

# CS145: Data Management and Data Systems

Stanford University, Fall 2020

## Getting Started with BigQuery

### Table of Contents

- [Overview](#)
- [Getting Credits](#)
- [Initial Setup](#)
- [Querying Public Datasets](#)
- [Best Practices](#)
- [References](#)

### Overview

CS145 uses BigQuery for its three 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

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

Credit policies & information:

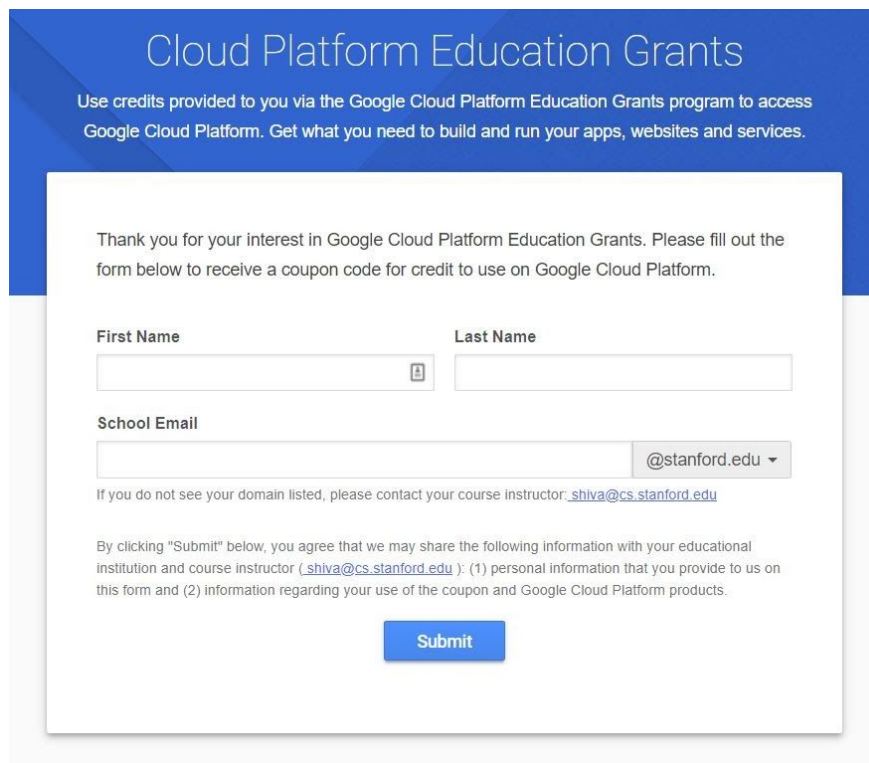
1. \$50 of credit is enough to query **10 Terabytes** of data (\$5/TB). This is a very large amount of data for the purposes of this class. You would need to run 1000 queries on 10GB in order to exhaust this, for example.
2. Google provides all users of BigQuery an additional **1TB free / month**. You 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.

- 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:

- Go to [this](#) link. You'll see the following page:



The screenshot shows a web form titled "Cloud Platform Education Grants" with a blue header. Below the header, it says "Use credits provided to you via the Google Cloud Platform Education Grants program to access Google Cloud Platform. Get what you need to build and run your apps, websites and services." The main content area is white and contains a thank-you message: "Thank you for your interest in Google Cloud Platform Education Grants. Please fill out the form below to receive a coupon code for credit to use on Google Cloud Platform." The form has three input fields: "First Name" and "Last Name" (each with a small icon to its right) and "School Email" (with a dropdown menu showing "@stanford.edu"). Below the email field, there is a note: "If you do not see your domain listed, please contact your course instructor: [shiva@cs.stanford.edu](mailto:shiva@cs.stanford.edu)". At the bottom of the form is a blue "Submit" button. A privacy notice at the bottom of the form states: "By clicking 'Submit' below, you agree that we may share the following information with your educational institution and course instructor ([shiva@cs.stanford.edu](mailto:shiva@cs.stanford.edu)): (1) personal information that you provide to us on this form and (2) information regarding your use of the coupon and Google Cloud Platform products."

- Enter your name and Stanford email address.
- Check your Stanford email. A verification link should have been sent to your Stanford email address (see below).



Dear Jennie,

Thank you for your interest in downloading a Google Cloud Platform Coupon Code. Please click on this [link](#) to verify your email address and a code will be sent to your email account.

