Getting Started with GCP and BigQuery

Adam Stine

U.S. National Library of Medicine
National Center for Biotechnology Information

Overview

- Initial Google Cloud Setup
- Make a Project
- Billing and Spending Alerts
- Service Account Credentials

Initial Setup

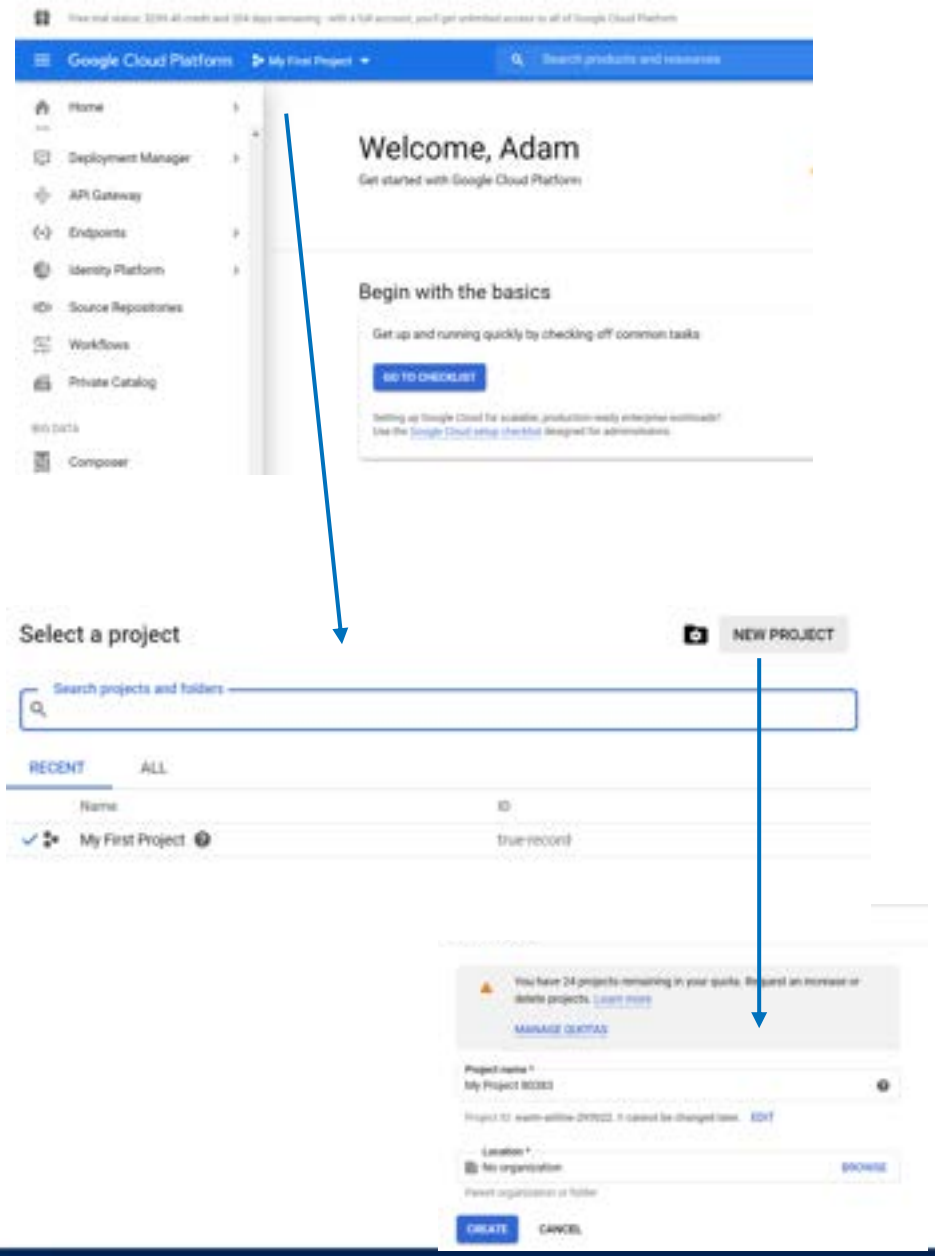
Much of this will depend on whether you are running a personally administered account or an institutional account.

We'll cover some basics in these slides to get you started but the cloud documentation from the service provider will be a more complete resource.

You can experiment with your own Google Cloud account by entering the necessary billing information or you may have access to the BigQuery sandbox or credits/discounts as part of NIH STRIDES Initiative.

Create a Project

- There are multiple optional layers of resource hierarchy in Google Cloud but Project is the most basic.
- Projects enable billing, managing permission for resources, adding and removing collaborators, and more.
- A project is required to use Google Cloud resources.



Projects

- You will need to create a project if you plan to use a personal GCP account. Institutional users will likely be added to an existing project.
- This guide in the GCP documentation should help you create and manage projects if you need to. https://cloud.google.com/resource-manager/docs/creating-managing-projects

Set Spending Alerts

Cost is a frequent concern when moving from local to cloud.

It is certainly possible to generate large bills when using cloud services from the major providers but there are quite a few trial credits and “free tier” services that allow for learning and experimenting with very low expense.

GCP billing has tools to set alerts for various spending levels either projected or actual costs spent to prevent surprises.

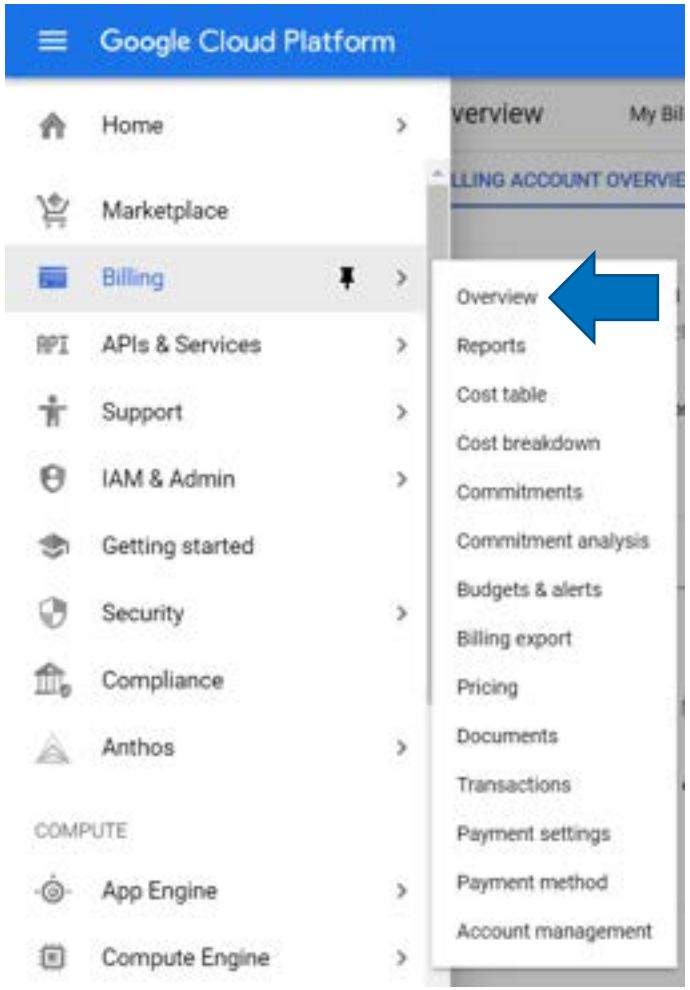
There is also a robust monitoring system to track spending

Billing

The Billing section of the console provides many views and tools to manage and track spending.

For institutional accounts some of these features may not be available to you.

The Billing Overview is a good place to start to talk about costs.

