BigQuery Intro Basics Exercise

Doc status: updated, Sep 19 21013

This exercise introduces you to BigQuery, and leads you through submitting queries using the interactive browser and the online tool.

Contents

What You Will Learn

You will learn:

- How to use the BigQuery browser to review schema and submit queries.
- How to use bq tool to access BigQuery from the command line.
- How to import a dataset

What You Will Do

- Query for weather patterns using the public weather dataset in BigQuery, first using the online interactive browser, and then using the bq tool.
- Import and query earthquake data

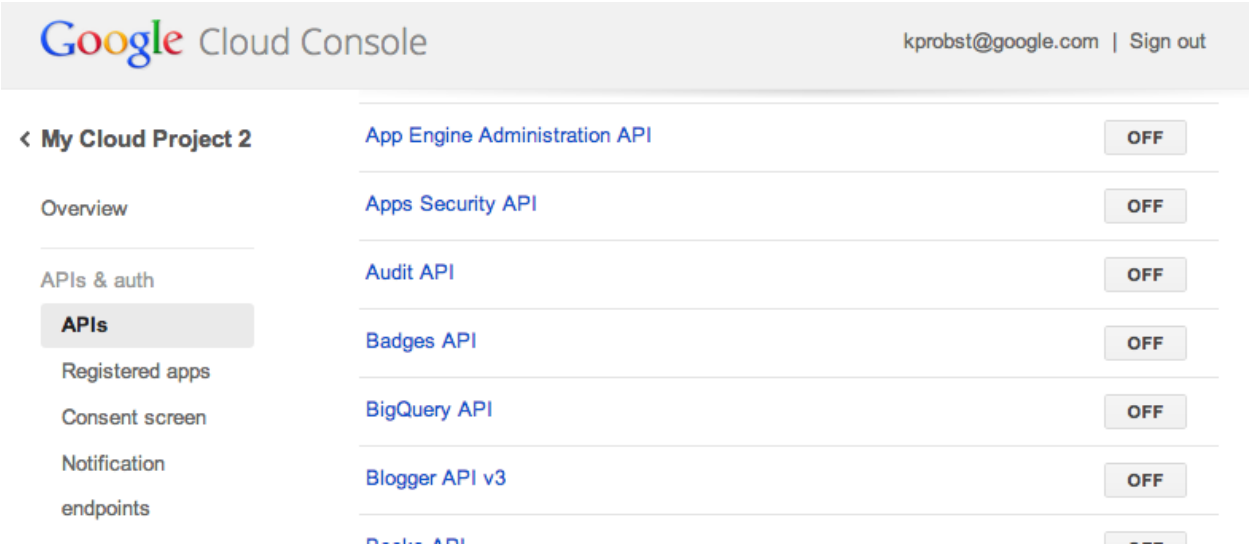
Get Ready

For this exercise, you will work with existing datasets in BigQuery. You do not need an App Engine application.

You need to have a Google APIs project and enable BigQuery API

Enable BigQuery in your Google Cloud Project

1. Go to the Google Cloud Console. Select your project or create a new one.
2. In the left hand navigation panel, select Services.
3. In the Services screen on the right, select the Google Cloud Platform tab.
4. Switch ON the BigQuery API.



5. If the Terms of Service appear, read and accept them.
6. If you have not already enabled billing for your project, enable it now.

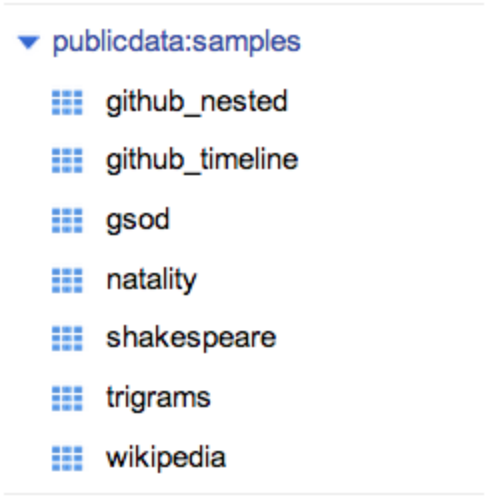
Use the BigQuery Interactive Browser

You will use the BigQuery browser to explore historical weather data. BigQuery provides a public sample dataset containing “global summary of the day” (GSOD) weather data. The data is collected from weather stations around the world.

> ***Note***: The public sample GSOD dataset provided in BigQuery is a snapshot of GSOD data, and is intended for learning purposes only; it is not guaranteed to be complete or up to date. Do not use it in a production application!