Instructor Name: Shiva Shivakumar

Email Address: [shiva@cs.stanford.edu](mailto:shiva@cs.stanford.edu)

School: Stanford

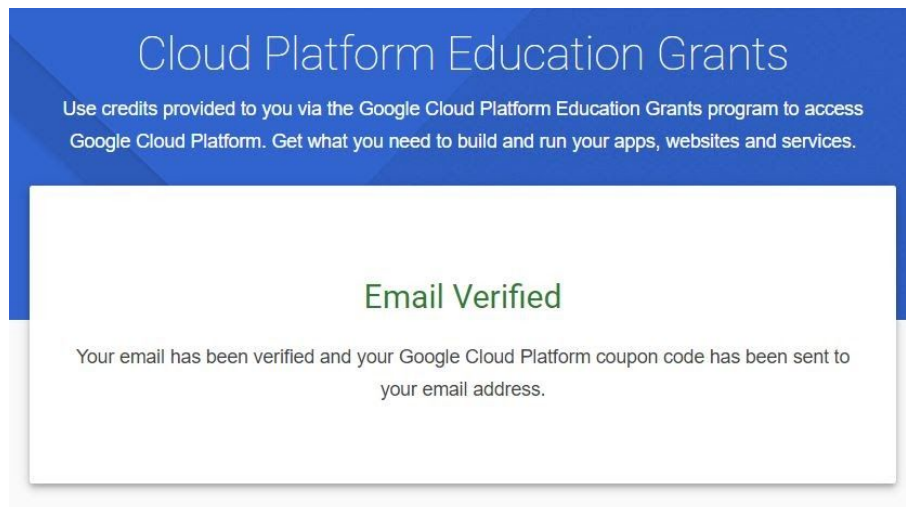
Course/project: CS 145 - Data Management and Data Systems

If you have any questions, please contact your course instructor as listed above.

Thanks,

Google Cloud Platform Education Grants Team

4. Click the link. You should see the following page:



5. Check your Stanford email again. You should see an email like the following:



Dear Jennie,

Here is your Google Cloud Platform Coupon Code: **0VU7-3WR5-ALUQ-MYFB**

Click [\[here\]](#) to redeem.

**Course/Project Information**

Instructor Name: Shiva Shivakumar

Email Address: [shiva@cs.stanford.edu](mailto:shiva@cs.stanford.edu)

School: Stanford

Course/project: CS 145 - Data Management and Data Systems

Activation Date: 9/23/2019

Redeem By: 1/23/2020

Coupon Valid Through: 9/23/2020

If you have any questions, please contact your course instructor as listed above.

Thanks,

Google Cloud Platform Education Grants Team

6. Click the link in the email. You will see a page to redeem the code (see below). **Make sure you are logged in to a personal Google account, NOT your Stanford account.**

## Education grants

Please enter the coupon code provided to you via the Google Cloud Platform Education Grants programme to receive credit for Google Cloud Platform. Get what you need to build and run your apps, websites and services.

### Coupon code

0VU7-3WR5-ALUQ-MYFB

Credit amount	Expiry date	Course
\$50.00	22 Sep 2020	CS 145 - Data Management and Data Systems

### Terms of Service

#### Country of residence

United States

#### Google Cloud Platform education grants credits terms and conditions

By clicking "Accept and continue" below, you, on behalf of yourself and the organization you represent ("You") agree to these terms and conditions:

The credit is valid for Google Cloud Platform products and is subject to Your acceptance of the applicable Google Cloud Platform License Agreement and any other applicable terms of service. The credit is non-transferable and may not be sold or bartered. Unused credit expires on the date indicated on the media conveying the promotion code. The credit may be issued in increments as You use the credit over the period of time during which the credit is valid. Offer void where prohibited by law.