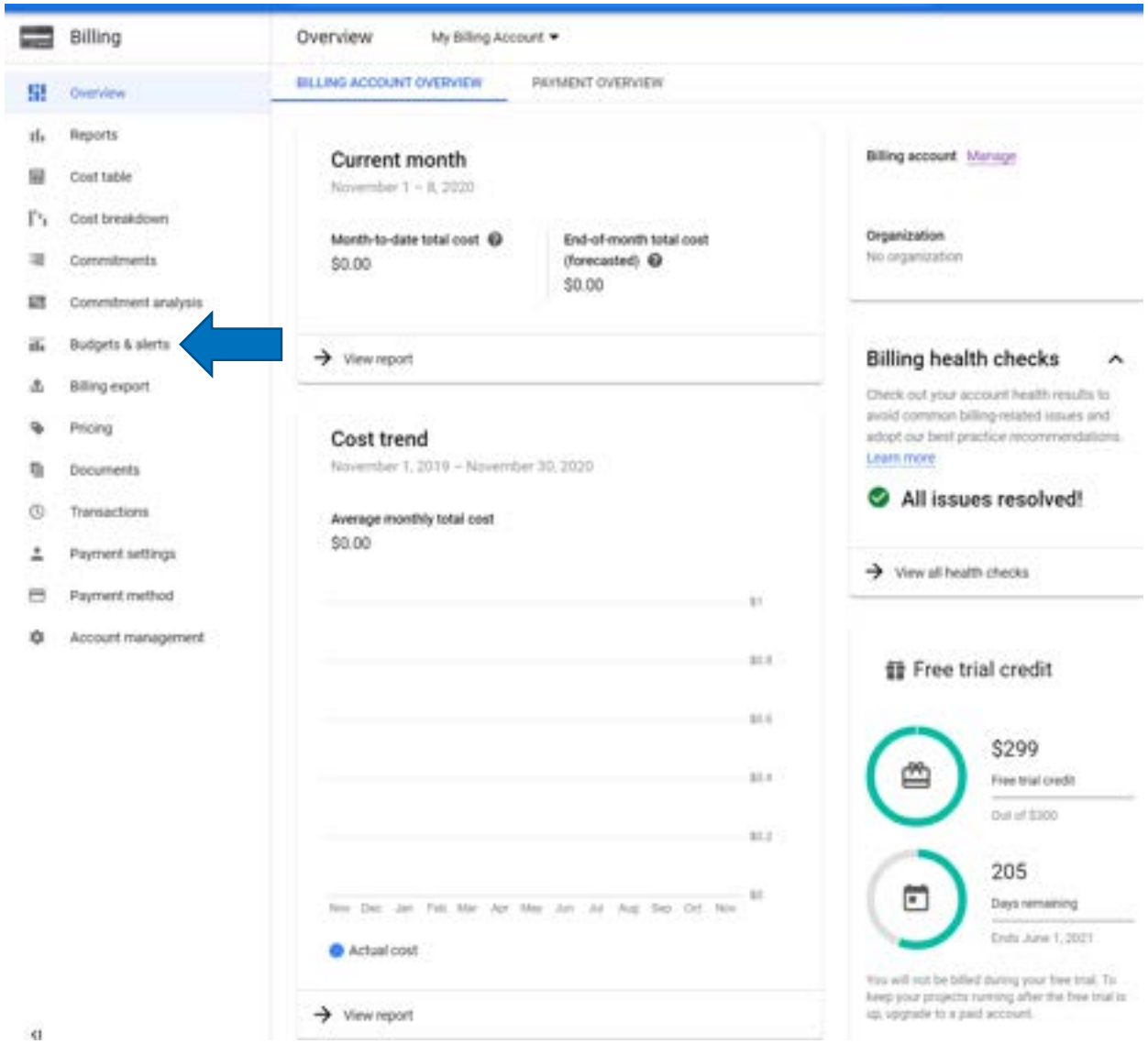


Billing Overview

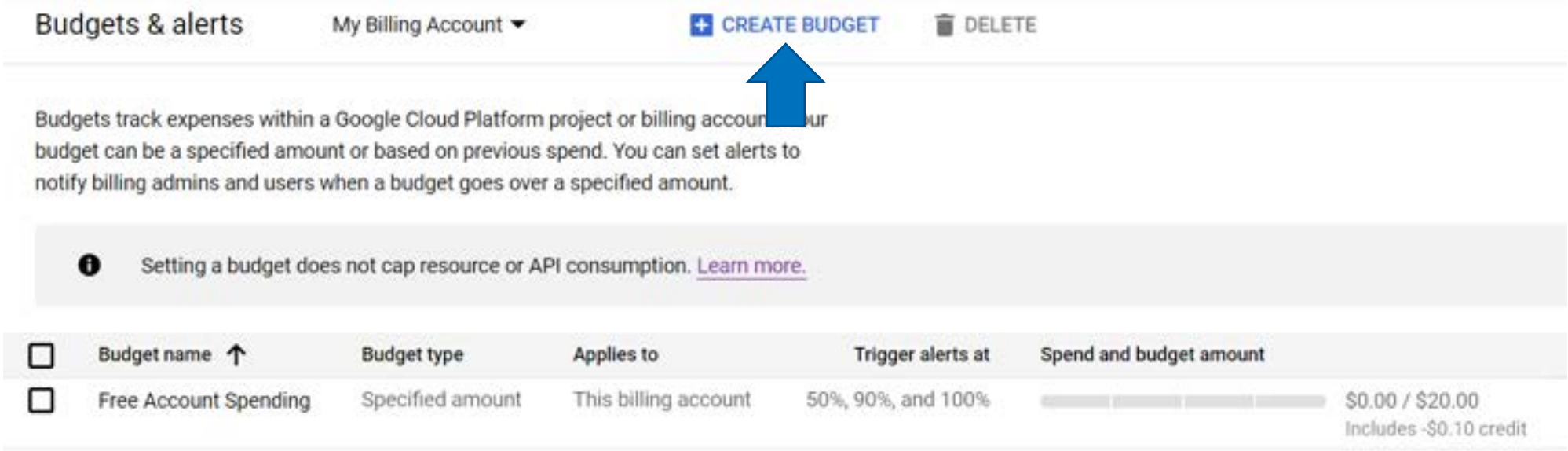
The Overview page shows current and projected usage for the billing period.

For personal accounts like the one shown it also provides information about the remaining trial credit.

We can set some spending alerts by clicking Budgets & alerts on the menu.



