CMPE-138/180B: Database Systems

Getting started with BigQuery

Overview

CMPE-138/180B uses BigQuery for its class projects; we also recommend playing around with BigQuery to enforce your understanding of class material. This is a guide about getting started with BigQuery, from getting the credits that we will provide to querying BigQuery public datasets. You are responsible for the information in this document, especially the portions about how to prevent yourself from burning all your credits.

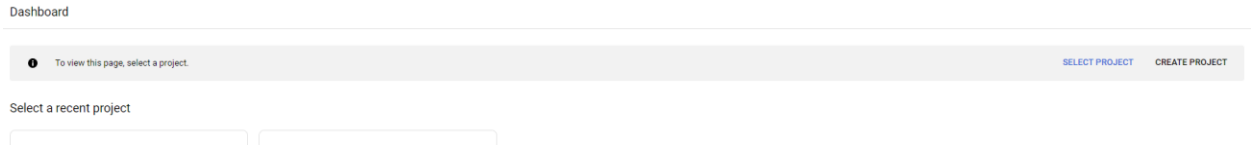
Getting Credits

Please follow the instructions in the Getting_Google_Cloud_Credits-CMPE-138.pdf document.

Initial Setup

This section will guide you through creating a BigQuery project and setting up your account so that you can easily query datasets. **Remember that all of this should be done on your personal Google account .**

1. Click this link . On the top right corner; click “Create Project” to make a GCP (Google Cloud Platform) project.



2. Fill in the information to make a new project. You can name your project anything, but we recommend something with a short project ID you can easily remember and type. Make sure to select the new billing account you should have after getting the class credits. (After you create the project, you can double check the linked billing account of your project through instructions here).

