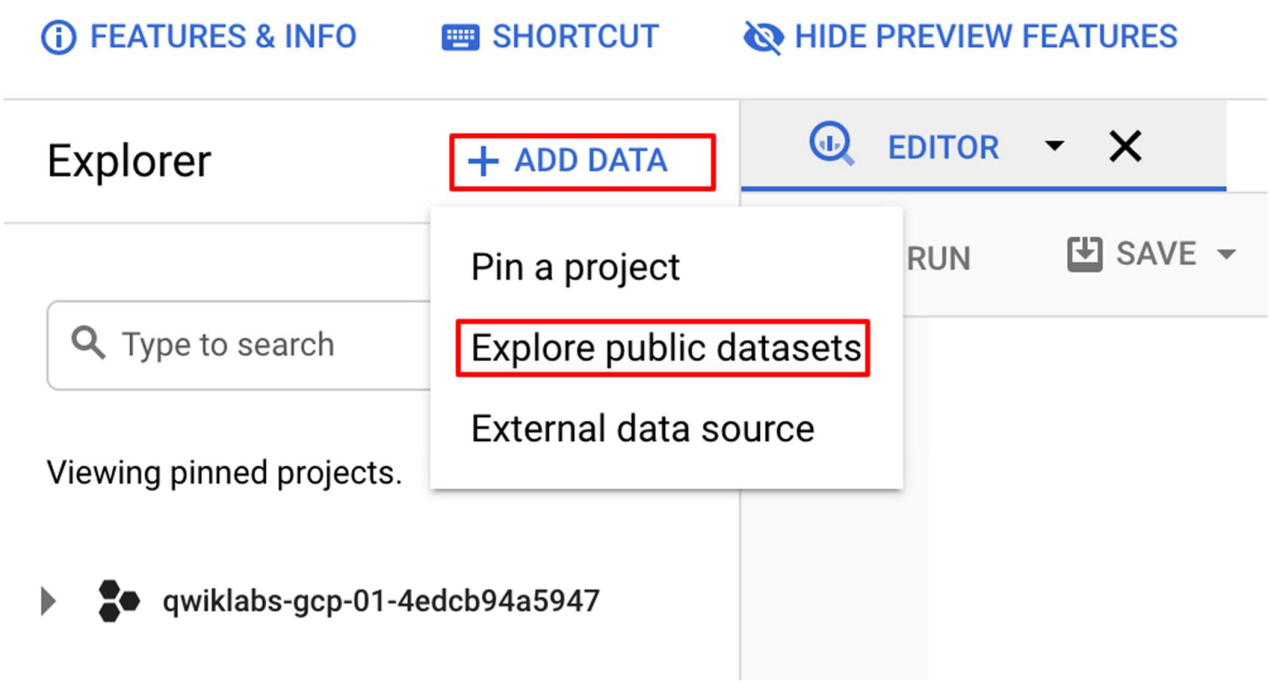
The **Welcome to BigQuery in the Cloud Console** message box opens. This message box provides a link to the quickstart guide and the release notes.

Click **Done**.

The BigQuery console opens.