View the GSOD Sample Dataset

1. In a browser, go to the BigQuery browser at http://bigquery.cloud.google.com/.
2. Make sure the BigQuery browser is displaying the correct project.
3. In the BigQuery browser, expand **publicdata:samples**.

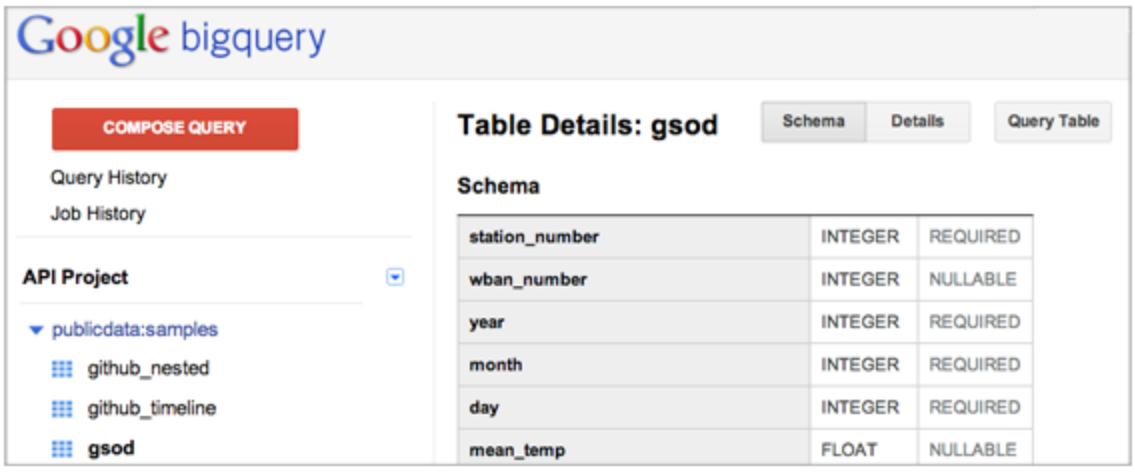


You will see the list of datasets that are available to the public.

Note: These sample datasets are provided to help people investigate BigQuery. The datasets are not guaranteed to be up to date.

4. Select **gsod**.

The schema for the GSOD (Global Summary of Day) table appears on the right.



5. Click the **Details** button on top right to see the summary details for the dataset.

6. Select the **Query Table** button on the top right.

7. Run the following query to get an idea of the contents of the table:

```
SELECT * FROM [publicdata:samples.gsod] LIMIT 10
```

Submit Queries About Historical Weather Data

In this section you will submit queries to answer questions about weather. Feel free to experiment and make up your own queries too. In case you need it, here are the station_numbers for various cities:

- San Francisco (SFO Airport) 724940
- Chicago O'Hare airport 725340
- Paris Orly airport 071490
- London (Heathrow Airport) 037720 c
- Tokyo International Airport 476710

If you want to find more station numbers, see http://berkeleyearth.lbl.gov/station-list/, and drill down to individual pages to find the GSOD number.

For help with query syntax, see the BigQuery Query Reference.

> ***Note***: At any time to view the schema, select the dataset name (in this case, **gsod**) from the list on the left.

Feel free to submit your own queries, or work through the queries listed below.

1. List the days that San Francisco Airport has had fog since 2000. (station_number 724940)

```
SELECT
  fog,
  year,
  month,
  day
FROM
  [publicdata:samples.gsod]
WHERE
  station_number = 724940
  AND year >= 2000
  AND fog=TRUE
ORDER BY
  year ASC
LIMIT
  10
```

2. How many foggy days has San Francisco Airport had since 2000? (staton_number 72490)

```
SELECT
  count(*)
FROM
  [publicdata:samples.gsod]
WHERE
  station_number = 724940
  AND year >= 2000
  AND fog=TRUE
```

3. How many times since 2000 has San Francisco Airport had fog on today's date (station_number 72490)?

(Substitue the appropriate values for month and day)

```
SELECT
  count (fog)
FROM
  [publicdata:samples.gsod]
WHERE
  station_number = 724940
  AND year >= 2000
  AND month = XX
  AND day=XX
  AND fog=TRUE
```

4. Count the number of foggy days per month in San Francisco since 2000, and order by the foggiest month.

```
SELECT
  year, month, count(day) fog_days
FROM
  [publicdata:samples.gsod]
WHERE
  station_number = 724940
  AND year >= 2000
  AND fog=TRUE
GROUP BY
  year, month
ORDER BY
  fog_days desc, year, month
LIMIT
  10
```