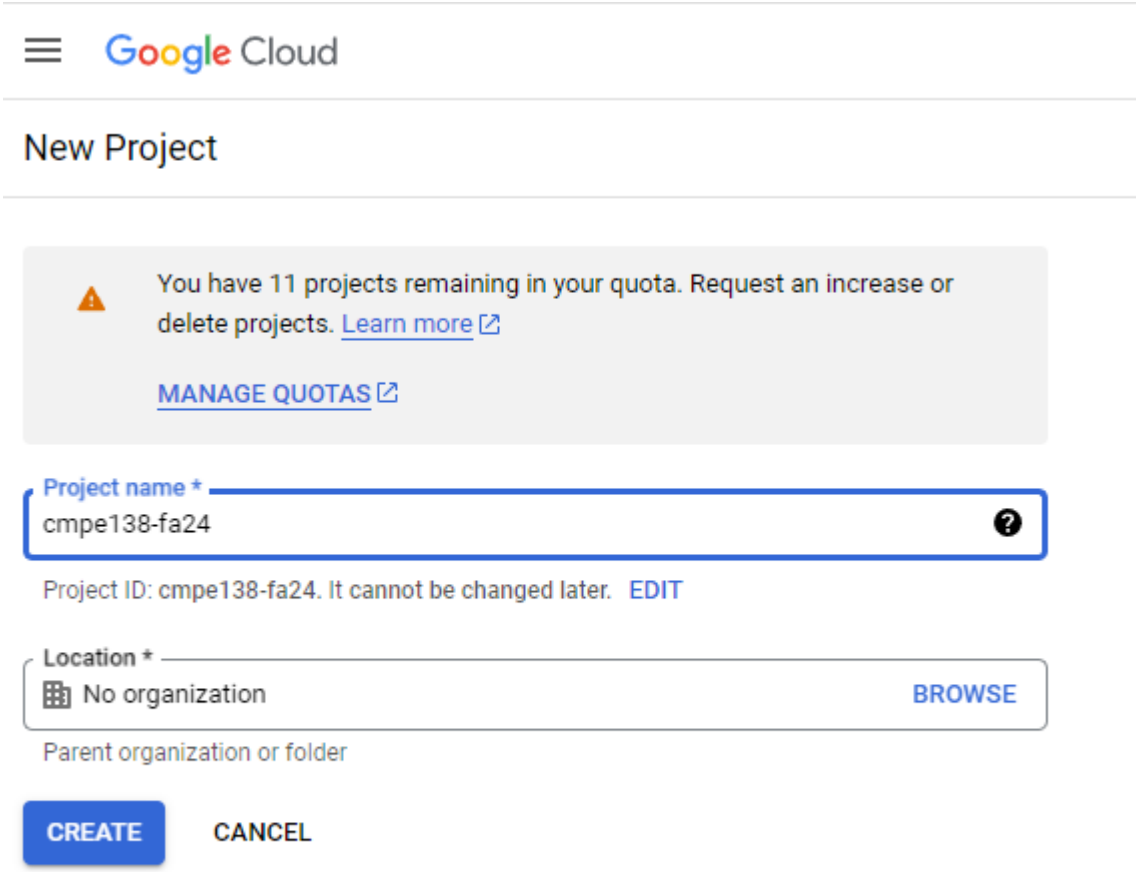
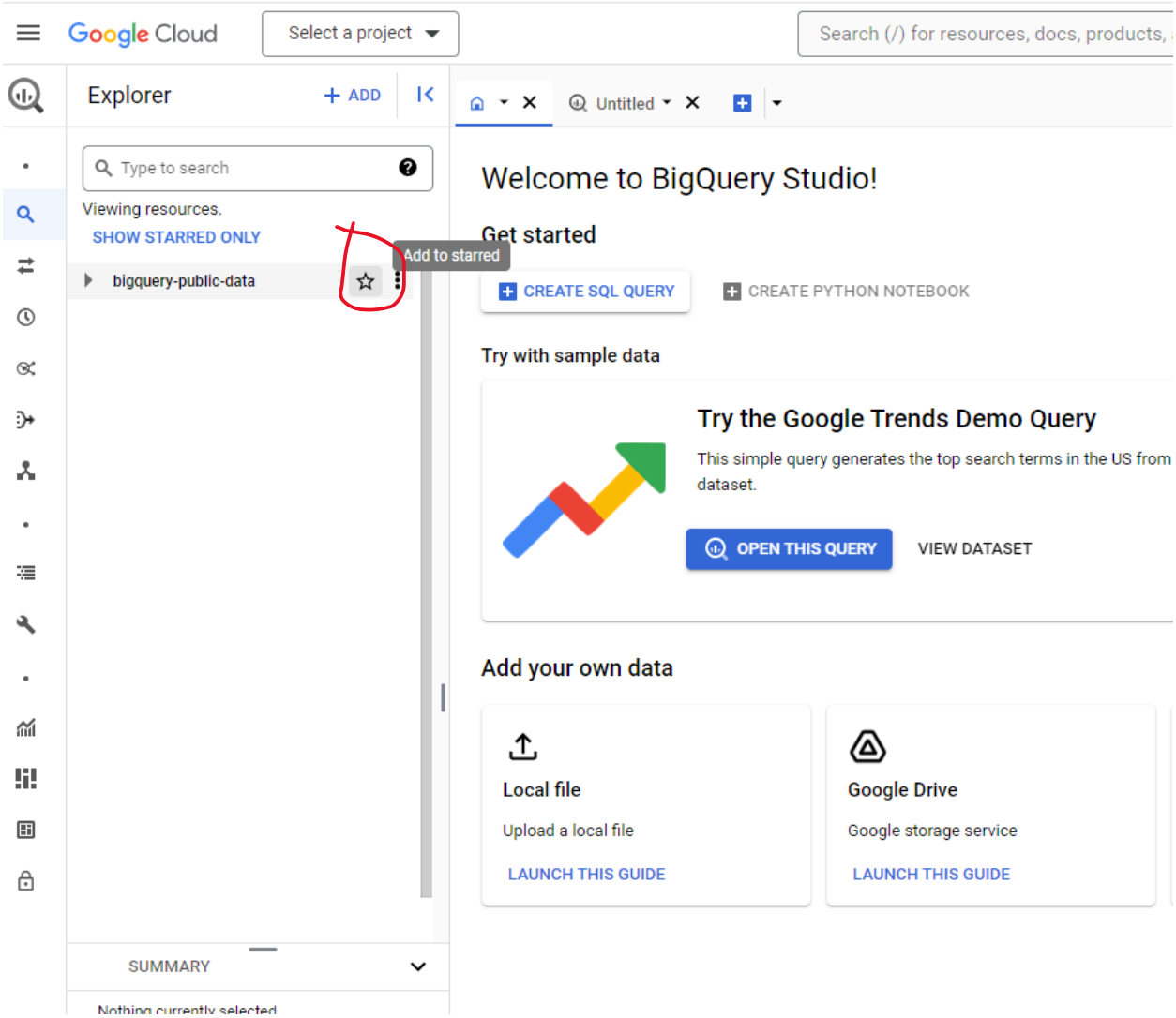