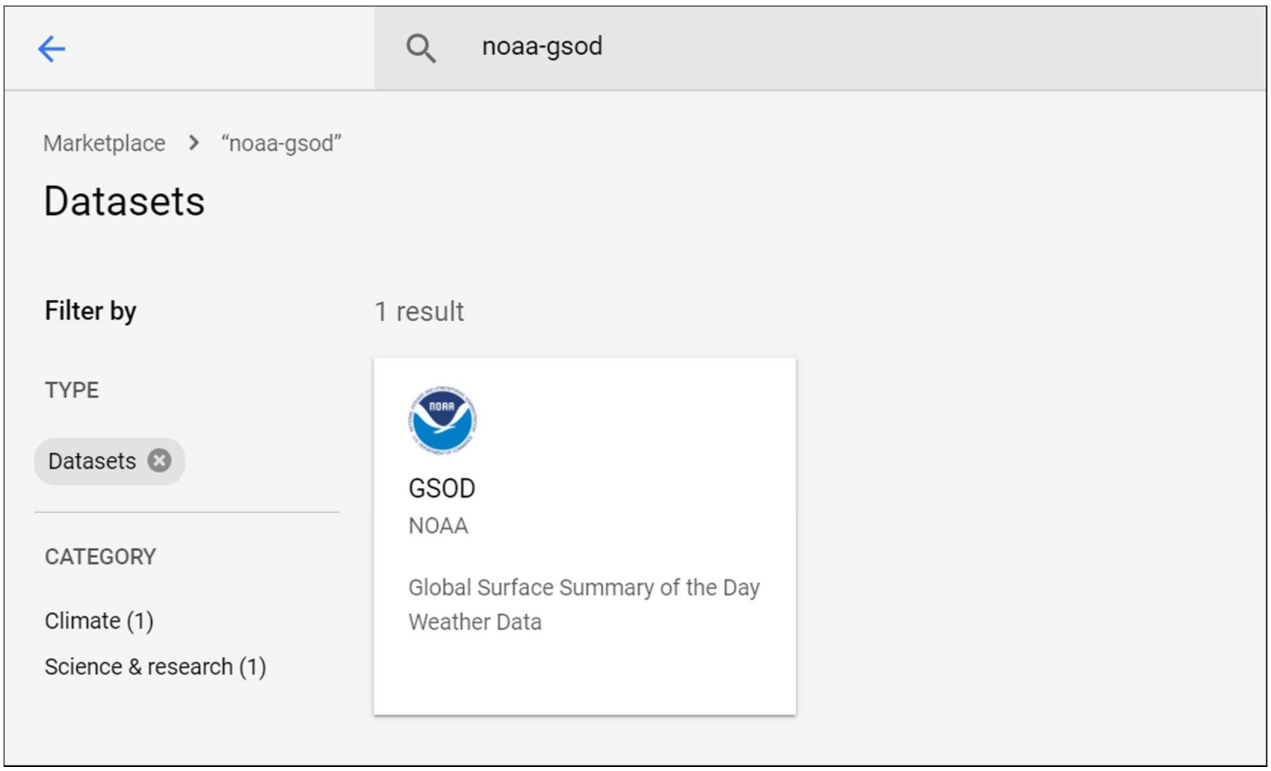
In the left pane, click **ADD DATA** > **Explore public datasets**.



The Datasets window opens.

In the **Search** bar, type "noaa_gsod" then press **Enter**.

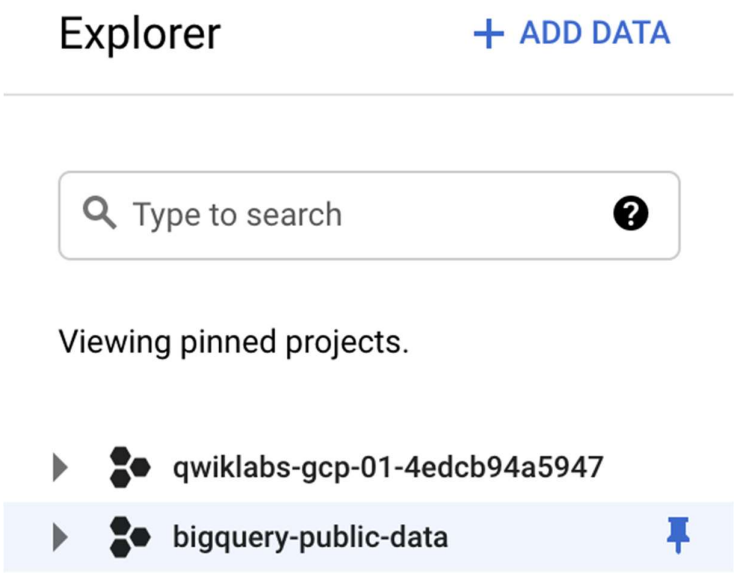
1 result, GSOD dataset, displays.



Click the GSOD dataset and then click **VIEW DATASET**.

