Introduction - What is BigQuery?

Google BigQuery is a great example of a modern cloud data warehouse that many companies today will use to store big data. It can interactively retrieve and run analysis on petabytes of data within seconds thanks to its technology behind the scenes.

Specifically, Google BigQuery is serverless, you can just make an account and have access to its powerful engine through the internet (what we call the “cloud”). You can upload your own data, or analyze public default datasets that Google puts on there and all of this is stored on the cloud.

This is in contrast to traditional databases which had infrastructure/server requirements to service locally. If you were working for a company that used BigQuery, you would see the same user interface that you would through a personal account. If the company uses other cloud data warehouse services, they would all have roughly the same features.

Behind the scenes, BigQuery leverages multiple clusters of thousands of machines behind the scenes to make all of this processing possible. These are hidden from the user, you won’t have to worry about these and can just focus on analysis!

BigQuery is only one part of the Google Cloud Platform, there are other services such as Google Cloud Storage, Google Datastore which have other properties which venture outside the realm of relational databases and the scope of this course.

Relational Databases & SQL

A relational database is just a database made of tables (which could have anywhere from one row to billions of rows), which can be linked to one another through columns they have in common. Most companies one would work at as a Data Analyst, Data Scientist or Data Engineer would have at least one database which follows this design model (if not all of them).

It is on this platform that we will learn Structured Query Language (or SQL). Like any real life human language, it can have many dialects but they are all similar to one another so learning it on BigQuery will allow you to apply it to all relational databases.

The reason BigQuery was chosen for this course is because many companies use it so learning it here mimics an actual work environment and the power of the SQL engine allows us to perform tasks on big data that would be otherwise impossible in say a spreadsheet.

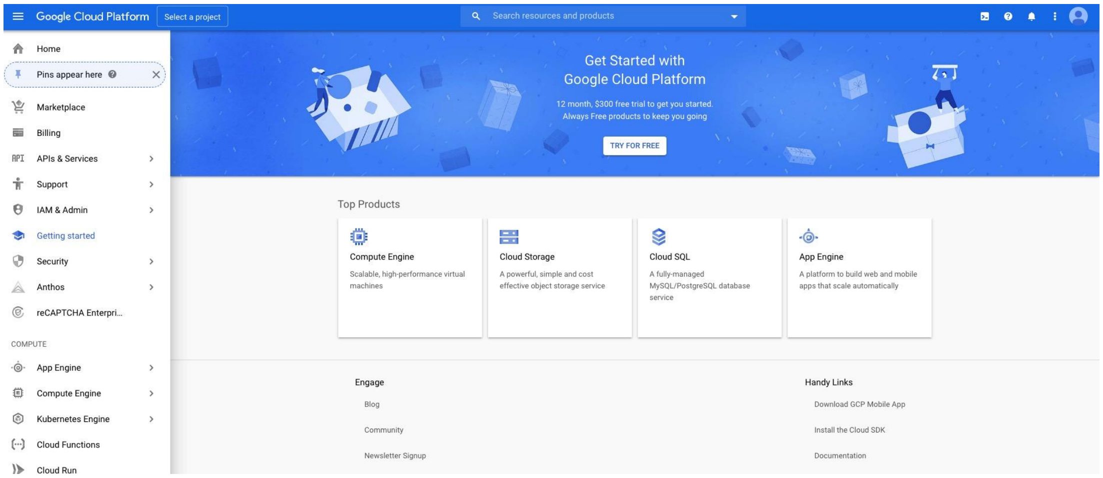
Setting Up Environment

Let us first set up a sandbox user account environment.

- Go to https://cloud.google.com/bigquery/docs/sandbox and click on “Go to the Cloud Console”

- Sign in with a google account, if you do not have one you can make one by making a gmail account very easily here https://www.google.com/gmail/about/

- You should see something similar to the following.