Fig. 2 Create new project

3. After creating your project, you'll be brought to an overview page for your project. Go to this link , which is the page for BigQuery's public datasets. Click on the star to the right of the "bigquery-public-data" resource. This will add the project to the list of your starred projects.

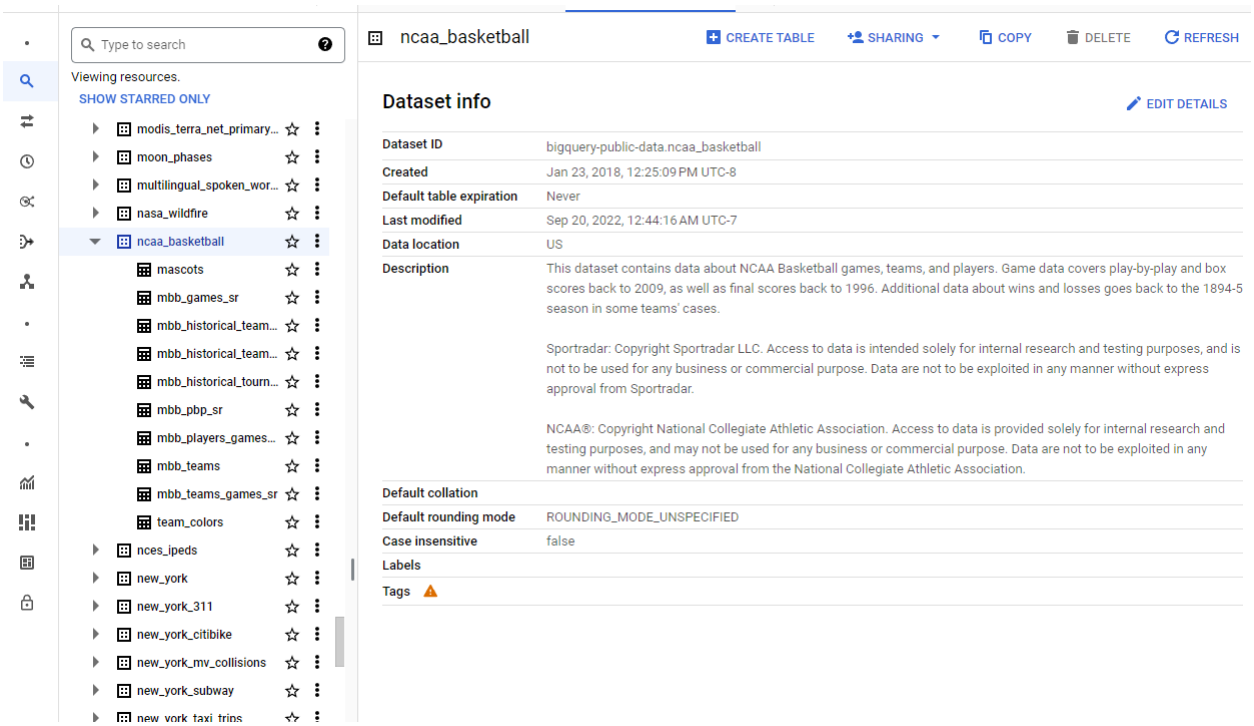


Now when you visit the console, you'll be able to easily find the datasets in your sidebar.