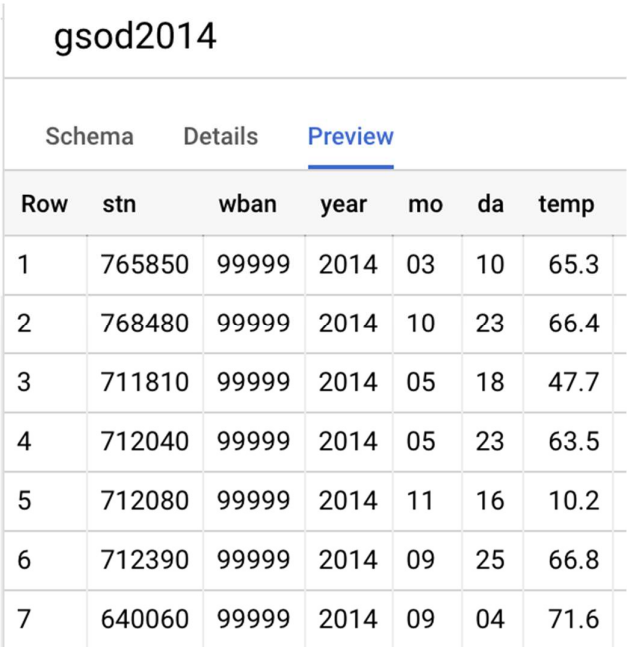
The BigQuery console opens in a new browser tab.To keep your workspace organized, close this new browser tab, go to **Navigation menu** > **BigQuery** in the first tab and refresh the browser.

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**.



In the left pane of the BigQuery console, select **bigquery-public-data** > **noaa_gsod** > **gsod2014** table.

In the Table (gsod2014) window, click the *Preview* tab.



gsod2014

Schema Details Preview

Row	stn	wban	year	mo	da	temp
1	765850	99999	2014	03	10	65.3
2	768480	99999	2014	10	23	66.4
3	711810	99999	2014	05	18	47.7
4	712040	99999	2014	05	23	63.5
5	712080	99999	2014	11	16	10.2
6	712390	99999	2014	09	25	66.8
7	640060	99999	2014	09	04	71.6

Examine the columns and some of the data values.

Paste the following in the Query editor textbox:

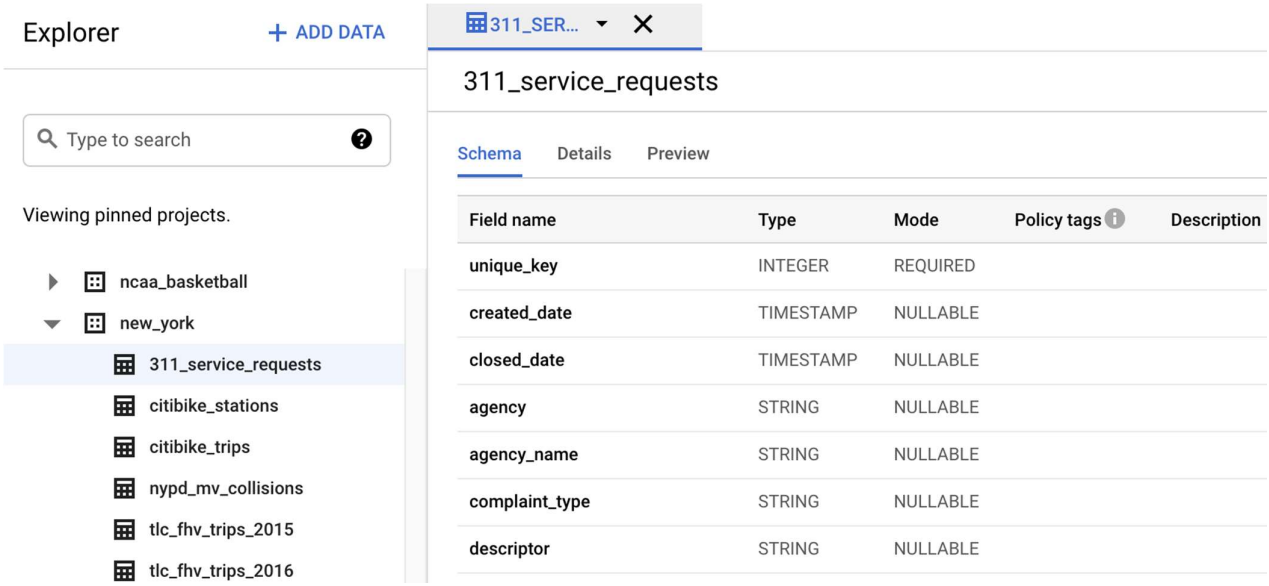
```
SELECT
  -- Create a timestamp from the date components.
  stn,
  TIMESTAMP(CONCAT(year,"-",mo,"-",da)) AS timestamp,
  -- Replace numerical null values with actual null
  AVG(IF (temp=9999.9,
      null,
      temp)) AS temperature,
  AVG(IF (wdsp="999.9",
      null,
      CAST(wdsp AS Float64))) AS wind_speed,
  AVG(IF (prcp=99.99,
      0,
      prcp)) AS precipitation
FROM
  `bigquery-public-data.noaa_gsod.gsod20*`
WHERE
  CAST(YEAR AS INT64) > 2010
  AND CAST(MO AS INT64) = 6
  AND CAST(DA AS INT64) = 12
  AND (stn="725030" OR  -- La Guardia
    stn="744860")    -- JFK
GROUP BY
  stn,
  timestamp
ORDER BY
  timestamp DESC,
  stn ASC
```

