

# Homework 1: SQL Refresher, and BigQuery

This assignment will give you some experience with Google BigQuery and a refresher of SQL queries. We will be using public datasets from [NYC OpenData](#), a repository of free and public datasets provided by the city, from the city government's agencies. In particular, we will be creating a data warehouse storing five tables, each uploaded from a different public dataset that concerns NYC Public Schools.

You will need: Internet access, a coupon code for GCP (which you should have retrieved during assignment 0 before class began), and a GCP account (which you should have retrieved during assignment 0 before class began). Please read through this assignment and make sure you have these three requirements EARLY. Contact me if you have any trouble!

## Setup:

First, sign in to GCP, create a new project, and enable BigQuery. You can call the project whatever you would like, but remember that if you need my help and come to office hours, you will likely need to share the project name with me.

## Enable BigQuery

If you don't already have a Google Account (Gmail or Google Apps), you must [create one](#).

- Sign-in to Google Cloud Platform console ([console.cloud.google.com](https://console.cloud.google.com)) and navigate to BigQuery. You can also open the BigQuery web UI directly by entering the following URL in your browser.

```
https://console.cloud.google.com/bigquery
```

- Accept the terms of service.
- Before you can use BigQuery, you must create a project. Follow the prompts to create your new project.

Choose a project name and make note of the project ID.

### New Project

---

Project Name \*

my-codelab-project ?

Project ID: my-codelab-project-207619. It cannot be changed later. [EDIT](#)

Billing account \*

The project ID is a unique name across all Google Cloud projects. It will be referred to later in this codelab as `PROJECT_ID`.

Next, create a new dataset, with the name `hw1`.

To create a dataset, click the **project name** under the resources pane, then click the **Create dataset** button:

Google Cloud Platform Demo Project

SANDBOX Set up billing to upgrade to the full BigQuery experience. [Learn more](#) DISMISS UPGRADE

BigQuery FEATURES & INFO SHORTCUTS + COMPOSE NEW QUERY

Query history  
Saved queries  
Job history  
Transfers  
Scheduled queries  
BI Engine  
Resources + ADD DATA  
Search for your tables and datas... ?

your-project-id ← Select your project ID from the resources pane

bigquery-public-data

Query editor HIDE EDITOR FULL SCREEN


1

Run Save query Save view Schedule query More

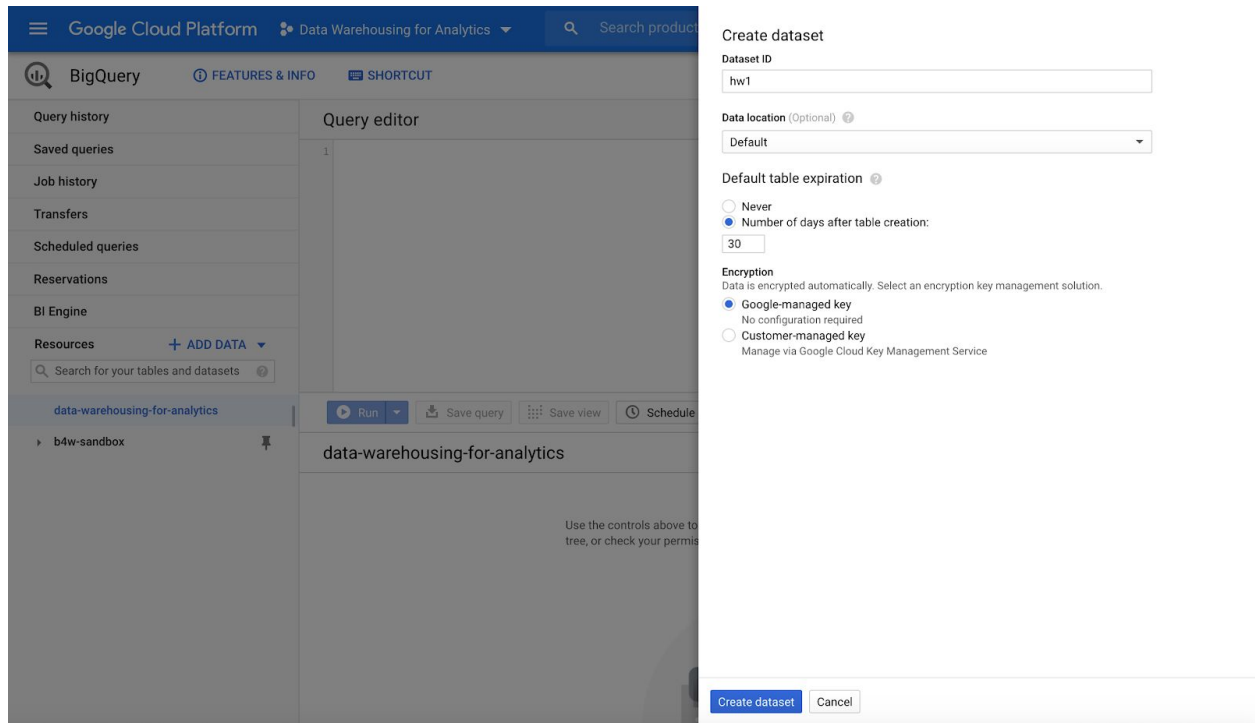
your-project-id Create a dataset CREATE DATASET PIN PROJECT