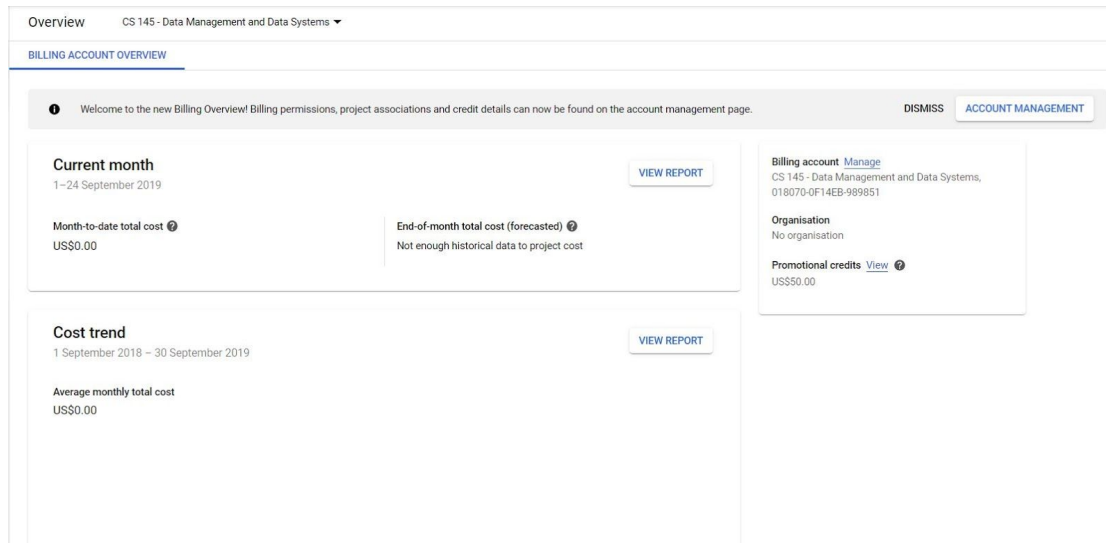
You represent that you are accepting the promotional credit on behalf of your educational institution and the credit can only be used on behalf of the educational entity and not for your personal use. You represent, on behalf of such educational entity, that (i) You are authorized to accept this credit; (ii) the credit is consistent with all applicable laws and regulations, including relevant ethics rules and laws; and (iii) the provision of credits will not negatively impact Google's current or future ability to do business with such educational entity.

You agree that we may share the following information with your educational institution and course instructor: (1) personal information that you provide to us during the coupon redemption process and (2) information regarding your use of the coupon and Google Cloud Platform products.

Accept and continue

Clear

7. Once you've verified that you are logged in to a personal Google account, click "Accept and continue". You'll be taken to a page with a billing overview (see below). This is where you can keep track of how much of your credit you have used and how much you have left.

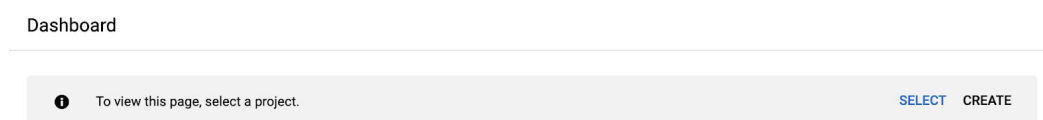


## Initial Setup

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.**


**Note:** This quarter, we will be focusing support on BigQuery's new Web UI. You are welcome to use the Classic Web UI if you prefer, but all instructions below will be for the new Web UI and most support will be for the new Web UI.

1. Click this [link](#). You'll see the page below; click "Create" 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](#)).


## New Project

 You have 23 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)


[MANAGE QUOTAS](#)

**Project name \***  
cs145-fa19 

Project ID: cs145-fa19. It cannot be changed later. [EDIT](#)

**Billing account \***  
CS 145 - Data Management and Data Systems 

Any charges for this project will be billed to the account that you select here.

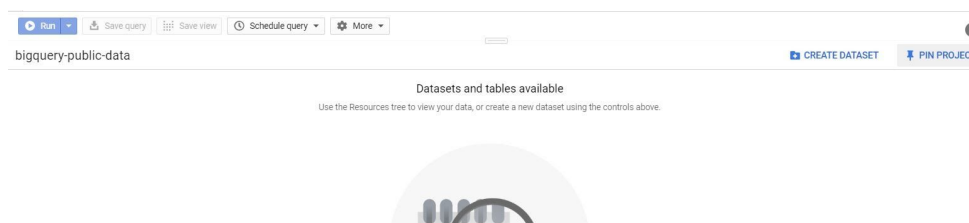
**Location \***  
 No organisation [BROWSE](#)