12 month, $300 free trial to get you started.
Always Free products to keep you going

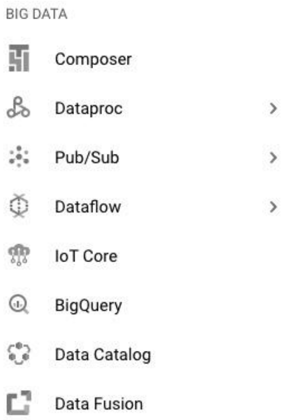
Scalable, high-performance virtual machines

A powerful, simple and cost effective object storage service

A fully-managed MySQL/PostgreSQL database service

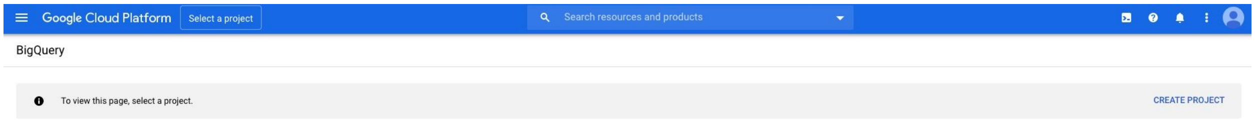
A platform to build web and mobile apps that scale automatically

- Go to the left hand panel and scroll down till you see the following under “Big Data”



- Click on BigQuery (Protip: click on the pin button beside BigQuery so that it is the first thing that appears in the left hand menu)

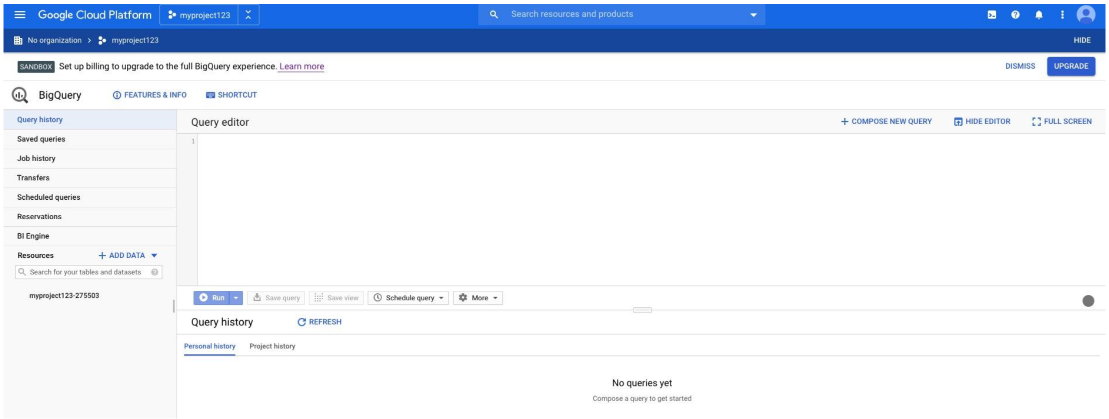
- You should now see something like this click on “CREATE PROJECT”



To view this page, select a project.

- Feel free to name the it whatever you like and then click “CREATE”

- You should now be taken to the UI feel free to bookmark this page! It should look like this. Notice the left hand column underneath “Resources” the name of your project will appear here (this is the resource tree).