No datasets available

Use the controls above to create a dataset and start building out your Resources tree, or check your permissions for this project.

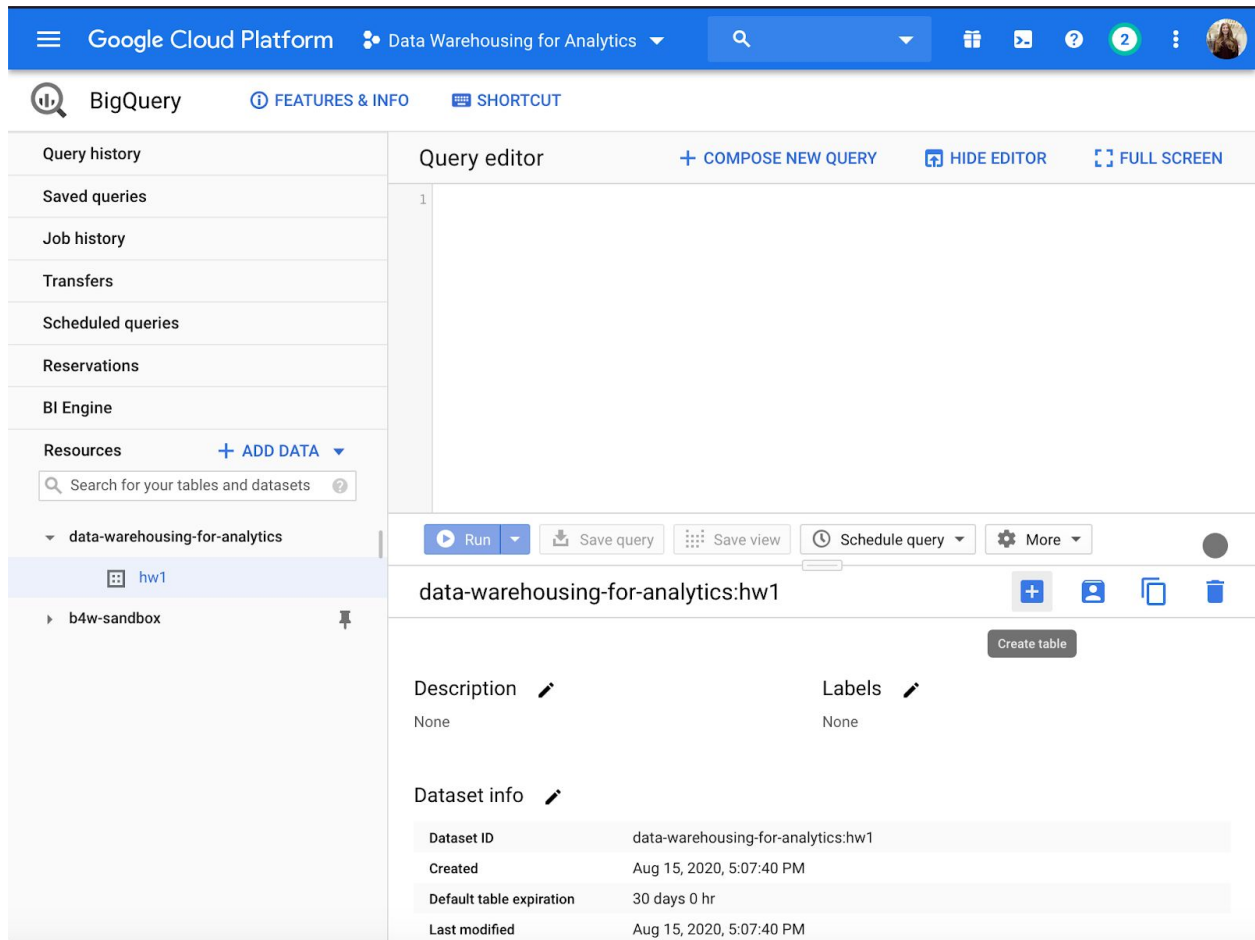


Next, name the dataset "hw1" and set the default expiration date to 30 days.

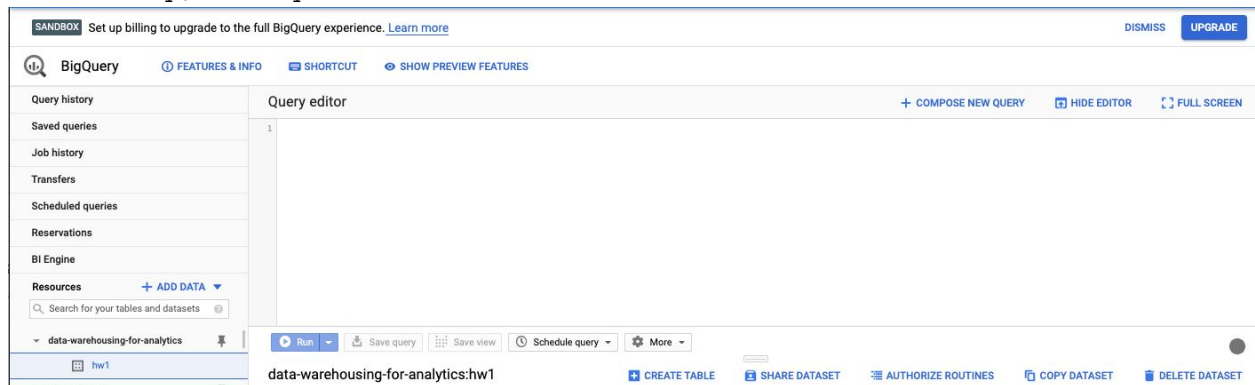


Next we're going to load 5 tables into our dataset.

In the lefthand sidebar, click on your project name and dataset name, so that you see a view like this:



Alternately, it may look like this:



In the navigation bar in the middle of the righthand side of the page, which says "data-warehousing-for-analytics:hw1" in the example image above, click "create table".

You will see a pop up menu like below: select "Google Cloud Storage" as the data source.

### Create table

Source

Create table from:

Empty table

Empty table

Google Cloud Storage

Upload

Drive

Google Cloud Bigtable

Destination

☒ Search for a project name

Project name

Data Warehousing for Analytics

Dataset name

hw1

Table type ?

Native table

Table name

Letters, numbers, and underscores allowed

Schema

☐ Edit as text

+ Add field

Partition and cluster settings

Partitioning: ?

No partitioning

Clustering order (optional): ?

Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Create table

Cancel

Copy this Google Cloud Storage (GCS) bucket object link into the "Select file from GCS bucket" input line:  
`gs://analytical_sql_homework/2018-2019_Daily_Attendance.csv`

### Create table

Source

Create table from:

Google Cloud Storage

Select file from GCS bucket: 

omework/Copy of 2018-2019\_Daily\_Attendance.csv

Browse

File format:

CSV

☐ Source Data Partitioning

Destination

☒ Search for a project

☐ Enter a project name

Project name

Data Warehousing for Analytics

Dataset name

hw1

Table type 

Native table

Table name

Letters, numbers, and underscores allowed

Schema

Auto detect

☐ Schema and input parameters

☒ Edit as text

+ Add field

Partition and cluster settings

Partitioning: 

No partitioning

Create table

Cancel

Under "Table name", put  
Daily\_Attendance

And check the box that is labeled "auto detect" to enable auto-detection of the schema of the table.

## Create table

### Source

Create table from:

Google Cloud Storage

Select file from GCS bucket: ?