Parent organisation or folder

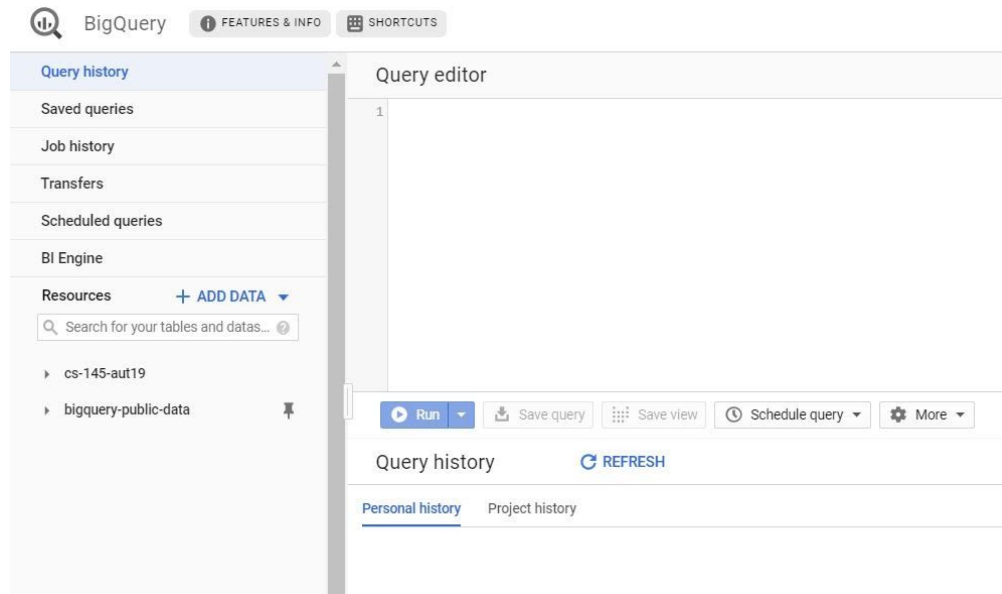
[CREATE](#) [CANCEL](#)

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. In the top-right corner of the bottom panel, click the "Pin Project".



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. This is what you will do for all three projects (and perhaps in your spare time).

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.



Scheduled queries

BI Engine

Resources + ADD DATA

nasa\_wildfire

ncaa\_basketball

mascots

mbb\_games\_sr

mbb\_historical\_teams\_g...

mbb\_historical\_teams\_s...

mbb\_historical\_tournam...

mbb\_pbp\_sr

mbb\_players\_games\_sr

mbb\_teams

mbb\_teams\_games\_sr

team\_colors

bigquery-public-data:ncaa\_basketball

Description

This dataset contains data about NCAA Basketball games, teams, and players. Game data covers play-by-play and box scores back to 2009, as well as final scores back to 1996. Additional data about wins and losses goes back to the 1894-5 season in some teams' cases.

Sportradar: Copyright Sportradar LLC. Access to data is intended solely for internal research and testing purposes, and is not to be used for any business or commercial purpose. Data are not to be exploited in any manner without express approval from Sportradar.

NCAA®: Copyright National Collegiate Athletic Association. Access to data is provided solely for internal research and testing purposes, and may not be used for any business or commercial purpose. Data are not to be exploited in any manner without express approval from the National Collegiate Athletic Association.

Dataset info

Dataset ID	bigquery-public-data:ncaa_basketball
Created	Jan 23, 2018, 12:25:09 PM
Default table expiration	Never
Last modified	Mar 20, 2019, 2:10:55 PM
Data location	US

- In the sidebar, click on one of the tables in the dataset (for example, `mascots`). 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.

mascots [QUERY TABLE](#) [CC](#)

Schema Details Preview

Field name	Type	Mode	Description
id	STRING	NULLABLE	University unique ID from Sportradar
market	STRING	NULLABLE	The university to which the mascot belongs
name	STRING	NULLABLE	The name of the university's team
mascot	STRING	NULLABLE	The name of the university's mascot
mascot_name	STRING	NULLABLE	The proper name of the university's mascot, if available (e.g. a character)
mascot_common_name	STRING	NULLABLE	The type of being or creature that the mascot embodies