Set up billing to upgrade to the full BigQuery experience. Learn more

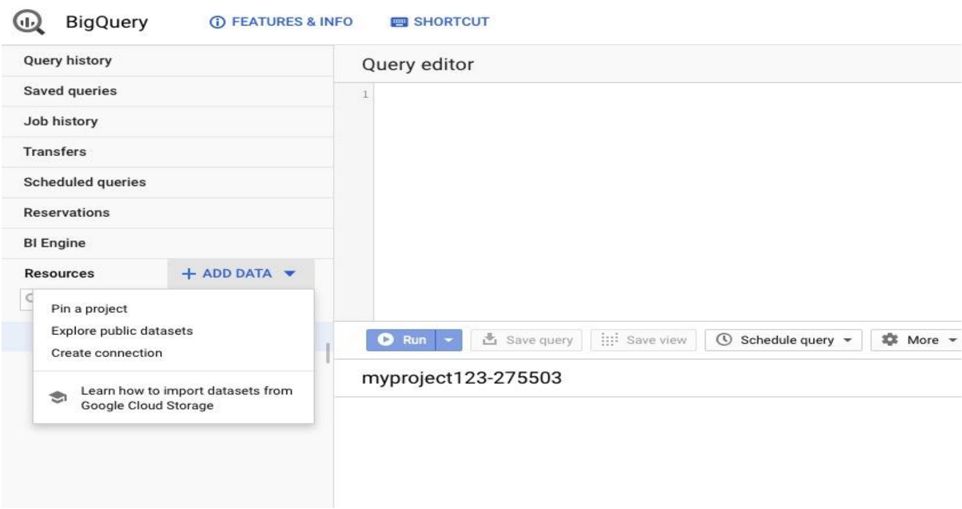
No queries yet

Compose a query to get started

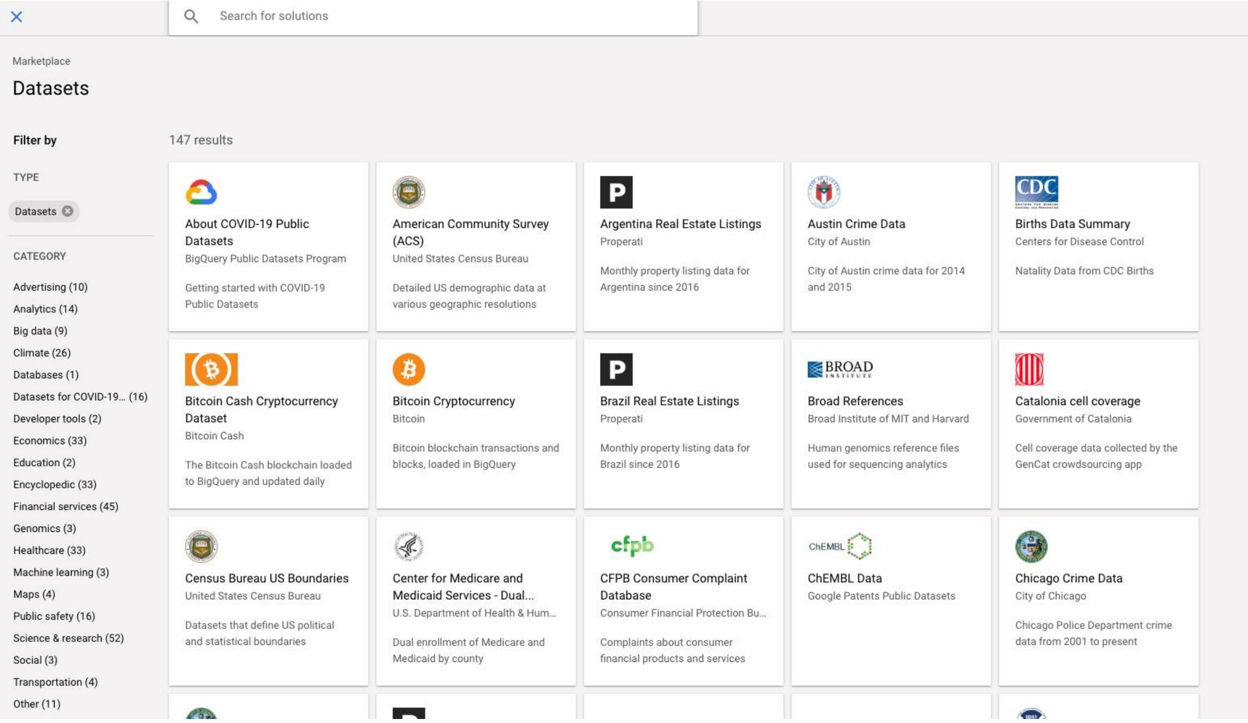
Public Datasets

Let’s populate our environment with some public data now.

In the left hand panel click the “ADD DATA” button. A menu will drop down, click explore public datasets



A popup will appear with numerous selections. Feel free to click on any that interests you but it won't matter too much



Search for solutions

Marketplace

Datasets

Filter by

TYPE

Datasets

CATEGORY

Advertising (10)
Analytics (14)
Big data (9)
Climate (26)
Databases (1)
Datasets for COVID-19... (16)
Developer tools (2)
Economics (33)
Education (2)
Encyclopedic (33)
Financial services (45)
Genomics (3)
Healthcare (33)
Machine learning (3)
Maps (4)
Public safety (16)
Science & research (52)
Social (3)
Transportation (4)
Other (11)

147 results

About COVID-19 Public Datasets
BigQuery Public Datasets Program
Getting started with COVID-19 Public Datasets

American Community Survey (ACS)
United States Census Bureau
Detailed US demographic data at various geographic resolutions

Argentina Real Estate Listings
Properati
Monthly property listing data for Argentina since 2016