☒ omework/Copy of 2018-2019\_Daily\_Attendance.csv

Browse

File format:

CSV

☐ Source Data Partitioning

### Destination

☒ Search for a project

☐ Enter a project name

Project name

Data Warehousing for Analytics

Dataset name

hw1

Table type ?

Native table

Table name

Daily\_Attendance

### Schema

Auto detect

☒ Schema and input parameters

*i* Schema will be automatically generated.

### Partition and cluster settings

Partitioning: ?

No partitioning

Clustering order (optional): ?

Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Comma-separated list of fields to define clustering order (up to 4)

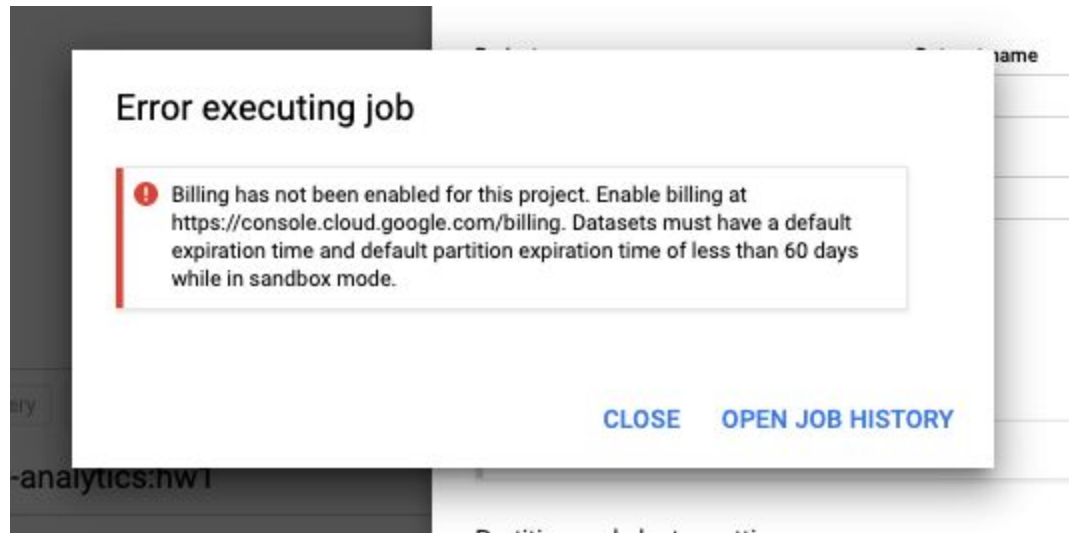
Create table

Cancel

Then click "Create table" to begin the import job.

You may see an error message that looks like this:





If so, you may have not redeemed your GCP coupon. Follow the instructions on Blackboard in the Week 2 Course Documents folder titled "GCP coupon instructions". If, once you have done that, you still get this error, go to <https://console.cloud.google.com/billing> and follow the instructions to *enable billing* for your project.

In a moment, the data will be copied from the public GCS bucket where the example data is stored (I put it there, it is a dataset open to the public) into your BigQuery instance. You can inspect the table by going to the lefthand sidebar in the BigQuery UI, expanding the project name, expanding the dataset name, and clicking on the table name. It will look like this:

The screenshot shows the Google Cloud BigQuery interface. On the left, the 'Resources' panel is expanded to show the 'data-warehousing-for-analytics' dataset, with the 'Daily\_Attendance' table selected. The main area displays the 'Daily\_Attendance' table schema with the following fields:

| Field name | Type    | Mode     | Policy tags | Description |
|------------|---------|----------|-------------|-------------|
| School_DBN | STRING  | NULLABLE |             |             |
| Date       | INTEGER | NULLABLE |             |             |
| Enrolled   | INTEGER | NULLABLE |             |             |
| Absent     | INTEGER | NULLABLE |             |             |
| Present    | INTEGER | NULLABLE |             |             |
| Released   | INTEGER | NULLABLE |             |             |

You should see the table schema (at this point, you will only have one table). Type a query like the following into the "query editor" box in the middle of the page:

```
select * from hw1.Daily_Attendance limit 1;
```

to test that your data was imported correctly. Click "Run". It should look like this:

