CS145: Data Management and Data Systems

Stanford University, Fall 2019

Getting Started with BigQuery

Table of Contents

- <u>Overview</u>
- <u>Getting Credits</u>
- Initial Setup
- <u>Querying Public Datasets</u>
- Best Practices
- <u>References</u>

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.

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. You'll see the following page:

Thank you for your interest in Goog	le Cloud Platform I	Education Grants. Please fill out the
		on ocogie oloud nationit.
First Name	Last Na	ime
	<u>ا</u>	
School Email		
		@stanford.edu -
If you do not see your domain listed, pleas	e contact your course in	nstructor: shiva@cs.stanford.edu
	we may share the follo	wing information with your educational

- 2. Enter your name and Stanford email address.
- 3. 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 <u>link</u> to verify your email address and a code will be sent to your email account.

Instructor Name: Shiva Shivakumar Email Address: <u>shiva@cs.stanford.edu</u> 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: <u>shiva@cs.stanford.edu</u> 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	S 🔻		

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.



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.

Welcome to the new Billing Overview! Billing p	ermissions, project associations and credit details can now be found of	on the account management page.	DISMISS ACCOUNT MANAGEM
Current month 1–24 September 2019		VIEW REPORT	Billing account Manage CS 145 - Data Management and Data Systems, 018070-0F14E9-998951
Month-to-date total cost 🍘 US\$0.00	End-of-month total cost (forecasted) 🖗 Not enough historical data to project cost		Organisation No organisation Promotional credits <u>View</u> US\$50.00
Cost trend 1 September 2018 – 30 September 2019		VIEW REPORT	
Average monthly total cost US\$0.00			

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 <u>link</u>. You'll see the page below; click "Create" to make a GCP (Google Cloud Platform) project.

Dashboard		
To view this page, select a project.	SELECT	CREATE

 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.

New Project

4	You have 23 projects remaining in your quota. F delete projects. <u>Learn more</u>	Request an increase or
	MANAGE QUOTAS	
Project	name *	
cs145-f	a19	0
Billing a	ccount *	
00 145	Data Management and Data Customer	
CS 145 Any cha	- Data Management and Data Systems rges for this project will be billed to the account that ye	▼ ou select here.
CS 145 Any cha	 Data Management and Data Systems rges for this project will be billed to the account that ye ation * 	▼ ou select here.
CS 145 Any cha Loc	- Data Management and Data Systems rges for this project will be billed to the account that yo ation * organisation	▼ ou select here. BROWSE
CS 145 Any cha Loc Darent of Parent of	- Data Management and Data Systems rges for this project will be billed to the account that yo ation * organisation organisation or folder	▼ ou select here. BROWSE

After creating your project, you'll be brought to an overview page for your project.

3. Go to this <u>link</u>, 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.

Query history	Query editor
Saved queries	1
Job history	
Transfers	
cheduled queries	
BI Engine	
tesources + ADD DATA ▼ 2 Search for your tables and datas @ > cs-145-aut19	n
bigquery-public-data	O Run ▼ ≛ Save query iii Save view ③ Schedule query ✿ More ▼
	Query history C REFRESH

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 <u>https://console.cloud.google.com/bigquery</u>. 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	bigquery-public-data:ncaa_basketball		
Bi Engine Resources + ADD DATA • Search for your tables and datasets • I nasa_wildfire • I ncaa_basketball I ncaa_basketba	Description This dataset contains data Game data covers play-by-scores back to 1996. Addit 1894-5 season in some teas Sportradar: Copyright Spor research and testing purpo commercial purpose. Data approval from Sportradar. NCAA®: Copyright Nationa provided solely for internal any business or commercial without express approval from Sportradar. Dataset info	about NCAA Basketball games, teams, and players. olay and box scores back to 2009, as well as final ional data about wins and losses goes back to the ms' cases. tradar LLC. Access to data is intended solely for internal ses, and is not to be used for any business or are not to be exploited in any manner without express I Collegiate Athletic Association. Access to data is research and testing purposes, and may not be used for al purpose. Data are not to be exploited in any manner rom the National Collegiate Athletic Association.	
mbb players games sr	Dataset ID	bigquery-public-data:ncaa_basketball	
ganeo_si	Created	Jan 23, 2018, 12:25:09 PM	
mbb_teams	Default table expiration	Never	
mbb_teams_games_sr	Last modified	Mar 20, 2019, 2:10:55 PM	
team_colors	Data location	US	

4. 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

Schema Details Preview				
Field name	Туре	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	null
7	dfe0d93f-94a5-47fb-b7aa-f74786e09acb	Illinois-Chicago	Flames	Dragon	Sparky

5. Click "Query Table" in the main window for the console to pop up, and try to run your query.

Q	uery editor				[] FULL SCREEN
1	SELECT market FROM `big	guery-public-data.ncaa_bas)	cetball.mascots` LIM	IT 1000	
	Processing location: US				
	🕞 Run 🔻 📩 Save que	y iiii Save view 🕓 Schedu	le query 👻 🏟 More	 This query will process 	4.5 KB when run. 🥑
Q	uery results	🛃 SAVE RESULTS 🔻 🛛 🍏	EXPLORE WITH DATA ST	TUDIO	
Quer	y complete (0.1 sec elapsed, ca	ached)			
Job i	information Results JS	ON Execution details			
Row	market				
1	Tulsa				
2	Arizona State				
3	Drexel				
4	Bradley				
5	Canisius				
6	Michigan				
		Rows per page: 1	00 🔻 1 - 100 of 351	First page 1	> 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".

0		My Project 81433	•	
BigQuery D FEATURES & INFO	HORTCUTS	Table name		
Query history	Query editor	Letters, numbers, and underscores allowed		
Saved queries	1 SELECT * FROM "bigquery-public-data.nca	Destination table write preference	Results size 🛞	
Job history		Append to table Overwrite table	Allow large results (no size limit)	
Transfers				
Scheduled queries		Resource management		
BI Engine		lab priority	Cache preference	
Resources + ADD DATA •		Interactive	✓ Use cached results	
Q. Search for your tables and datasets 🛛 🔞	h	O Batch		
> carbon-vault-254121 > bigquery-public-data ∓	Processing location: US	Additional settings		
	Query results AVE RESULTS	SQL dialect Standard	Processing location	
	Query complete (0.4 sec elapsed, 55.6 KB processed)	C Legacy	United States (US)	
	Job information Results JSON Execution deta	Advanced options $$		
	Row id ma			
	1 2959bd24-7007-41ae-a3a3-abdf26888cfc Tu	Save Cancel		

References

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