Austin Crime Data
City of Austin
City of Austin crime data for 2014 and 2015

Births Data Summary
Centers for Disease Control
Natality Data from CDC Births

Bitcoin Cash Cryptocurrency Dataset
Bitcoin Cash
The Bitcoin Cash blockchain loaded to BigQuery and updated daily

Bitcoin Cryptocurrency
Bitcoin
Bitcoin blockchain transactions and blocks, loaded in BigQuery

Brazil Real Estate Listings
Properati
Monthly property listing data for Brazil since 2016

Broad References
Broad Institute of MIT and Harvard
Human genomics reference files used for sequencing analytics

Catalonia cell coverage
Government of Catalonia
Cell coverage data collected by the GenCat crowdsourcing app

Census Bureau US Boundaries
United States Census Bureau
Datasets that define US political and statistical boundaries

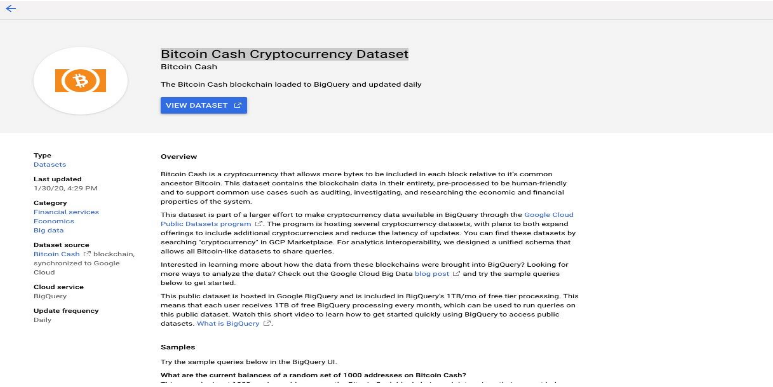
Center for Medicare and Medicaid Services - Dual...
U.S. Department of Health & Hum...
Dual enrollment of Medicare and Medicaid by county

CFPB Consumer Complaint Database
Consumer Financial Protection Bu...
Complaints about consumer financial products and services

ChEMBL Data
Google Patents Public Datasets

Chicago Crime Data
City of Chicago
Chicago Police Department crime data from 2001 to present

For the purposes of this demonstration, we will select "Bitcoin Cash Cryptocurrency Dataset"



Bitcoin Cash Cryptocurrency Dataset
Bitcoin Cash

The Bitcoin Cash blockchain loaded to BigQuery and updated daily

VIEW DATASET

Type
Datasets

Last updated
1/30/20, 4:29 PM

Category
Financial services
Economics
Big data

Dataset source
Bitcoin Cash blockchain, synchronized to Google Cloud

Cloud service
BigQuery

Update frequency
Daily

Overview

Bitcoin Cash is a cryptocurrency that allows more bytes to be included in each block relative to it's common ancestor Bitcoin. This dataset contains the blockchain data in their entirety, pre-processed to be human-friendly and to support common use cases such as auditing, investigating, and researching the economic and financial properties of the system.

This dataset is part of a larger effort to make cryptocurrency data available in BigQuery through the Google Cloud Public Datasets program. The program is hosting several cryptocurrency datasets, with plans to both expand offerings to include additional cryptocurrencies and reduce the latency of updates. You can find these datasets by searching "cryptocurrency" in GCP Marketplace. For analytics interoperability, we designed a unified schema that allows all Bitcoin-like datasets to share queries.

Interested in learning more about how the data from these blockchains were brought into BigQuery? Looking for more ways to analyze the data? Check out the Google Cloud Big Data blog post and try the sample queries below to get started.