The screenshot shows the Google Cloud BigQuery interface with the query editor and results. The query entered is:

```
select * from hw1.Daily_Attendance limit 1;
```

The query has been executed successfully, and the results are displayed below. The query complete message indicates that 12.7 MB of data was processed in 0.3 seconds.

| Row | School_DBN | Date     | Enrolled | Absent | Present | Released |
|-----|------------|----------|----------|--------|---------|----------|
| 1   | 01M019     | 20190314 | 256      | 82     | 174     | 0        |

Now, repeat the same steps for the rest of the five tables for this dataset, in the table below:

**NOTE:** BigQuery has been buggy lately. If you get pop-up error messages when following these instructions, try doing whatever you were doing again, just in case the issue was transient (and is gone on the second try). If you get the same error message multiple times, email me for help, with a screenshot of the error message.

**NOTE:** For the last table, you may need to expand the "advanced settings" option at the bottom of the "Create Table" righthand sidebar, and click "allow quoted newlines".

| Table name                                   | GCS link  | NYC Open data link  |
|--|---|---|
| Daily_Attendance                             | gs://analytical_sql_home<br>work/2018-2019_Daily_Att<br>endance.csv                       | <a href="https://data.cityofnewyork.us/Education/2018-2019-Daily-Attendance/x3bb-kg5j">https://data.cityofnewyork.us/Education/2018-2019-Daily-Attendance/x3bb-kg5j</a>                                       |
| School_Demographic_Snapshots                 | gs://analytical_sql_home<br>work/2018-2019_School_De<br>mographic_Snapshot.csv            | <a href="https://data.cityofnewyork.us/Education/2018-2019-School-Demographic-Snapshot/45j8-f6um">https://data.cityofnewyork.us/Education/2018-2019-School-Demographic-Snapshot/45j8-f6um</a>                 |
| DOE_Gifted_and_Talented_Adm<br>issions_Guide | gs://analytical_sql_home<br>work/2019_DOE_Gifted_and<br>_Talented_Admissons_Gui<br>de.csv | <a href="https://data.cityofnewyork.us/Education/2019-DOE-Gifted-and-Talented-Admissions-Guide/k65y-fzgq">https://data.cityofnewyork.us/Education/2019-DOE-Gifted-and-Talented-Admissions-Guide/k65y-fzgq</a> |
| DOE_High_School_Director<br>y                | gs://analytical_sql_home<br>work/2019_DOE_High_Schoo<br>l_Directory.csv                   | <a href="https://data.cityofnewyork.us/Education/2019-DOE-High-School-Directory/uq7m-95z8">https://data.cityofnewyork.us/Education/2019-DOE-High-School-Directory/uq7m-95z8</a>                               |
| DOE_Teacher_Responses                        | gs://analytical_sql_home<br>work/DOE_Teacher_Respon<br>ses.csv                            | <a href="https://data.cityofnewyork.us/Education/2019-Public-Data-File-Teachers/hvwa-bi3h">https://data.cityofnewyork.us/Education/2019-Public-Data-File-Teachers/hvwa-bi3h</a>                               |

Open the NYC Open Data link provided in the table above to inspect the data dictionary provided for each of these data sets. This will give you an idea of what each table holds.

## Questions:

1. Click on each table in the hw1 dataset in the lefthand sidebar. Under the name of the table, there are three pages, "Schema," "Details," and "Preview." Look at the Schema page, and note the name of each column and the data type of

each column in the table. These were *inferred* automatically by BigQuery when we imported the data in each table from Google Cloud Storage, during setup.

Write a CREATE TABLE statement for each of the five tables, using the same column names and types that BigQuery inferred. I want you to create new table names by appending "\_1" to the end of the original table name, i.e.

**Daily\_Attendance\_1.** In addition, while the *Schema* page will show you that a data type is "Integer," that is not actually a valid data type in BigQuery (that's weird! I know! I wish they didn't do that!). You should instead use INT64, which you can see is listed on this page in the BigQuery documentation as a valid data type:

[https://cloud.google.com/bigquery/docs/reference/standard-sql/data-types#numeric\\_types](https://cloud.google.com/bigquery/docs/reference/standard-sql/data-types#numeric_types)

Note: When you write the table name in the CREATE TABLE statement, you need to include the dataset name (hw1). It will look like this:

```
CREATE TABLE hw1.Daily_Attendance_1 ....
```

Play around with the queries in the Query Editor, and when you have one that works for each table, copy and paste it into the document you are using for your homework submission (see the instructions at the bottom of this document for submission instructions and formatting). Remember, I only want the query. (5pts)

2. BigQuery doesn't support primary or foreign keys, but I want you to tell me what columns you would choose to use for primary and foreign keys, for each of the five tables.