Budgets & Alerts



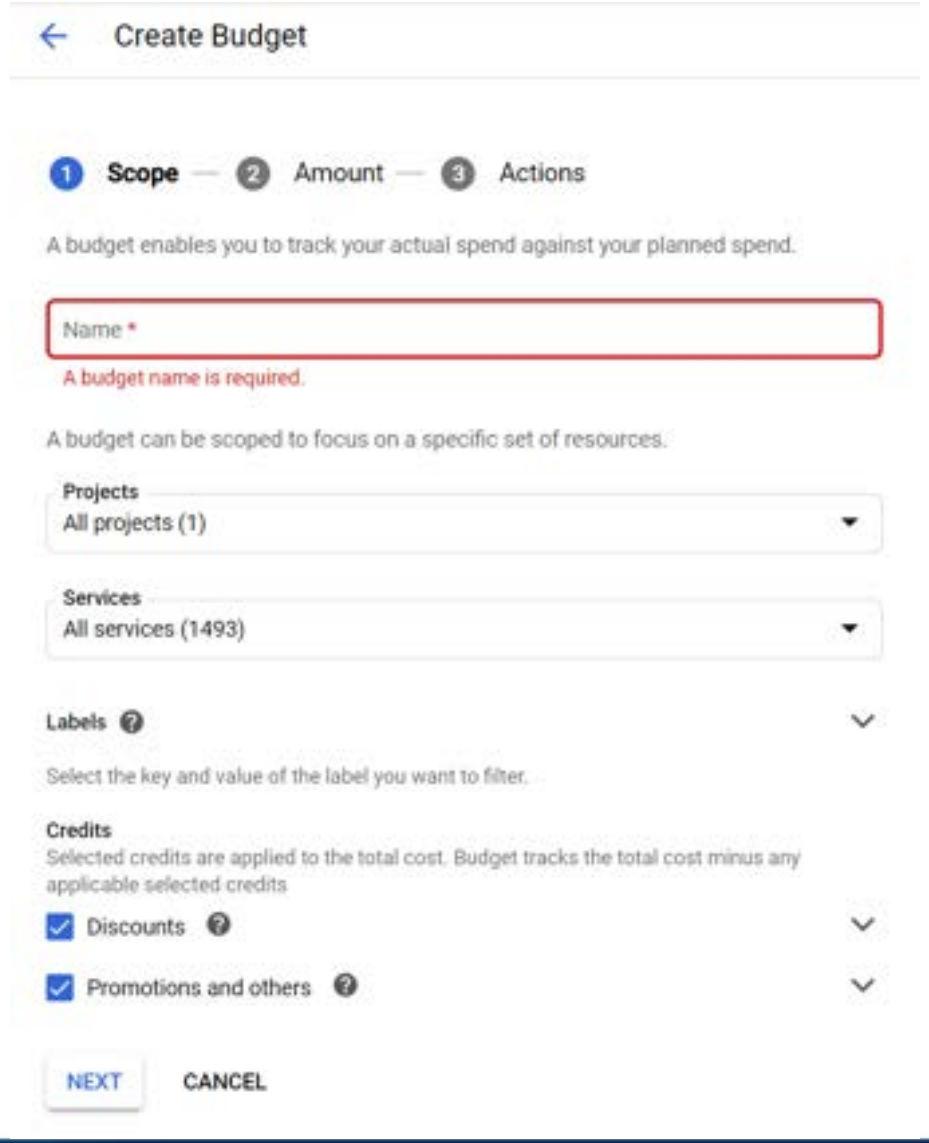
- From this page we can create a budget and use it to send us alerts at certain milestones or if we are spending at a faster monthly rate than previously.

Create a Budget

We can create a budget that tracks one or more of our Projects as well as one or more of the GCP services.

This can provide very granular tracking to optimize spending.

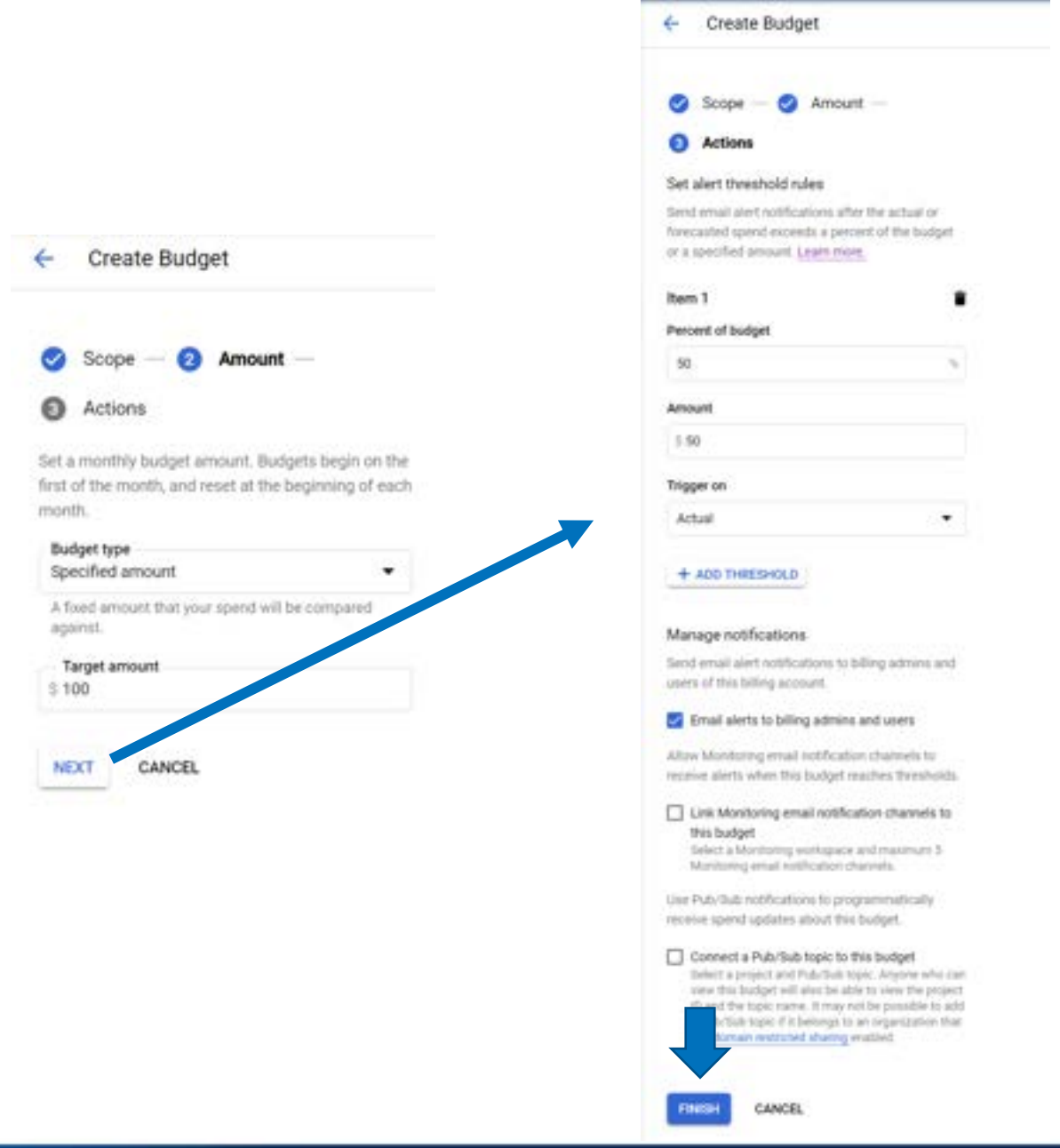
Or it can simply ensure you don't spend too much money without being alerted.



Simple Budget Alert

A simple budget sets a target spending limit.

Then we can trigger alerts to email when certain thresholds are hit, for example 50% of the budget.



API Credentials

- Sometimes APIs in GCP will need you to provide billing credentials in order to charge usage to your project.
- This is going to be important later when installing the SRA Toolkit.
- GCP handles this with a JSON file containing billing project information.

Get JSON Credentials for API

A description of how to make credentials needed for Google or AWS can be found here.

https://github.com/ncbi/sra-tools/wiki/04.-Cloud-Credentials

Review

- Projects are a fundamental part of GCP.
- Budgets allow you to track spending and set alerts to prevent unexpected expenses.
- Applications may require credentials to bill your service account. For the SRA Toolkit this takes the form of a JSON file that can be downloaded from the GCP Console.

Additional Resources

- GCP Intro Video https://www.youtube.com/watch?v=4D3X6XI5c_Y
- GCP Documentation Page https://cloud.google.com/docs
- GCP Billing Documentation https://cloud.google.com/billing/docs/how-to
- GCP Service Accounts https://cloud.google.com/iam/docs/creating-managing-service-accounts

Searching with BigQuery

Adam Stine

U.S. National Library of Medicine
National Center for Biotechnology Information

Overview

- What is BigQuery?
- Pinning Data in the Console
- Search SRA Metadata
- Taxonomy Searches of SRA

What is BigQuery?