Click **Run**. Look at the result and try to determine what this query does.

Click **Check my progress** below to verify you're on track in this lab.

Explore New York citizen complaints data

In the left pane of the BigQuery Console, select the newly added **bigquery-public-data** project and select **new_york** > **311_service_requests**. Then click on the **Preview** tab. Your console should resemble the following:



Examine the columns and some of the data values.

If editor has been closed, click **COMPOSE NEW QUERY** in the upper right.

Paste the following into the Query editor:

```
SELECT
  EXTRACT(YEAR
  FROM
    created_date) AS year,
  complaint_type,
  COUNT(1) AS num_complaints
FROM
  `bigquery-public-data.new_york.311_service_requests`
GROUP BY
  year,
  complaint_type
ORDER BY
  num_complaints DESC
```

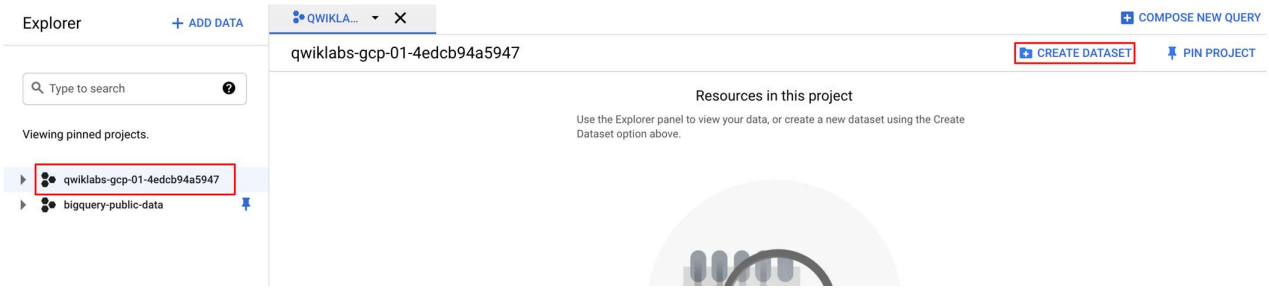
Click **Run**.

Look at the results to determine what the most common complaints are. You will try to determine if these complaints correlate to weather in a later part of this lab.

Click **Check my progress** below to verify you're on track in this lab.

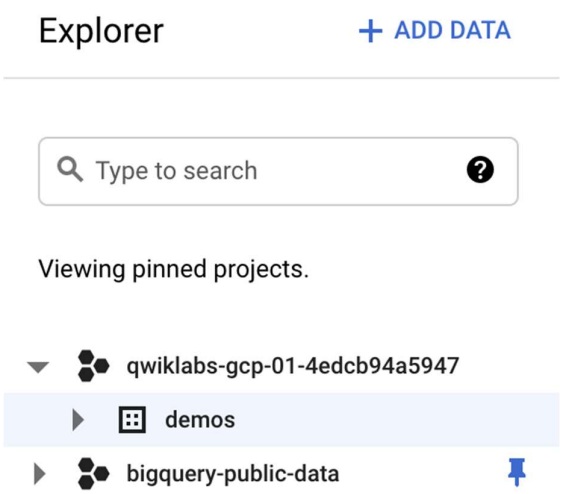
Saving a new table of weather data

In the left pane of the BigQuery Console, select your qwiklabs project and select **CREATE DATASET**.