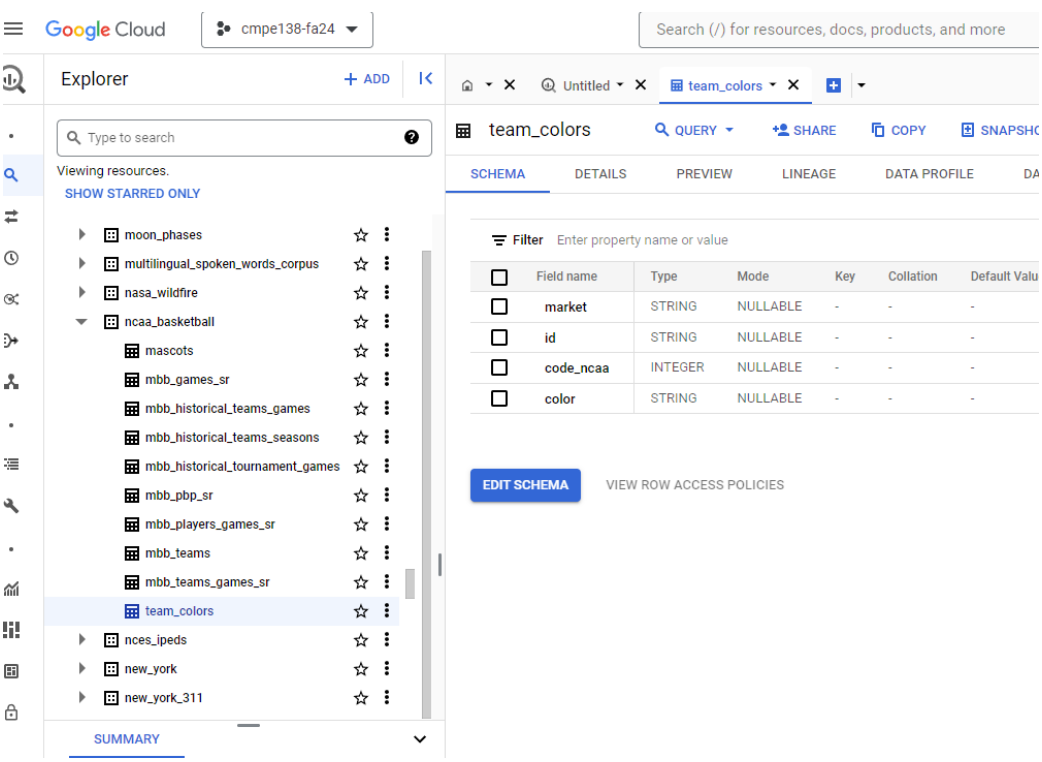
Querying Public Datasets

Here are some step-by-step instructions on how to get started with making queries on BigQuery's public datasets. We will use these for our class project.

1. Go to https://console.cloud.google.com/bigquery. You should see "bigquery-public-data" pinned on the left menu.
2. Click on "bigquery-public-data" and scroll until you find the dataset "ncaa_basketball" .
3. Click on the dataset. You'll see a brief description of what information the dataset contains, as well as a brief overview of information such as the dataset size.



4. In the sidebar, click on one of the tables in the dataset (for example, 'team_colors'). Here, you can find the table schema with a description of what each column represents. You can also find table details (such as the size of the table, which will give you a sense of how safe it is to query repeatedly given your data limits) and a preview of the table.



Google Cloud | cmpe138-fa24 | Search (/) for resources, docs, products, and more

Explorer + ADD

Type to search

Viewing resources.

SHOW STARRED ONLY

- moon_phases
- multilingual_spoken_words_corpus
- nasa_wildfire
- ncaa_basketball
  - mascots
  - mbb_games_sr
  - mbb_historical_teams_games
  - mbb_historical_teams_seasons
  - mbb_historical_tournament_games
  - mbb_pbp_sr
  - mbb_players_games_sr
  - mbb_teams
  - mbb_teams_games_sr
  - team_colors
- nces_ipeds
- new_york
- new_york_311