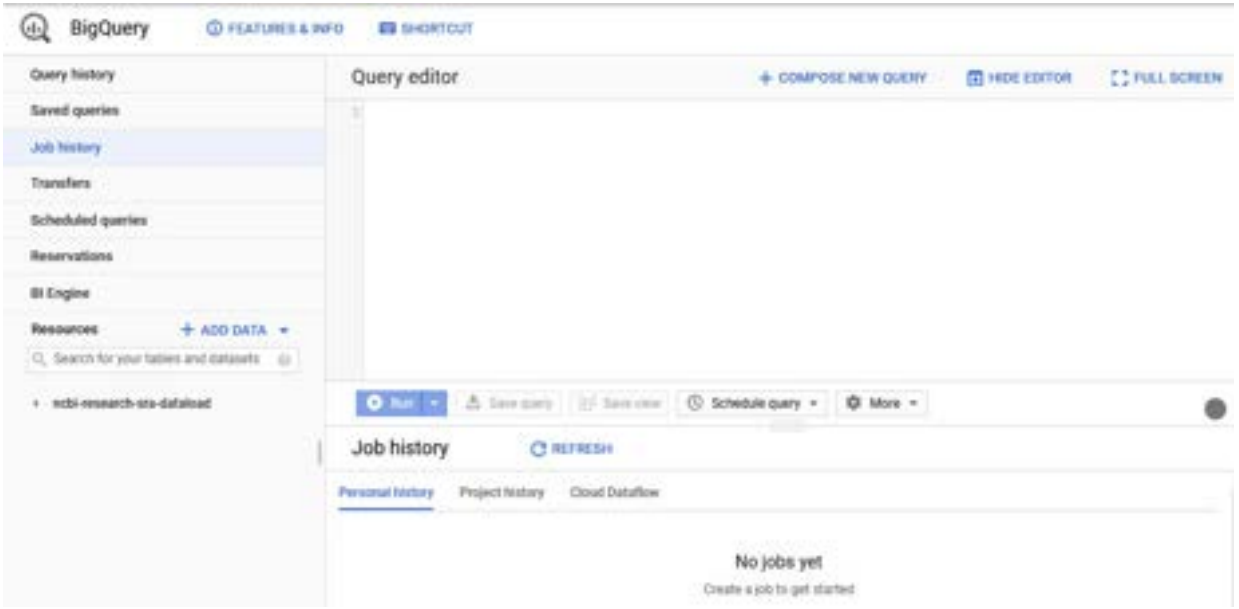
BigQuery is the Google Cloud Data Warehouse product.

It supports features like queries, access control, and data management.

The Sequence Read Archive (SRA) maintains a copy of the sequence metadata from submissions in BigQuery

This allows SQL queries to search the metadata in ways that are hard or impossible with the NCBI Entrez search engine.

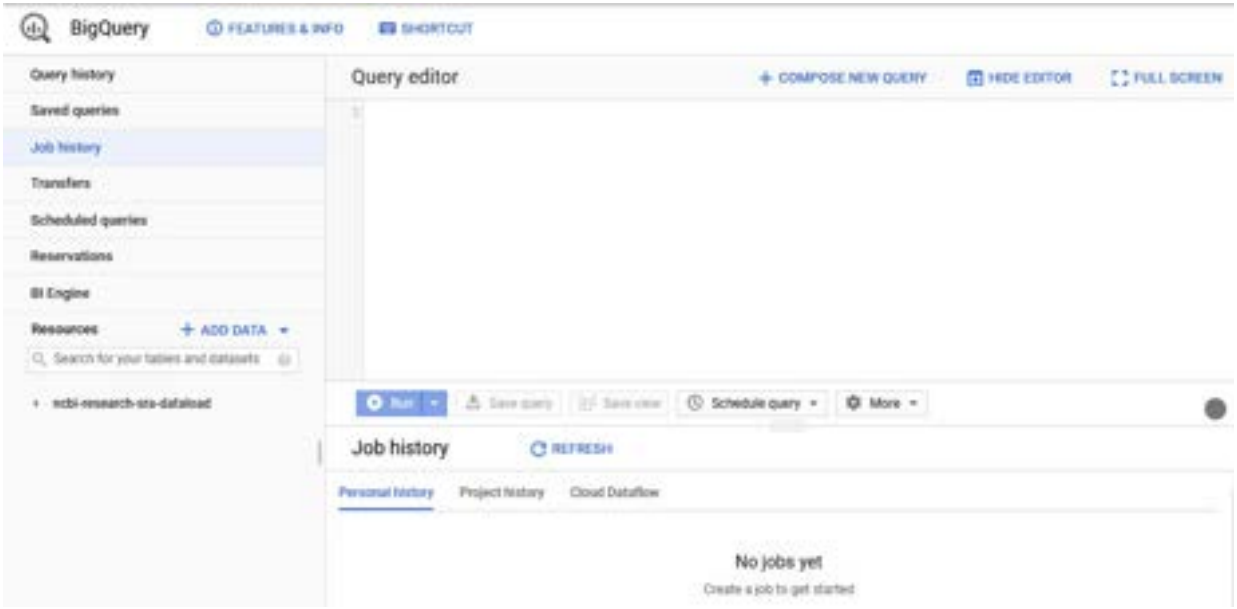
https://cloud.google.com/bigquery/docs/how-to



Intro to BigQuery

You can either use the ‘bq’ client from Google in a virtual machine or the GCP console in a web browser.

This guide will show you how you would setup a query in the GCP console.

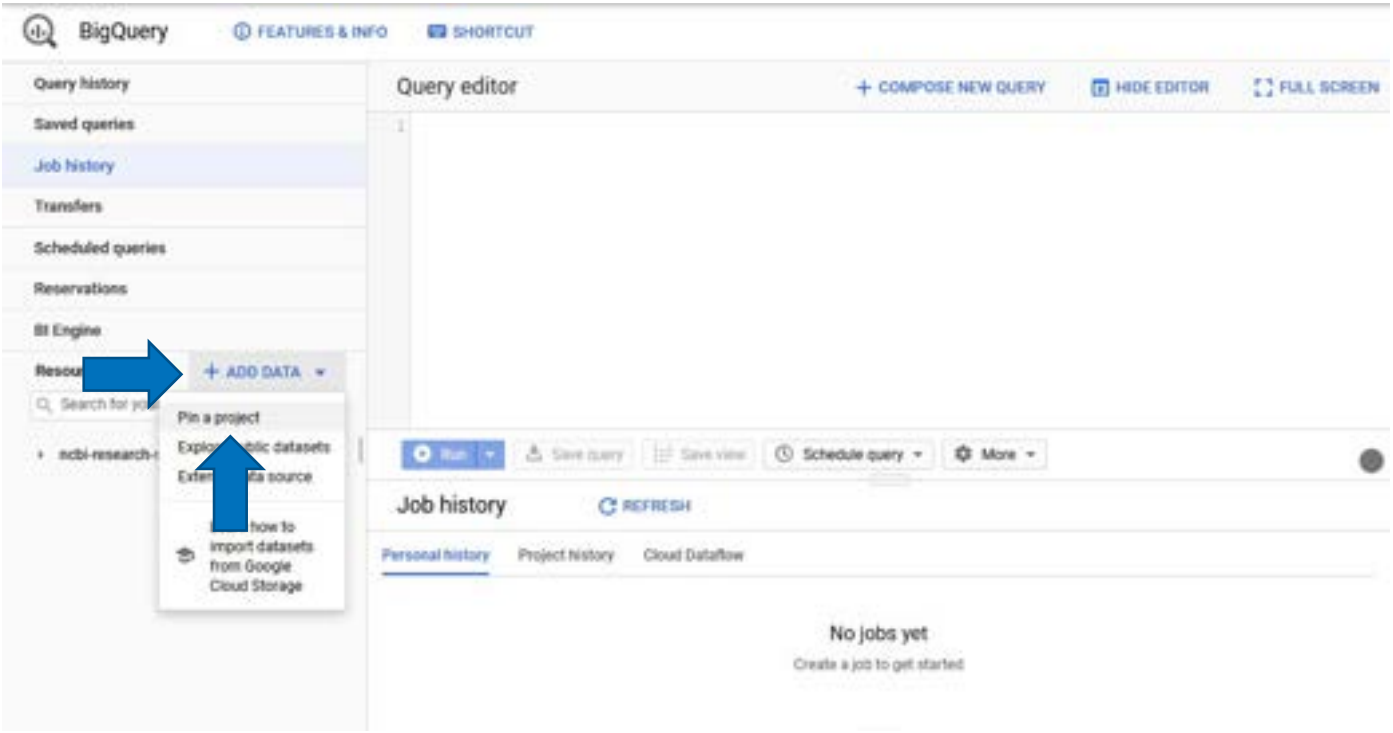


Adding Data

We can add the public SRA data set to our BigQuery project.

