Google Cloud Platform

About COVID-19 Public Datasets

BigQuery Public Datasets Program

Getting started with COVID-19 Public Datasets

VIEW DATASET

Type
Datasets

Last updated
3/31/20, 4:21 PM

Category
Healthcare
Science & research

Dataset source
Various

Cloud service
BigQuery

Region
US/EU

Update frequency
Daily

Overview

In an effort to help combat COVID-19, we created a COVID-19 Public Datasets program to make data more accessible to researchers, data scientists and analysts. The program will host a repository of public datasets that relate to the COVID-19 crisis and make them free to access and analyze. These include the Johns Hopkins Center for Systems Science and Engineering (JHU CSSE) dataset, Global Health Data from the World Bank, and OpenStreetMap data.

Free hosting and queries of COVID datasets
As with all data in the Google Cloud Public Datasets Program ↗, Google pays for storage of datasets in the program. BigQuery also provides free queries over certain COVID-related datasets to support the response to COVID-19. Queries on COVID datasets will not count against the BigQuery sandbox free tier ↗, where you can query up to 1TB free each month.

Limitations and duration
Queries of COVID data are free. If, during your analysis, you join COVID datasets with non-COVID datasets, the bytes processed in the non-COVID datasets will be counted against the free tier, then charged accordingly, to prevent abuse. Queries of COVID datasets will remain free until Sept 15, 2020.

The contents of these datasets are provided to the public strictly for educational and research purposes only. We are not onboarding or managing PHI or PII data as part of the COVID-19 Public Dataset Program. Google has

practices & policies ⤢ in place to ensure that data is handled in accordance with widely recognized patient privacy and data security policies.

Issues/questions
For questions or issues, contact us at bq-public-data@google.com

See here for all datasets ⤢ included in the program

Samples

<u>More Information - Getting Started - FAQ</u>

How do I get started using COVID-19 Public Datasets?
The public datasets available as part of this program are available for query in BigQuery by default. There's no need to sign up or register for the program specifically. Just log on to BigQuery (see Sandbox below if you are new) and run your query. Each of the datasets have sample queries linked in the description so you can start there.

I'm new to BigQuery. How can I get started quickly?
BigQuery's Sandbox ⤢ allows a new user to store up to 10 GB of data and query up to 1 TB of data each month for free without having to provide a credit card. This can be useful for citizen data scientists and researchers from organizations that don't have agreements with Google Cloud. Since queries over COVID-related datasets will be free, this is a quick and efficient way to get started.

What is the difference between a COVID-19 public dataset and a normal public dataset?
Google Cloud has an extensive Cloud Public Datasets Program ⤢ that launched over four years ago. Datasets listed as COVID-19 public datasets are hosted by Google so that customers can more easily discover and access them as part of that program. However, we have taken the additional step of making queries over the COVID-19 datasets free until September 15th to support response to the COVID-19 pandemic. To see which datasets are part of the COVID-19 public datasets, see the list here ⤢.

If queries over COVID-19 datasets are free, why does the UI still tell me how many bytes the query will process?
We've left this feature in place because users find this helpful to help to optimize performance. So you can still see the bytes processed in your queries even though you will not be billed for bytes processed on datasets that are part of the COVID-19 Public Datasets Program. BigQuery reports two different numbers in the query statistics: the bytes processed and the bytes billed. For the most part, these numbers are the same. However, for queries against COVID-19 Public Datasets, the bytes billed will be lower, since they will not include any bytes read from those tables. To see the actual bytes billed, click on "Job Information" on the query results pane. It will show both the bytes scanned and bytes billed metrics.

I'm a researcher looking to do my work on Google Cloud. Are credits available for researchers?

Google Cloud offers credits for qualifying researchers. Please take a look here to see if you qualify and at the FAQ for instructions on how to apply.

How do I tell if a dataset is part of the COVID-19 Public Dataset Program?

All datasets in the COVID-19 public dataset program have a "COVID-19" label. These labels are searchable in Marketplace, the BigQuery UI, and a full list is available here ⤢.

How do I nominate a dataset for inclusion in the Covid-19 Public Datasets Program?

To nominate a dataset for inclusion please file a ticket in our Public Issue Tracker here ⤢.

I submitted a dataset but it requires engineering. How do I get engineering help to onboard a dataset?

After you submit the request in the public issue tracker, if the dataset gets pre-selected for inclusion in the program, someone from the Public Datasets program will reach out to you to discuss options for ingesting the dataset.

Where can I find a list of datasets that are included in this program?

For a full list of all datasets included in the COVID-19 Public Datasets Program, see the Marketplace listing ⤢. From the BigQuery UI, if you have the bigquery-public-datasets project pinned, you can search for the "freebqcovid" tag, and it will filter the results to just include COVID-19 Public Datasets.

Contact us

To nominate a dataset for inclusion please file a ticket in our Public Issue Tracker here ⤢. For other questions or issues, contact us at bq-public-data@google.com

Terms of service

Each dataset included in the COVID-19 Public Datasets Program has its own licensing terms. Please see the page for the datasets you intend to use for the terms of use. These datasets are provided "as is" without any warranties or representations of any kind. You are responsible for determining the suitability of this data for your purposes.