CS145: Data Management and Data Systems

Stanford University, Fall 2018

Getting Started with BigQuery

Overview

This is information from us to you about how to get started with BigQuery and the credits we provide this quarter. **You are responsible for the information in this document, especially the portions about how to prevent yourself from burning all your credits.**

Getting your credits

Google has provided all students in this class with $50 of credit to use for BigQuery. This should be *more than ample* to finish the course, possibly even with credit remaining.

Credit policies & information:

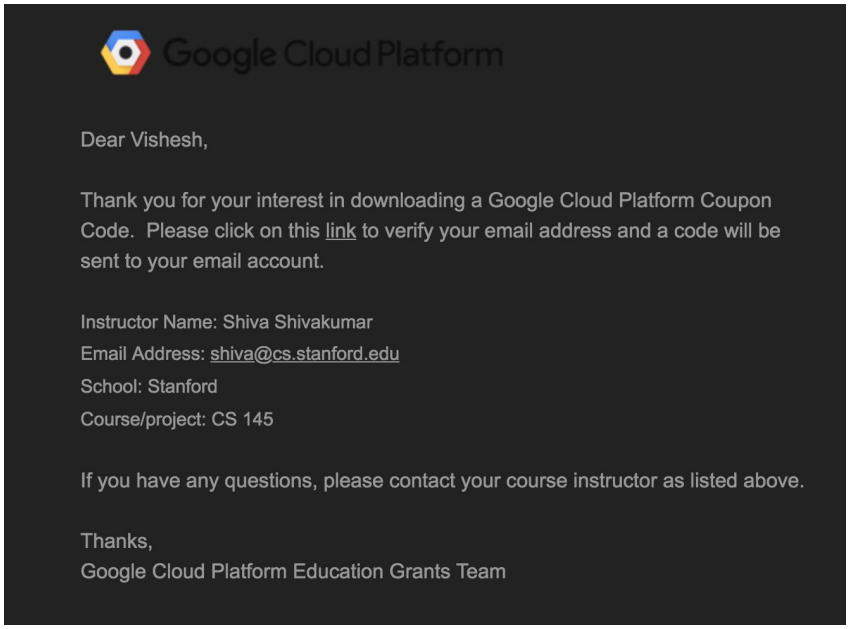
1. $50 of credit is enough to query over **10 Terabytes** of data ($5/TB). This is a very large amount of data for the purposes of this class. You would need to run 1000 queries on 10GB in order to exhaust this, for example.
2. Google provides all users of bigquery an additional **1TB free / month**. You may find that you don't even use 1TB over the course of the first two projects.
3. You are responsible for your credit. If you are in danger of running out (eg, you are running $2 queries) please contact the TAs. We are able to help students *before* they use up their credits, but there's not much we can do *after* you've used them up.
4. Google charges by **# of rows * # of columns * size of column** for each query. The easiest and best way to keep the amount of data you handle down is to use **only the columns you need for your query.** It can be a little verbose at times, but if you stick to the practice of writing **`SELECT column1, column2 ...`**, you will save lots of credits over the course of the quarter.

Note: AVOID USING **`SELECT *`**. Google will charge the query as scanning the whole table, even if it doesn't.

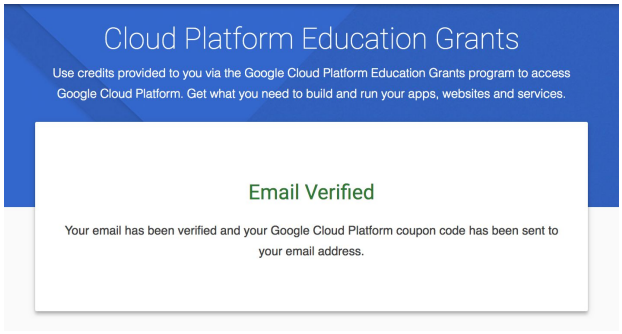
In order to get your bigquery credits, you will need to:

1. Go to this link.
2. Enter your name and Stanford email address

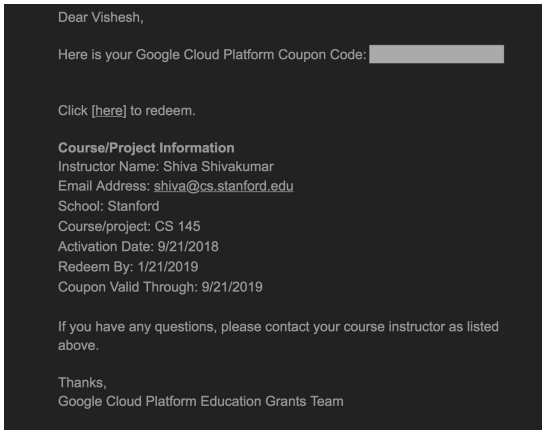
3. A verification link should be sent to your Stanford email address



4. After clicking the link, you should see a page which says “Email Verified”



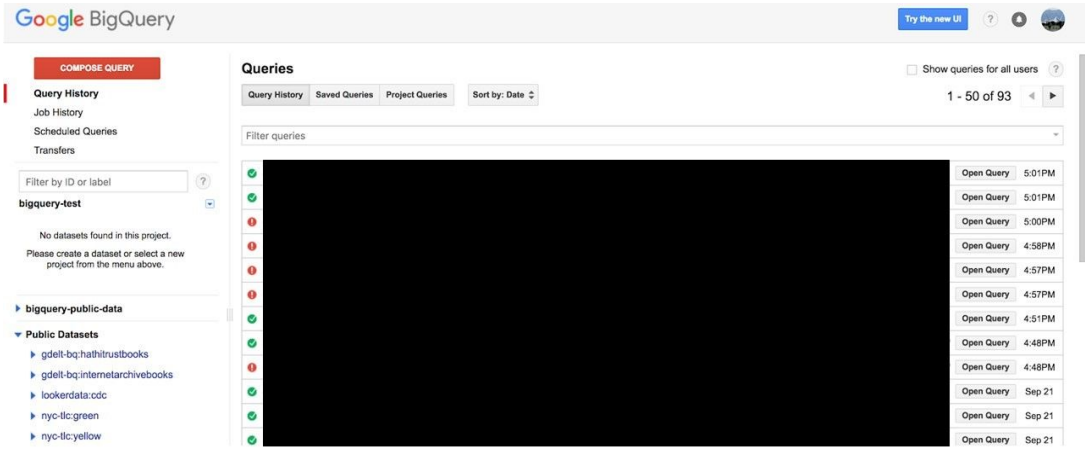
5. Receive the coupon in the email:

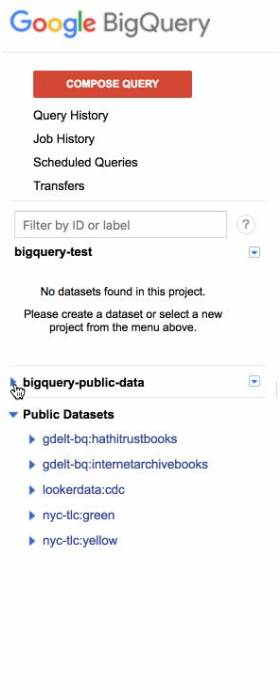


6. Go to **https://console.cloud.google.com/education**
7. Verify that your **personal google** account is being used. If you don't have one, you may need to set one up in order for billing to work correctly.
8. Enter your code in the box and click accept. This should add the credits to your billing info in https://console.cloud.google.com/billing.

Finding public datasets

1. Go to https://bigquery.cloud.google.com
2. You should see "bigquery-public-data" in the left menu:





3. Open the menu and find **`ncaa_basketball`** (or another dataset)
4. Click on any table to view its schema along with description of what the columns represent. Note that you can see the *size* of the table, which will give you a sense for how safe it is to query repeatedly given your data limits
5. Click "Query Table" in the main window for the console to pop up

Query setup and best practices

1. Make sure that you are using **Standard SQL**, not legacy SQL[1].
2. Set a query limit to 5GB, especially for the first two assignments.
3. If you are just exploring/trying out queries, use **`LIMIT`** to query less data.
4. 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.
5. 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.
6. Remember - in a declarative language, you say **what** to do, not how to do it. Think in terms of writing a function that takes a table as an input.

[1] You can select to use Standard SQL/Legacy SQL via a check box on the "Query settings" pane

References

1. https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators for a list of functions that BigQuery supports
2. https://cloud.google.com/bigquery/docs/best-practices-costs for more best practices to save cost