Click + ADD DATA on the left side of the console.

In the menu that appears click Pin a project.

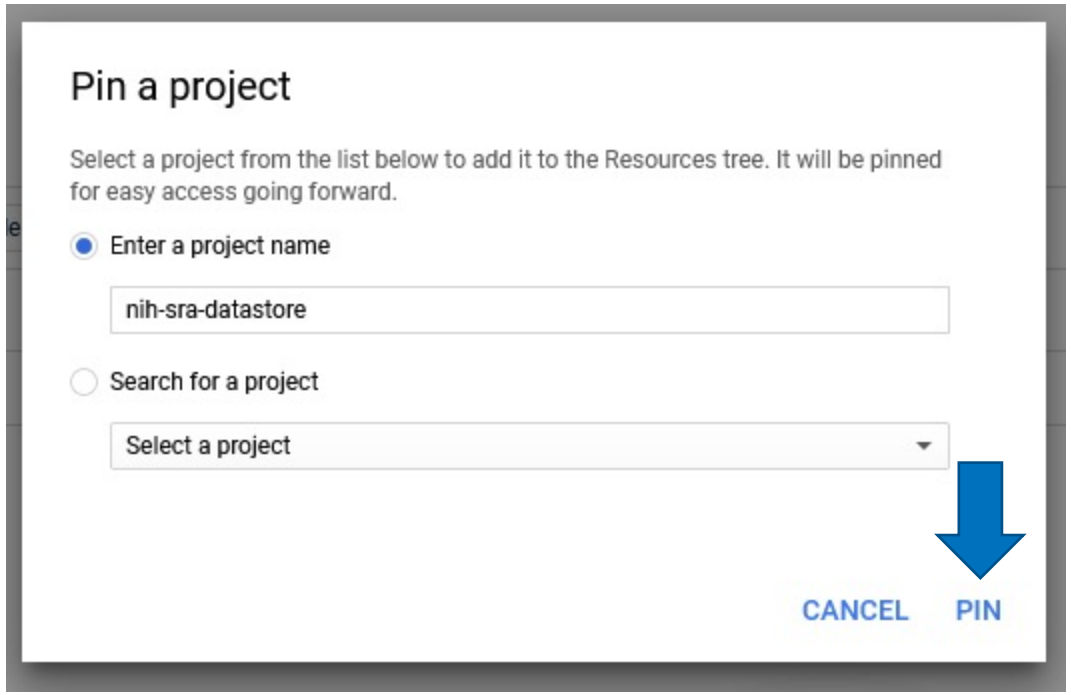


Adding Data

Enter nih-sra-datastore in the project name box.

Click PIN.

This will pin the public metadata datastore to your BigQuery console.



Simple SQL Queries

A very basic overview of SQL queries.
SELECT – command to extract data from a table.
FROM – specify which tables to extract data from
WHERE – filters the data
ORDER BY – order the results by specified column(s)

https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax

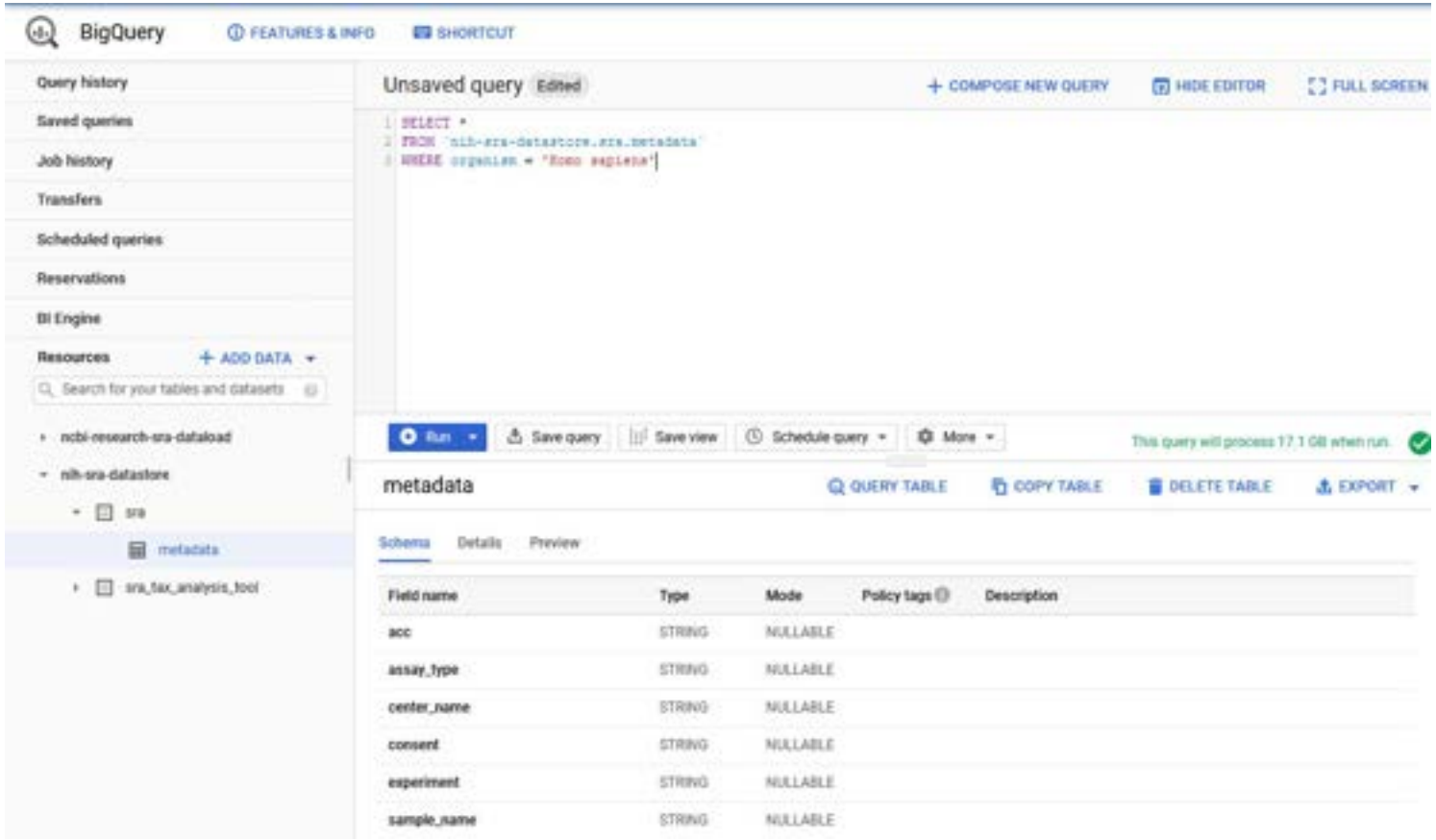
Searching in BigQuery

Here is an example search for the SRA Data.

**SELECT *
FROM `nih-sra-datastore.sra.metadata`
WHERE organism = 'Homo sapiens'**

This query will search in the metadata table that is part of the SRA dataset contained in the nih-sra-datastore project.

The query will look for all (Select *) records with 'Homo sapiens' in the organism column.