## mascots

Schema [Details](#) Preview

### Description

None

### Table info

Table ID	bigquery-public-data:ncaa_basketball.mascots
Table size	55.57 KB
Long-term storage size	55.57 KB
Number of rows	351
Created	Apr 1, 2018, 9:55:30 AM
Table expiration	Never
Last modified	Apr 1, 2018, 9:55:30 AM
Data location	US

mascots [QUERY TABLE](#)

Schema Details [Preview](#)

Row	id	market	name	mascot	mascot_name
1	2959bd24-7007-41ae-a3a3-abdf26888cfc	Tulsa	Golden Hurricane	Hurricane	Captain Cane
2	ad4bc983-8d2e-4e6f-a8f9-80840a786c64	Arizona State	Sun Devils	Devil	Sparky
3	f2d01b77-0f5d-4574-9e49-2a3eaf822e44	Drexel	Dragons	Dragon	Mario the Magnificent
4	b47d10b8-a2a5-47df-a2f9-7bd0b9d51beb	Bradley	Braves	Gargoyle	Kaboom!
5	0113eea0-c943-4fff-9780-ae0fb099e7ef	Canisius	Golden Griffins	Griffin	Petey
6	bdc2561d-f603-4fab-a262-f1d2af462277	Michigan	Wolverines	None	<i>null</i>
7	dfe0d93f-94a5-47fb-b7aa-f74786e09acb	Illinois-Chicago	Flames	Dragon	Sparky

- Click “Query Table” in the main window for the console to pop up, and try to run your query.

Query editor HIDE EDITOR FULL SCREEN

```
1 SELECT market FROM `bigquery-public-data.ncaa_basketball.mascots` LIMIT 1000
```

Processing location: US

Run Save query Save view Schedule query More This query will process 4.5 KB when run. ✓

Query results SAVE RESULTS EXPLORE WITH DATA STUDIO

Query complete (0.1 sec elapsed, cached)

Job information **Results** JSON Execution details

Row	market
1	Tulsa
2	Arizona State
3	Drexel
4	Bradley
5	Canisius
6	Michigan

Rows per page: 100 1 - 100 of 351 First page < > > Last page

## Best Practices

1. Pay attention to the estimated number of bytes read by the query. 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.

Query editor HIDE EDITOR FULL SCREEN

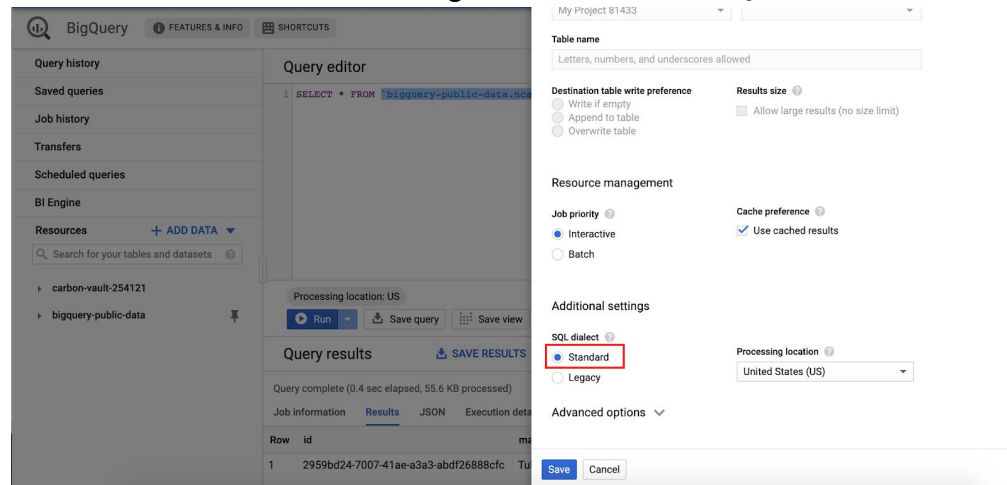
```
1 SELECT market FROM `bigquery-public-data.ncaa_basketball.mascots` LIMIT 1000
```

Processing location: US

Run Save query Save view Schedule query More This query will process 4.5 KB when run. ✓

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).
6. Make sure that you are using **Standard SQL**, not legacy SQL. This should be the default with BigQuery's new Web UI.
  - a. Check the SQL dialect by navigating through “More” -> “Query settings” on the top of the bottom panel.
  - b. Scroll down to “Additional settings”. Make sure the SQL dialect is “Standard”.



## 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