In the Create dataset dialog, set the **Dataset ID** to "demos" and leave the other options at their default values.

Click **Create dataset**. Your project now has a dataset named "demos"



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

```
SELECT
  -- Create a timestamp from the date components.
  timestamp(concat(year,"-",mo,"-",da)) as timestamp,
  -- Replace numerical null values with actual nulls
  AVG(IF (temp=9999.9, null, temp)) AS temperature,
  AVG(IF (visib=999.9, null, visib)) AS visibility,
  AVG(IF (wdsp="999.9", null, CAST(wdsp AS Float64))) AS wind_speed,
  AVG(IF (gust=999.9, null, gust)) AS wind_gust,
  AVG(IF (prcp=99.99, null, prcp)) AS precipitation,
  AVG(IF (sndp=999.9, null, sndp)) AS snow_depth
FROM
  `bigquery-public-data.noaa_gsod.gsod20*`
WHERE
  CAST(YEAR AS INT64) > 2008
  AND (stn="725030" OR  -- La Guardia
       stn="744860")    -- JFK
```

```
GROUP BY timestamp
```

Along the bottom of the Query editor, click **More** > **Query settings**.

In the Query settings dialog, set the following fields. Leave all others at their default value.

Destination: check **Set a destination table for query results**

Project name: Project ID

Dataset name: **demos**

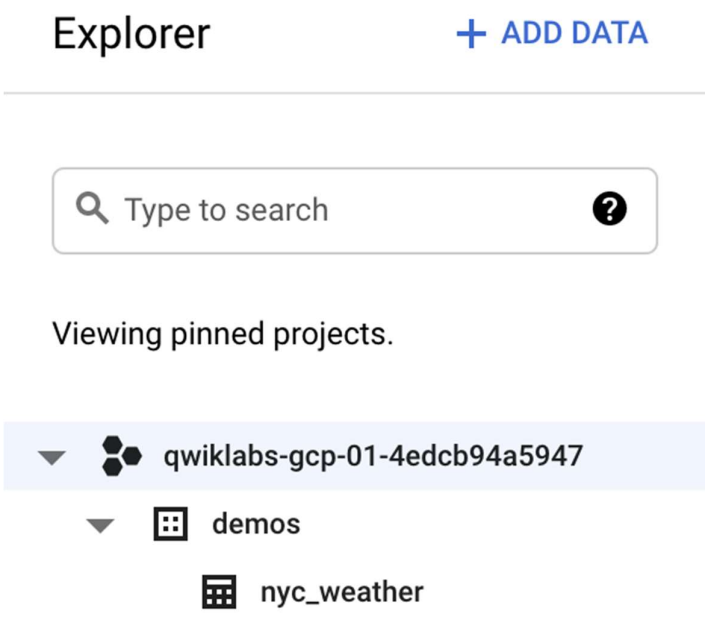
Table name: type **nyc_weather**

Results size: check **Allow large results (no size limit)**

Click **Save**

Click **Run**.

The results are now saved in the dataset you created (demos).



Navigate back to **More** > **Query settings** and, in the *Destination field* select **Save query results in a temporary table**. This removes the demos dataset as a destination for future queries.

Click **Save** to close the query.

Click **Check my progress** below to verify you're on track in this lab.