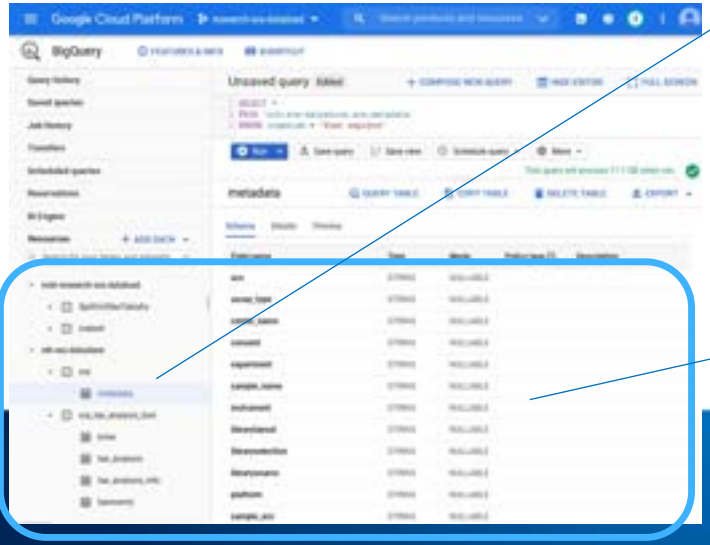


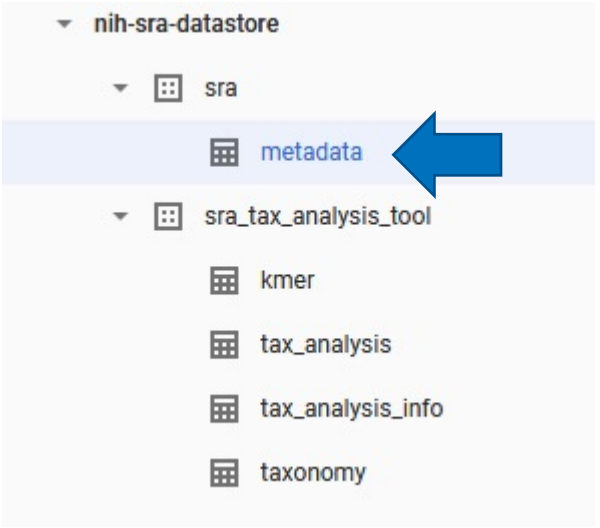
Console Previews and Info

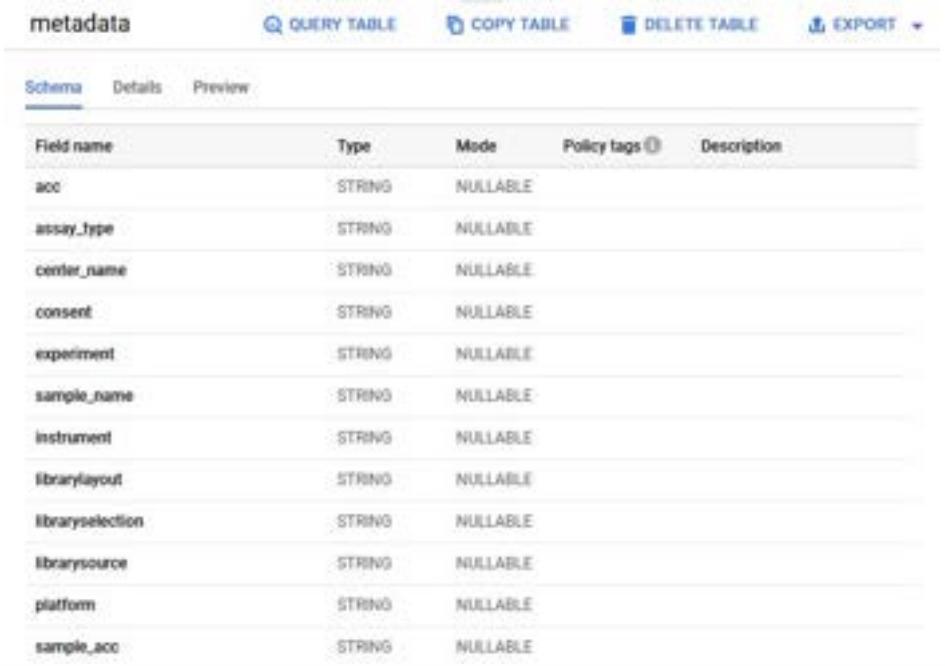
The console has some very useful features we can explore.

We can look through the tables available to query in the pinned dataset.

By clicking on a single table we can view the columns and data type of those columns.







NCBI

Documentation of SRA Cloud Metadata

The SRA documentation page includes:

- Available Tables List
- Text descriptions of the contents of the columns
- Additional example queries

Column name	Type	Description
acc	STRING	SRA Run accession in the form of SRR######## (ERR or DRR for INSDC partners)
assay_type	STRING	Type of library (i.e. AMPLICON, RNA-Seq, WGS, etc)
center_name	STRING	Name of the sequencing center
consent	STRING	Type of consent need to access the data (i.e. public is available to all, others are for dbGaP)
experiment	STRING	The accession in the form of SRX######## (ERX or DRX for INSDC partners)
sample_name	STRING	Name of the sample
instrument	STRING	Name of the sequencing instrument model
librarylayout	STRING	Whether the data is SINGLE or PAIRED
libraryselection	STRING	Library selection methodology (i.e. PCR, RANDOM, etc)
librarysource	STRING	Source of the biological data (i.e. GENOMIC, METAGENOMIC, etc)
platform	STRING	Name of the sequencing platform (i.e. ILLUMINA)
sample_acc	STRING	SRA Sample accession in the form of SRS######## (ERS or DRS for INSDC partners)
biosample	STRING	BioSample accession in the form of SAM#######? (SAMEA###### or SAMD###### for INSDC partners)
organism	STRING	Scientific name of the organism that was sequenced (as found in the NCBI Taxonomy Browser)
sra_study	STRING	SRA Study accession in the form of SRP######## (ERP or DRP for INSDC partners)
releasedate	TIMESTAMP	The date on which the data was released
bioproject	STRING	BioProject accession in the form of PRJNA######## (PRJEB###### or PRJDB###### for INSDC partners)
mbytes	INTEGER	Number of mega bytes of data in the SRA Run
loaddate	TIMESTAMP	The date when the data was loaded into SRA
avgspotlen	INTEGER	Calculated average read length
mbases	INTEGER	Number of mega bases in the SRA Runs
insertsize	INTEGER	Submitter provided insert size
library_name	STRING	The name of the library
biosamplemodel_sam	STRING	The BioSample package/model that was picked
collection_date_sam	STRING	The collection date of the sample
geo_loc_name_country_calc	STRING	Name of the country where the sample was collected
geo_loc_name_country_continent_calc	STRING	Name of the continent where the sample was collected
geo_loc_name_sam	STRING	Full location of collection

https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud-based-examples/

BigQuery Charges

We also get a preview of how much data will be searched.

Our example query will look through 17.1 GB of data to get a result.

