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 \texttt{SELECT column1, column2 ...}, you will save lots of credits over the course of the quarter.

\textbf{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:

   ![Image of Cloud Platform Education Grants](image)

   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.

   First Name
   Last Name

   School Email
   @stanford.edu

   If you do not see your domain listed, please contact your course instructor: steve@cs.stanford.edu

   By clicking “Submit” below, you agree that we may share the following information with your educational institution and course instructor: steve@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.

   Submit

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 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
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
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). Unlike step 2, make sure you are logged in to a **personal Google account, NOT your Stanford account**. From now on, you’ll use your personal Google account with BigQuery, not your Stanford account.
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](#)).
After creating your project, you’ll be brought to an overview page for your project.

3. Go to this link, which is the page for BigQuery’s public datasets. Click the pin in the left sidebar, next to the name “bigquery-public-data” (see screenshot below).
4. Now when you visit the console, you’ll be able to easily find the datasets in your sidebar (see screenshot below).

```
[Image: screenshot of BigQuery console]
```

**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).

2. Click on “bigquery-public-data” and scroll until you find the dataset `ncaa_basketball`. Alternatively, search for `ncaa_basketball` in the search bar above your pinned projects.
3. Click the three dots next to the dataset name, then click “Open”. 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, `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 (see these three tabs in the screenshot below).
<table>
<thead>
<tr>
<th>Field name</th>
<th>Type</th>
<th>Mode</th>
<th>Policy Tags</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>University unique ID from SportRadar</td>
</tr>
<tr>
<td>market</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The university to which the mascot belongs</td>
</tr>
<tr>
<td>name</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The name of the university's team</td>
</tr>
<tr>
<td>mascot</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The name of the university's mascot</td>
</tr>
<tr>
<td>mascot_name</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The proper name of the university's mascot (if available e.g. a character)</td>
</tr>
<tr>
<td>mascot_common_name</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The type of being or creature that the mascot embodies</td>
</tr>
<tr>
<td>tax_subspecies</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The subspecies to which the mascot belongs</td>
</tr>
<tr>
<td>tax_species</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The species to which the mascot belongs</td>
</tr>
<tr>
<td>tax_genus</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The genus to which the mascot belongs</td>
</tr>
<tr>
<td>tax_family</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The family to which the mascot belongs</td>
</tr>
<tr>
<td>tax_order</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The order to which the mascot belongs</td>
</tr>
<tr>
<td>tax_class</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The class to which the mascot belongs</td>
</tr>
<tr>
<td>tax_phylum</td>
<td>STRING</td>
<td>NULLABLE</td>
<td></td>
<td>The phylum to which the mascot belongs</td>
</tr>
</tbody>
</table>

**Table Info**

Table ID: bigquery-public-data.ncaab.basketball_mascots

Table size: 55.57 KB

Long-term storage size: 55.57 KB

Number of rows: 385

Created: Apr 1, 2018, 9:55:00 AM UTC-7

Last modified: Apr 1, 2018, 9:55:00 AM UTC-7

Table expiration: NEVER

Data location: US

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

2. If you are just exploring/trying out queries, use \texttt{LIMIT} to query fewer data. Also, avoid using \texttt{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. BigQuery has a shortcut to run queries. This is CTRL-ENTER on Windows or CMD-RETURN on Mac.

7. Make sure that you are using \texttt{Standard SQL}, not legacy SQL. This should be the default with BigQuery’s Web UI.
   a. Check the SQL dialect by navigating through “More” -> “Query settings” on the top of the bottom panel (see screenshot below).
b. Scroll down to “Additional settings”. Make sure the SQL dialect is “Standard” (see screenshot below).

References

1. [https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators](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](https://cloud.google.com/bigquery/docs/best-practices-costs) for more best practices to save cost