5. What was the average mean wind speed, maximum gust wind speed, and maximum sustained wind speed per month in Chicago O'Hare airport in 2000? (station_number 725340)

```
SELECT
 year, month,
  ROUND(AVG(mean_wind_speed)) mean_wind_speed,
  ROUND (MAX(max_gust_wind_speed)) max_gust_wind_speed,
  ROUND (MAX(max_sustained_wind_speed)) max_sustained_wind_speed
FROM
  [publicdata:samples.gsod]
WHERE
  station_number = 725340
  AND year =2000
GROUP BY
  year, month
ORDER BY
  year, month ASC
LIMIT
  10
```

6. Enough of the fog and wind! Which months on which years did London Heathrow airport have the highest mean temperature? (station_number 037720)

```
SELECT
 MAX (ROUND (mean_temp)) mean_temp, year, month, day
FROM
  [publicdata:samples.gsod]
WHERE
  station_number = 037720
 GROUP BY
 year, month, day
 ORDER BY
 mean_temp DESC, year DESC, month DESC, day DESC
LIMIT 10
```

Use the bq tool to query datasets

An alternative to using the BigQuery online browser to interrogate datasets is to use the bq command line tool.

1. Install the bq tool if necessary.

1. At the command line, check if the bq tool is installed by running bq. If it is installed, you will see the help for the bq tool.
2. If it is not installed, install it now from https://cloud.google.com/sdk/ .
3. List tables in the public datasets to check the bq tool is working

   ```
   bq ls publicdata:samples
   ```

4. You might need to grant authorization access. If so, follow the instructions.

2. Set the default project.

The first time you run the bq tool, it asks for the default project. Thereafter, it uses that project.

One way to check the default project is to go into the bq shell by entering:

```
bq shell
```

The prompt shows the default project.bq

To exit the bq shell, type exit.

To change the default project, edit ~/.bigqueryrc

> **Note**: need to run gcloud auth init to switch projects... Otherwise the bq command will not pick up the .bigqueryrc file...

3. Run some queries.

- Show tables in the GSOD public sample dataset

  ```
  bq show publicdata:samples.gsod
  ```

- What are the most recent days in the gsod public data that show both rain and thunder at London Heathrow airport? station_number is 037720. Limit the results to 10

  ```
  bq query "SELECT year, month, day, thunder, rain, FROM [publicdata:samples.gsod] WHERE thunder=TRUE AND rain=TRUE AND station_number=037720 ORDER BY year DESC, month DESC, day DESC LIMIT 10"
  ```

4. Run the bq shell:

```
bq shell
```

In the shell, you can run bq commands without needing to prefix bq to the command.

Try it by submitting the following query to find which months on which years in Paris had the most

days where the mean temperature exceeded 80 (station_number for Paris Orly airport is 071490).

```
SELECT year, month, count(round(mean_temp)) num_hot_days FROM
[publicdata:samples.gsod] WHERE station_number = 071490 AND mean_temp > 83
GROUP BY year, month ORDER BY num_hot_days DESC, year DESC, month DESC
LIMIT 10
```

Do the results correlate to heat waves in Paris?

To exit the shell, type exit.

Import Your Own Dataset

So far you have used public datasets. The next step is to import a dataset into BigQuery.

For this part of the exercise, you will use earthquake data.

BigQuery can import:

- CSV files
- JSON files
- Datastore backup files

from:

- local files
- Google Cloud Storage.

In this exercise, you will import the data from a local CSV file.

First you need to get the data to import.

Get the Earthquake Data

1. In a browser, go to http://earthquake.usgs.gov/earthquakes/map/.

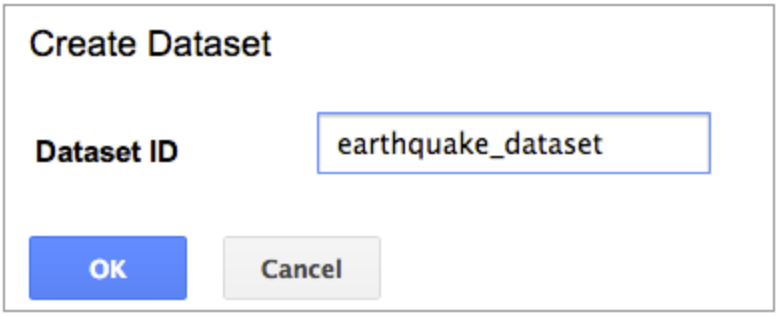
2. In the Setting section, choose a range of earthquakes (such as **30 days, Magnitude 2.5+, Worldwide**)
3. In the top left, click the Download link, then choose **CSV**.
4. Open the downloaded file and take a quick look at it it, to get an idea of the data it contains. Review the column headers -- you will use these labels when specifying the schema in BigQuery.