SUMMARY

team_colors

bigquery-public-data.ncaa_basketball

Last modified	May 31, 2018, 7:49:40 AM UTC-7

Untitled | team_colors

team_colors QUERY SHARE COPY SNAPSHOT

SCHEMA DETAILS PREVIEW LINEAGE DATA PROFILE DATA QUALIT

Table info

Table ID	bigquery-public-data.ncaa_basketball.team_colors
Created	Mar 5, 2018, 5:56:23 PM UTC-8
Last modified	May 31, 2018, 7:49:40 AM UTC-7
Table expiration	NEVER
Data location	US
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Description	Hex color codes for the 351 current men's D1 basketball teams.
Labels	
Primary key(s)	
Tags	

Storage info

Number of rows	351
Total logical bytes	23.35 KB
Active logical bytes	0 B
Long term logical bytes	23.35 KB
Total physical bytes	13.49 KB
Active physical bytes	0 B
Long term physical bytes	13.49 KB
Time travel physical bytes	0 B

Google Cloud | cmpe138-fa24 | Search (/) for resources, docs, products, and more

Explorer + ADD

Type to search

Viewing resources.

SHOW STARRED ONLY

- moon_phases
- multilingual_spoken_words_corpus
- nasa_wildfire
- ncaa_basketball
  - mascots
  - mbb_games_sr
  - mbb_historical_teams_games
  - mbb_historical_teams_seasons
  - mbb_historical_tournament_games
  - mbb_pbp_sr
  - mbb_players_games_sr
  - mbb_teams
  - mbb_teams_games_sr
  - team_colors
- nces_ipeds
- new_york
- new_york_311