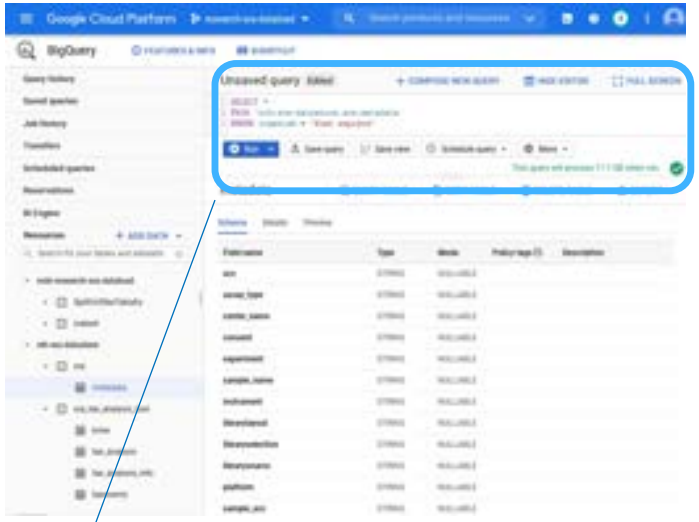
This is one of the ways you generate cost in BigQuery

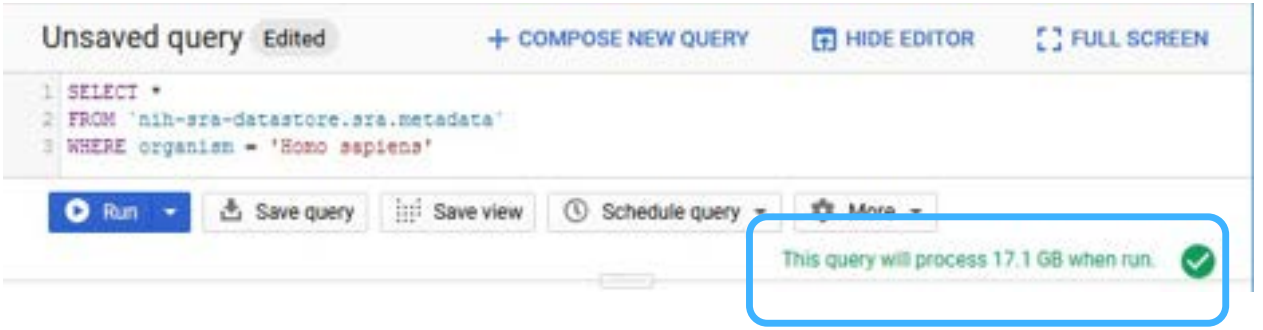
https://cloud.google.com/bigquery/pricing

Our example is an on-demand query which currently costs $5.00 per TB searched.

This query would be ~$0.10 to run

The first 1 TB each month is free.



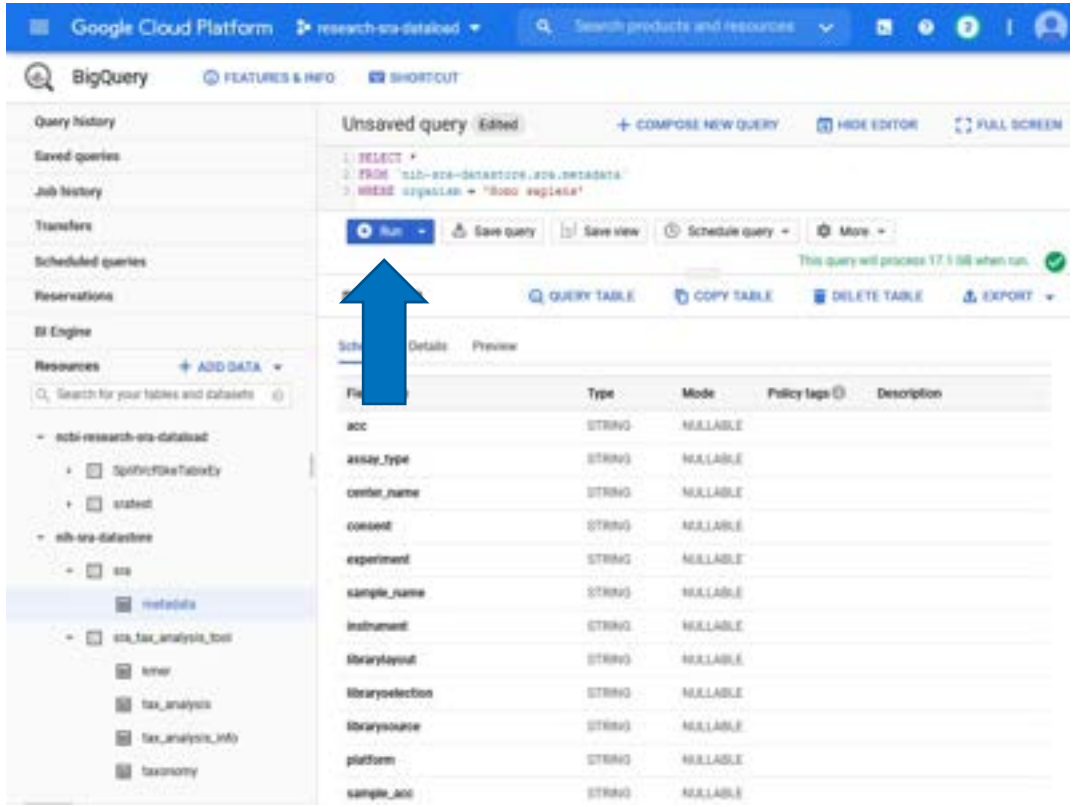


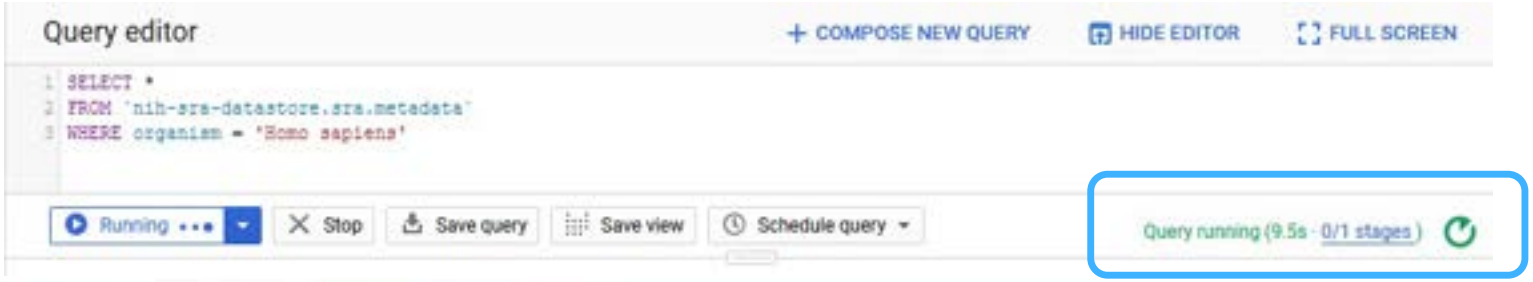
Running a Query

Click RUN to run the query.

While the query is running the console will show live updates on the status of the query.

Once the query finishes the results will display in the lower right portion of the console.



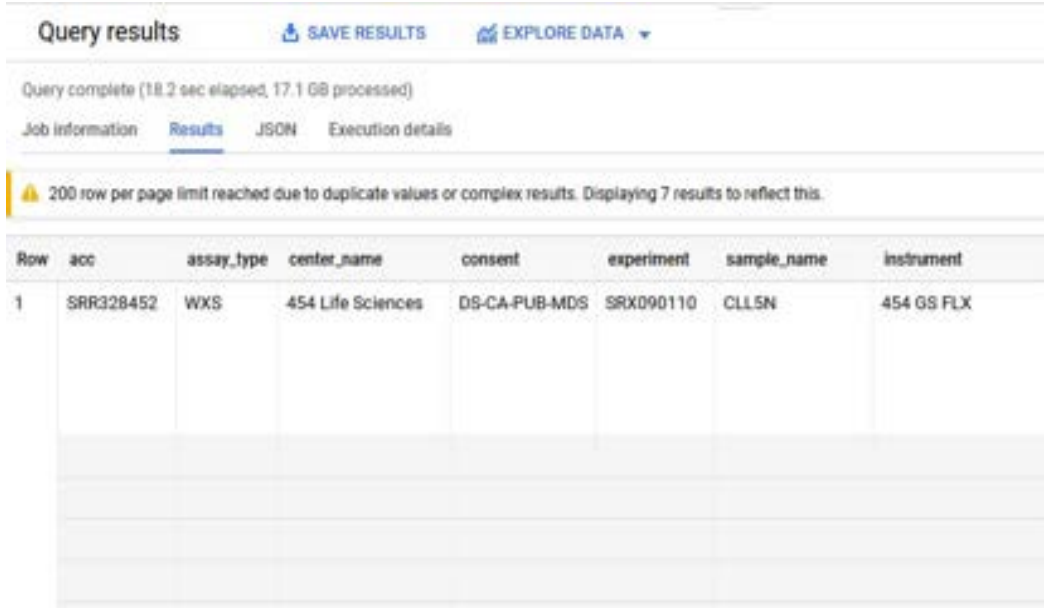


BigQuery Arrays

You might expect to see one record per row on the results but BigQuery supports an array feature for multi-value fields.