Find correlation between weather and complaints

Compare the number of complaints and temperature using the CORR function.
Go back to Query editor and run the following query:

```
SELECT
  descriptor,
  sum(complaint_count) as total_complaint_count,
  count(temperature) as data_count,
  ROUND(corr(temperature, avg_count),3) AS corr_count,
  ROUND(corr(temperature, avg_pct_count),3) AS corr_pct
From (
SELECT
  avg(pct_count) as avg_pct_count,
  avg(day_count) as avg_count,
  sum(day_count) as complaint_count,
  descriptor,
  temperature
FROM (
  SELECT
    DATE(timestamp) AS date,
    temperature
  FROM
    demos.nyc_weather) a
  JOIN (
  SELECT x.date, descriptor, day_count, day_count / all_calls_count as pct_count
  FROM
    (SELECT
      DATE(created_date) AS date,
      concat(complaint_type, ": ", descriptor) as descriptor,
      COUNT(*) AS day_count
    FROM
      `bigquery-public-data.new_york.311_service_requests`
    GROUP BY
      date,
      descriptor)x
    JOIN (
      SELECT
        DATE(timestamp) AS date,
        COUNT(*) AS all_calls_count
      FROM `demos.nyc_weather`
      GROUP BY date
    )y
  ON x.date=y.date
)b
ON
  a.date = b.date
GROUP BY
  descriptor,
  temperature
)
GROUP BY descriptor
HAVING
  total_complaint_count > 5000 AND
  ABS(corr_pct) > 0.5 AND
  data_count > 5
ORDER BY
  ABS(corr_pct) DESC
```

Click **Run**.

The results indicate that Heating complaints are negatively correlated with temperature (i.e., more heating calls on cold days) and calls about dead trees are positively correlated with temperature (i.e., more calls on hot days).

Next, compare the number of complaints and wind speed with the CORR function.

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

```
SELECT
  descriptor,
  sum(complaint_count) as total_complaint_count,
  count(wind_speed) as data_count,
  ROUND(corr(wind_speed, avg_count),3) AS corr_count,
  ROUND(corr(wind_speed, avg_pct_count),3) AS corr_pct
From (
SELECT
  avg(pct_count) as avg_pct_count,
  avg(day_count) as avg_count,
  sum(day_count) as complaint_count,
  descriptor,
  wind_speed
FROM (
  SELECT
    DATE(timestamp) AS date,
    wind_speed
  FROM
    demos.nyc_weather) a
  JOIN (
  SELECT x.date, descriptor, day_count, day_count / all_calls_count as pct_count
  FROM
    (SELECT
      DATE(created_date) AS date,
      concat(complaint_type, ": ", descriptor) as descriptor,
      COUNT(*) AS day_count
    FROM
      `bigquery-public-data.new_york.311_service_requests`
    GROUP BY
      date,
      descriptor)x
    JOIN (
      SELECT
        DATE(timestamp) AS date,
        COUNT(*) AS all_calls_count
      FROM `demos.nyc_weather`
      GROUP BY date
    )y
  ON x.date=y.date
)b
ON
  a.date = b.date
GROUP BY
  descriptor,
  wind_speed
)
GROUP BY descriptor
HAVING
  total_complaint_count > 5000 AND
  ABS(corr_pct) > 0.5 AND
  data_count > 5
ORDER BY
  ABS(corr_pct) DESC
```

Notice that the Corr columns are both negative for noise related complaints — do you have a hypothesis for why noise complaints reduce on windy days? Are the coefficients statistically sufficient?

As you can see, BigQuery can give you insights into many different problems from many different angles.

Click **Check my progress** below to verify you're on track in this lab.

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!

Finish your Quest

This self-paced lab is part of the Qwiklabs Scientific Data Processing Quest. A Quest is a series of related labs that form a learning path. Completing this Quest earns you the badge above, to recognize your achievement. You can make your badge (or badges) public and link to them in your online resume or social media account. Enroll in this Quest and get immediate completion credit if you've taken this lab. See other available Qwiklabs Quests.

Take your next lab

Continue your Quest with Distributed Image Processing in Cloud Dataproc, or try one of these:

- Predict Baby Weight with TensorFlow on AI Platform
- Analyzing Natality Data using AI Platform and BigQuery

Next steps / learn more

- For more fun analysis of the NYC data and how it is correlated with weather, see Reto Meier's blog post
- Learn more about BigQuery public data sets.

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 February 11, 2021
Lab Last Tested February 11, 2021