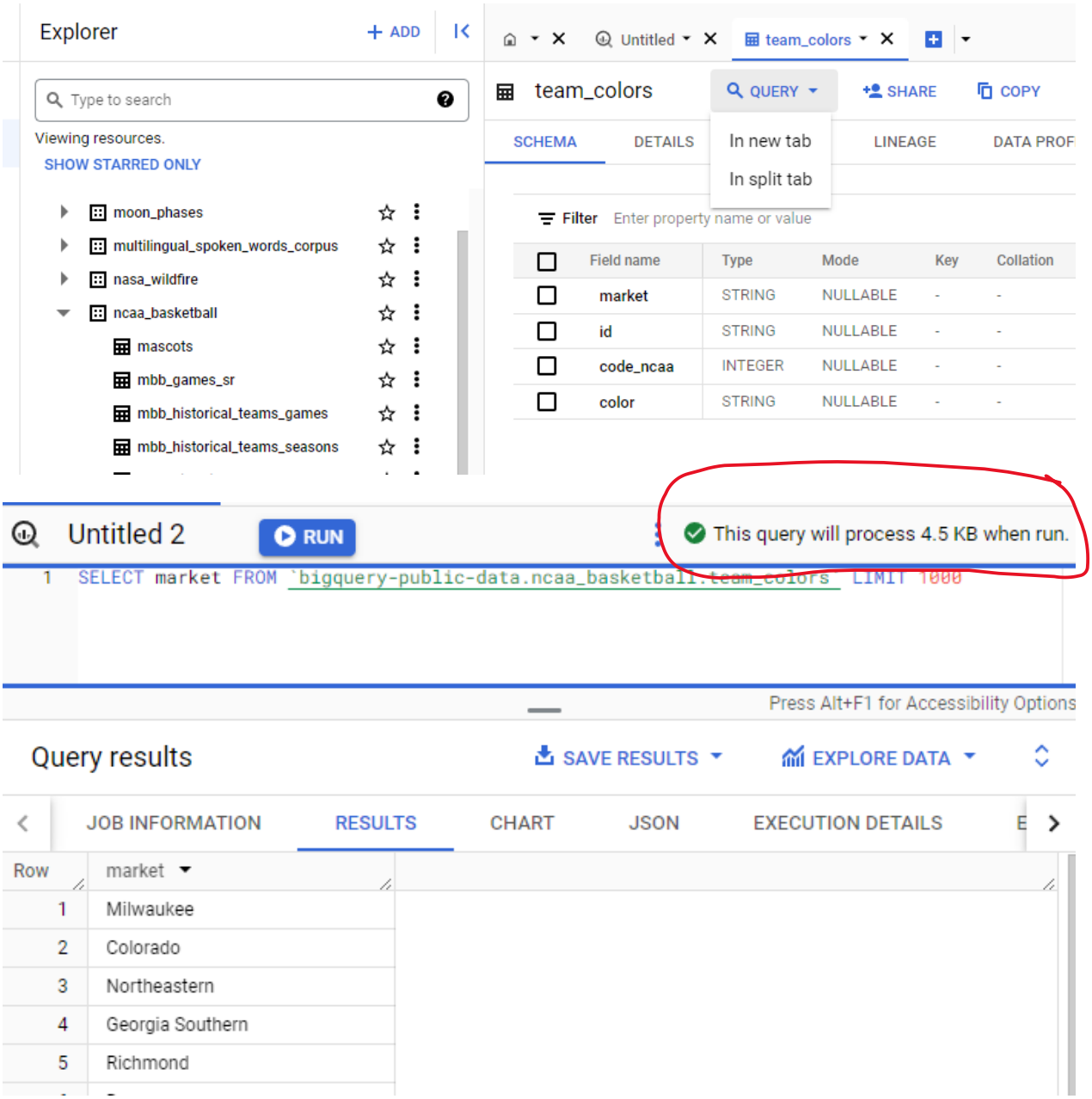
Untitled | team_colors

team_colors QUERY SHARE COPY SNAPSHOT DELETE

SCHEMA DETAILS PREVIEW LINEAGE DATA PROFILE DATA QUALITY

Row	market	id	code_ncaa	color
1	Milwaukee	5d77800f-1ae6-4b66-8e97-b0d...	797	#000000
2	Colorado	9fccbf28-2858-4263-821c-fdef...	157	#000000
3	Northeastern	93df9b18-e9fc-42a7-bb45-a73...	500	#000000
4	Georgia Southern	6b955b96-b736-475e-bffd-e4a...	253	#000066
5	Richmond	9b66e1e0-aace-4671-9be2-54c...	575	#000066
6	Duquesne	fea46ac5-6dad-43cd-a770-755...	194	#000144
7	North Florida	09920a5f-1b25-466c-b5ae-616...	2711	#001a49
8	Belmont	a0a22502-0d84-440c-84af-1fb...	14927	#00205c
9	Murray State	77a69fb0-1355-4342-ac09-b4c...	454	#002144
10	Pittsburgh	24051034-96bb-4f78-a3a6-312...	545	#002144
11	Virginia	56913910-87f7-4ad7-ae3b-5cd...	746	#002147
12	Penn State	4aebd148-8119-4875-954c-66...	539	#002147
13	George Washington	d52c3640-069c-4554-982e-e65...	249	#002147
14	Monmouth	3db7336c-c18a-441b-912e-e2a...	439	#002147
15	New Hampshire	93cb009e-5bbf-4081-b2c9-440...	469	#002162
16	Drake	aed211c3-23a4-4188-ad70-22c...	189	#002181
17	BYU	c31455b2-8a45-4248-aa8f-ce7...	77	#002255
18	Dayton	632616c5-2dbb-4017-a449-c9...	175	#002453
19	Utah State	7672ff16-8436-47e6-8546-0fb5...	731	#00264a

5. Click “Query” then select “In split tab” and try to run a query



Best Practices

1. Pay attention to the estimated number of bytes read by the query (Circled in the image above). Once you compose your query, you should see the number on the right side of the bottom panel.
   a. You will be billed by the number of bytes read by the query.
   b. If the estimated number of bytes is greater than 1GB, try to put on some constraints on your query. For example, only select the columns that you need.

2. If you are just exploring/trying out queries, use **LIMIT** to query fewer data. Also, avoid using **SELECT *** . Google will charge the query as scanning the whole table.
3. It's always helpful to use the "Preview" pane on a BigQuery table to see the first few rows of the table to see what data you're dealing with when writing your query.
4. In declarative languages, it's easier to build up the query piece by piece. Start with a basic frame of what you're looking for (maybe write the conditions, or do a join). Then add complexity to your query one bit at a time. It's much easier to debug this way as well.
5. BigQuery can auto-format your SQL queries with CTRL-SHIFT-F on Windows or CMD-SHIFT-F on Mac. This might be nice to learn about conventional SQL style guidelines (and will also make your queries more readable, which we appreciate).