Scrolling to the right on this record we can find several columns that have multiple lines for this one record.

We won't go into depth on the data structure in these columns but it is an important feature of BigQuery



Unnesting Arrays

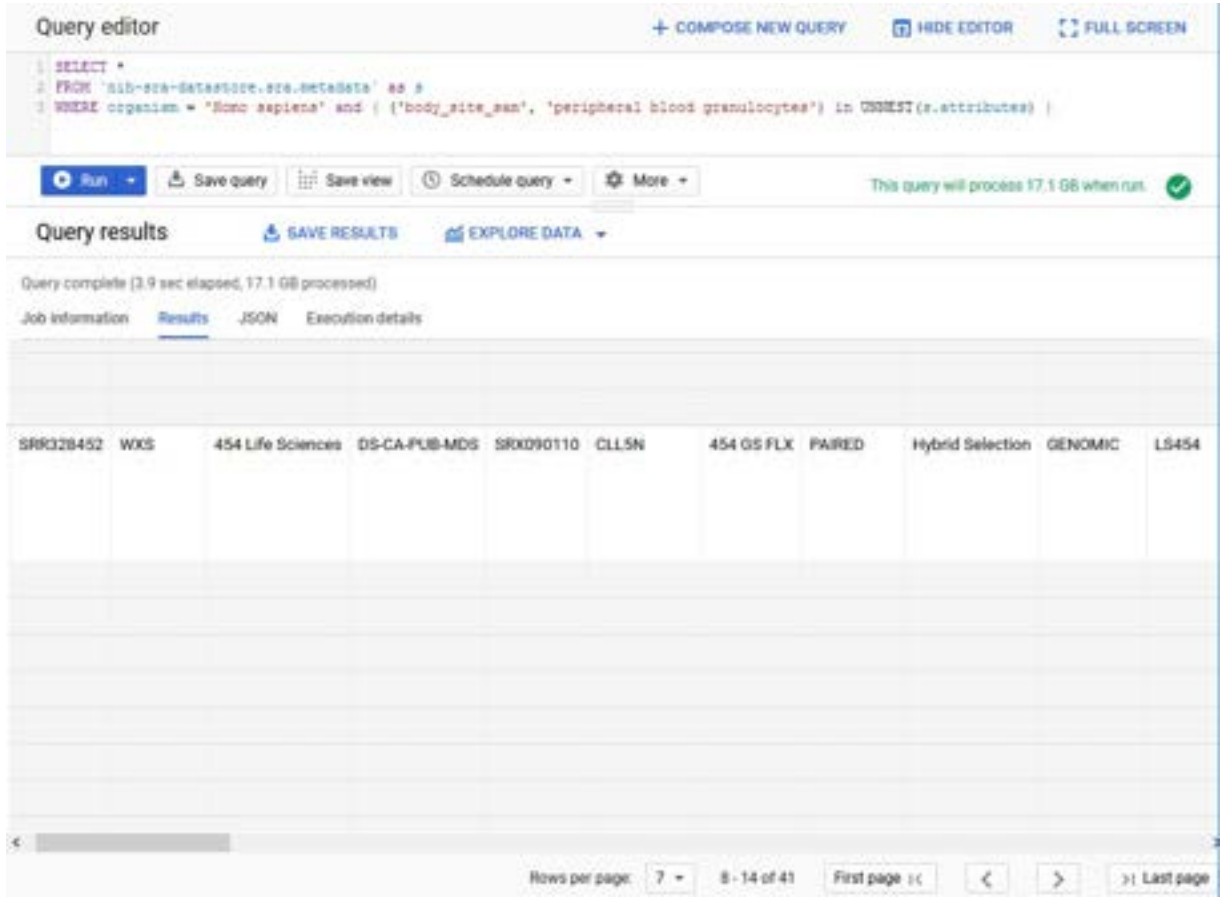
Here is an example search that will filter for a value in one of those columns.

SELECT *

FROM `nih-sra-datastore.sra.metadata` as s

WHERE organism = 'Homo sapiens' and (('body_site_sam', 'peripheral blood granulocytes') in UNNEST(s.attributes))

The unnest function is allowing us to search the key ‘body_site_sam’ and find only records with the value ‘peripheral blood granulocytes’



Taxonomy Data

In addition to metadata from the SRA database, there is also the sra_tax_analysis_tool dataset.

There are four tables in this dataset:

- tax_analysis_info: a summary table for the results of the STAT tool
- tax_analysis: use the taxonomy analysis table to locate any number of runs based on kmer hits to a particular organism or branch in a taxonomic tree.
- taxonomy: NCBI Taxonomy database where you can locate the taxid based on organism names.
- kmer: contains kmers mapped to a particular organism and allows you to continue exploring organismal content further. You can use kmer tables in your downstream analysis by building custom kmer libraries.

Searching in BigQuery

We can join the metadata and the taxonomy data and to search for sequence data in SRA that is from the coronaviridae family.

```
SELECT m.bioproject, m.biosample, m.acc, m.collection_date_sam,
m.geo_loc_name_sam,

(select v from unnest(m.attributes) where k = 'collected_by_sam') as collected_by,

(select v from unnest(m.attributes) where k = 'host_sam') as host

FROM `nih-sra-datastore.sra.metadata` m , `nih-sra-
datastore.sra_tax_analysis_tool.tax_analysis` tax

WHERE m.acc = tax.acc

and tax.name = 'Coronaviridae'
```

Searching in BigQuery

This query includes the project, sample, and run accessions for all the data we found. It also shows submitter supplied metadata for date, location, host, and collected by fields.

Note that some of this metadata might not have been provided by the submitter so the presence of certain metadata on one record does not imply it will be present on all records.

collection_date_sam	geo_loc_name_sam	collected_by
	China: Qingdao	null
	China: Qingdao	null
	China: Qingdao	null
	China: Qingdao	null
	India	null
	Canada: Saskatoon, SK	null
2019-08	China: shanghai	null
	Australia	null
2017-08-02	Hong Kong	null

Searching in BigQuery

The last query was part of this NCBI Minute webinar on using BigQuery

https://www.youtube.com/watch?v=DkNz-RCCm-M

The NCBI YouTube channel has additional videos on many topics that range in length from a few minutes to over an hour long.

https://www.youtube.com/channel/UCvJHVo5xGSKejBbBj0A5AyQ

Review

- BigQuery allows for SQL searches of large data sets quickly.
- BigQuery can be used to search the publicly accessible SRA metadata using SQL searches.
- SRA has generated taxonomy data using the SRA Taxonomy Analysis Tool (STAT) that can be used to find the organism content of runs in SRA.

Additional Resources

- STAT Description https://www.ncbi.nlm.nih.gov/sra/docs/sra-taxonomy-analysis-tool/
- Additional Examples for SRA Searches in BigQuery https://www.ncbi.nlm.nih.gov/sra/docs/sra-bigquery-examples/
- Google Quickstart for BigQuery https://cloud.google.com/bigquery/docs/quickstarts/quickstart-web-ui
- A video from Google that looks at nested data in BigQuery https://www.youtube.com/watch?v=STo98QUKDS8