Look at the schemas you wrote out in problem 1. Which columns in each of those tables could be a primary key? Which one makes the most sense to choose? Which columns could be foreign keys to other tables?

Write out new CREATE TABLE queries for each of the five original tables, this time appending "\_2" to the table names (i.e. **Daily\_Attendance\_2**), and this time, with keys. Remember- BigQuery won't actually let you run these queries, just write them out here. [5pts]

3. How many unique schools are there in the DOE\_High\_School\_Directory table? (1pt)

4. How many schools listed in the DOE\_Teacher\_Responses table had a total parent response rate of 100%? (1pt)

5. This question has several parts, and is meant to illustrate the difference between inner joins ("JOIN") and outer joins ("LEFT JOIN" or "RIGHT JOIN"). (5pts):

- How many rows are in the table DOE\_Gifted\_and\_Talented\_Admissions\_Guide?
- How many rows are in the table DOE\_Teacher\_Responses?
- Perform an inner join on these two tables. What column should you join on? How many rows are in the result?

- d. Perform a left join on the two tables, where the left side is DOE\_Teacher\_Responses. How many rows are in the result?
- e. Perform a left join on the two tables, where the left side is DOE\_Gifted\_and\_Talented\_Admissions\_Guide. How many rows are in the result?

6. Find the number of public schools with gifted and talented programs that teachers believe have a "Rigorous Instruction Score" below 3.0. You should inspect the data using the "Preview" page for each of the datasets- this will help you figure out if you need to filter out any data from your results in your query. (hint: you may also want to use a CAST clause, which we did not cover in class. Try Googling "cast clause MySQL, to look for tips.) [5pts]

## Instructions on how to format the resulting homework assignment:

Please submit your assignment as a link to a google doc that I have permission to view (email address: emily.mazo@baruch.cuny.edu). I will submit feedback by creating a copy of the document, make notes, and sending you a link to my annotated copy.

For each question, please submit the following:

- \* The query
- \* The results (in the BigQuery UI's "Query Results" header, click "Save Results" and then in the dropdown menu select "copy to clipboard", and paste the results from your clipboard into the google doc)

Copy the query into your google doc submission using Courier New font and at least 12pt font. Please make sure your queries, in your submission, do not cross line boundaries in the document.

## Grading rubric:

The assignment grade will be calculated with the following formula:

$$\text{sum}((\text{category score}) * (\text{category percentage of grade}))$$

For example: An assignment submitted on time, with all questions correct and formatted properly except for question 4, which was not attempted at all-  
 $(0.05 * 1) + (0.9 * (21/22)) + (0.05 * 1) = 96\%$

|          |      |     |     |     |    |
|----------|------|-----|-----|-----|----|
| Category | 100% | 75% | 50% | 25% | 0% |
|----------|------|-----|-----|-----|----|

|   |  |   |   |     |                                |
|---|--|---|---|-----|--------------------------------|
| Timeliness<br>(5% of grade)   | On time or late with approval from instructor  | N/A   | N/A   | N/A | Any other time                 |
| Correctness<br>(90% of grade. This row of the rubric will be applied to each question separately and summed at the end) | The question is answered correctly, the query provided runs in the instructor's BigQuery instance    | The question is not answered correctly, but the query provided runs in the instructor's BigQuery instance and it is clear an attempt was made (i.e., the syntax is correct and the query provided is valid SQL, but perhaps the wrong table/column/SQL function/clause was used). | The question is not answered correctly, and the query provided does not run in the instructor's BigQuery instance, but it is clear an attempt was made (i.e. the correct table/column/SQL function/clause was used, but there is a syntax bug). | N/A | The question was not attempted |
| Formatting<br>(5% of grade, to make my life easier, please be nice to me)   | Formatting instructions are followed, Google doc submission can be opened by instructor on first try | N/A   | N/A   | N/A | Anything else                  |