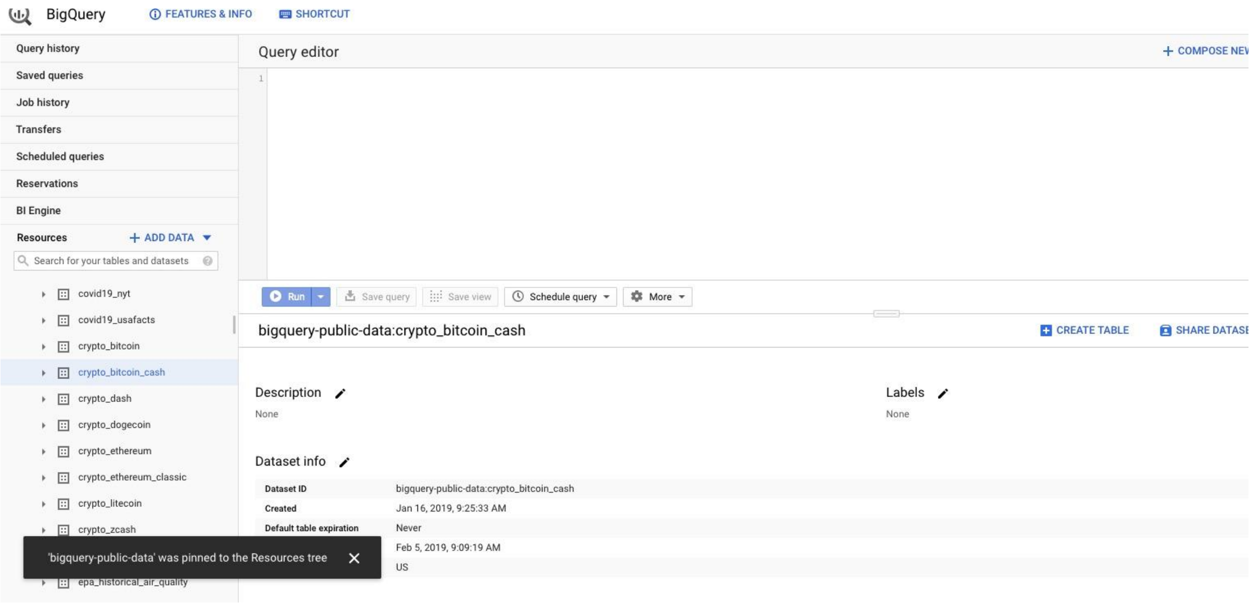
This public dataset is hosted in Google BigQuery and is included in BigQuery's 1TB/mo of free tier processing. This means that each user receives 1TB of free BigQuery processing every month, which can be used to run queries on this public dataset. Watch this short video to learn how to get started quickly using BigQuery to access public datasets. What is BigQuery.

Samples

Try the sample queries below in the BigQuery UI.

What are the current balances of a random set of 1000 addresses on Bitcoin Cash?

Click "View Dataset". This should take you back to the BigQuery UI. The resources tree should now have the selected dataset with a whole bunch of other public datasets (they have all been added as a new project called "bigquery-public-data"). Note the message that says the Public Datasets have been pinned now. This is good, you will be able to easily access these under the resource tree each time you login now.



bigquery-public-data:crypto_bitcoin_cash

Description

None

Labels

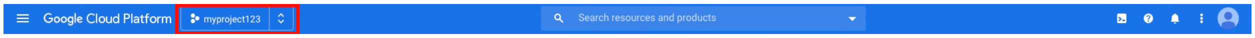
None

Dataset info

Dataset ID	bigquery-public-data:crypto_bitcoin_cash
Created	Jan 16, 2019, 9:25:33 AM
Default table expiration	Never
	Feb 5, 2019, 9:09:19 AM
	US

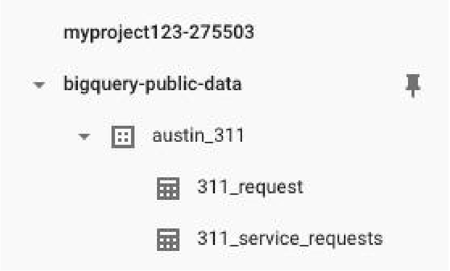
'bigquery-public-data' was pinned to the Resources tree

Note: If for some reason this doesn't work make sure you have a project selected in the very top bar on the UI



Similar to a file structure on your computer or laptop. The BigQuery Web UI Resources tree and most database systems will be organized under the following levels, by **Project → Dataset → Table.** A project can have many datasets, a dataset can have many tables. Tables are where you can actually see the data and where it is "stored". Tables from different datasets or projects can all interact with each other, more on this later. Likewise, tables can move or be saved to and from different projects or datasets.

The layout from project to dataset to table can be seen here



Note that the project you created while setting up the sandbox account is still here; there just isn't any data saved to it yet. We are only adding the "bigquery-public-data" to our environment right now.