Import the Data into BigQuery

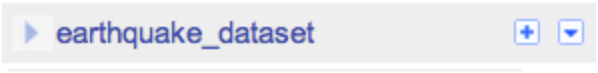
1. In a browser, go to http://bigquery.cloud.google.com/.
2. Make sure the BigQuery browser is displaying the correct project.

If not, expand the drop down menu next to the project name, select Switch to Project, and select the appropriate project.

3. Expand the drop down menu next to the project name, and select Create New Dataset.
4. Enter the ID for the new dataset (which is whatever you want), such as earthquake_dataset. Press OK.



5. In the list of datasets, select earthquake_dataset. Press the **+** button to start the process of adding a table to the dataset.

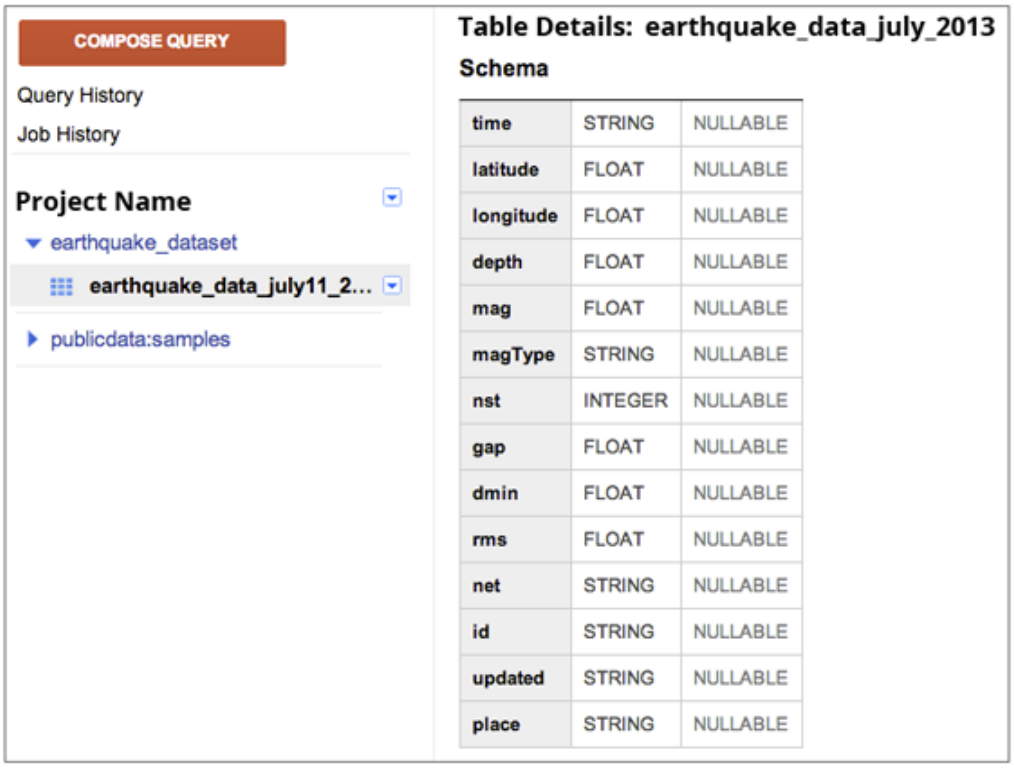


6. Step through the Create and Import Dialog box.
   - Choose Destination:
     - The data you downloaded contains data for the past 30 days, so give the table an ID something like earthquake_data_july11_2013 (substituting today's date).
   - Select Data:
     - The source format is CSV.
     - Upload the file you downloaded.
   - Specify Schema:
     - Enter the schema as: time:string,latitude:float,longitude:float,depth:float,mag:float,magType:string,nst:integer,gap:float,dmin:float,rms:float,net:string,id:string,updated:string,place:string,type:string
   - Advanced Options:
     - Field Delimiter is comma.
     - Skip 1 header row.
     - Allow 0 errors.
   - Submit

     After you submit the new table, you see table listed. However, you cannot click the new table to view the schema until it has finished loading.

     Note: it may take a few minutes for the data to upload. A faster way to upload data is to first upload the data file to Google Cloud Storage, and from there upload to BigQuery.

7. When the data has finished loading, click the table name to view the schema, as shown here:



8. Query the data. Where and when was the highest magnitude earthquake, and how deep was it?

9. Feel free to submit more queries.

Summary

In this chapter you learned how to use Google Big Query interactively and from the command line. You also learned how to create your own dataset.

Resources

- Google BigQuery Developer Documentation
- BigQuery Query Reference

bq Tool

- Documentation
- bq Tool Tutorial