Graduate School of Business

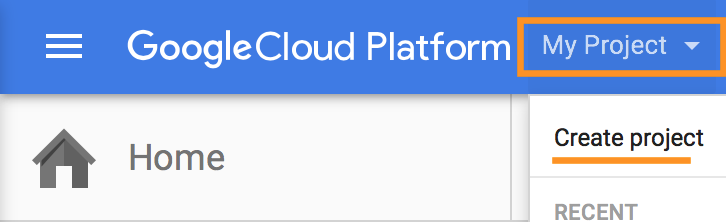
Fordham University

ISGB 7990 – Big Data Analytics – Fall 2016

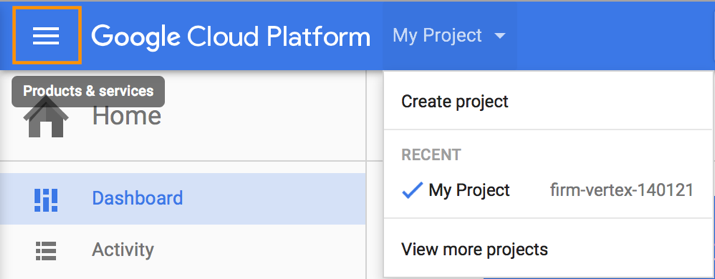
**Workshop: Getting Started with Google BigQuery**

**HOW TO SIGN UP**

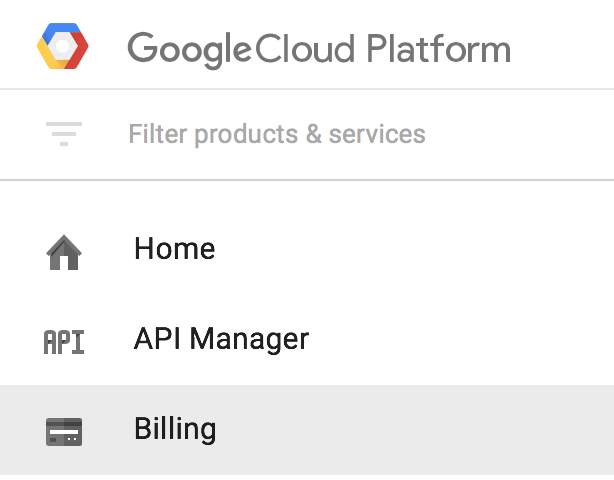
1. Log in to the Google Cloud Platform. (<https://console.cloud.google.com/home/dashboard>) and log in your Google account. You can log in with your Fordham email account.
2. Next to Google Cloud Platform title, Click the drop down menu and choose “**Create project**” to start



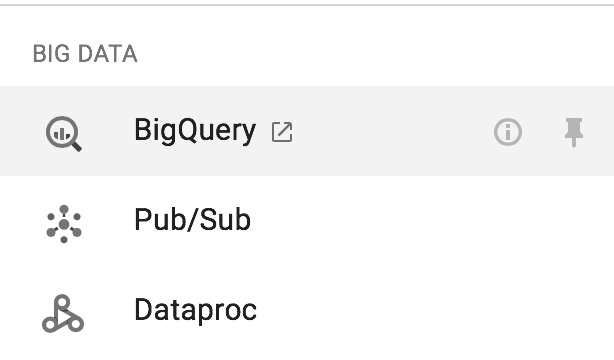
1. After you create the project, Click the product and service button on the right up corner to open navigation bar.



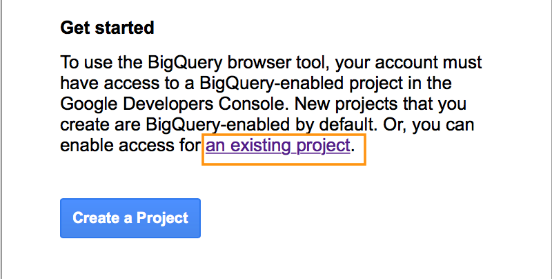
1. Click the billing tag and add a billing account, follow the instruction on the web page. Don't forget to link your project to billing account



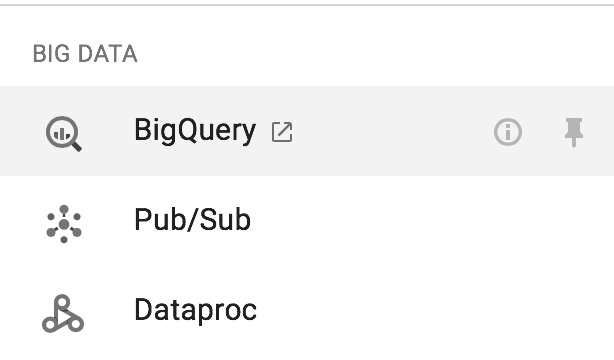
1. Open the product and service bar again and scroll down to the bottom. Click the BigQuery to go to google big query console.



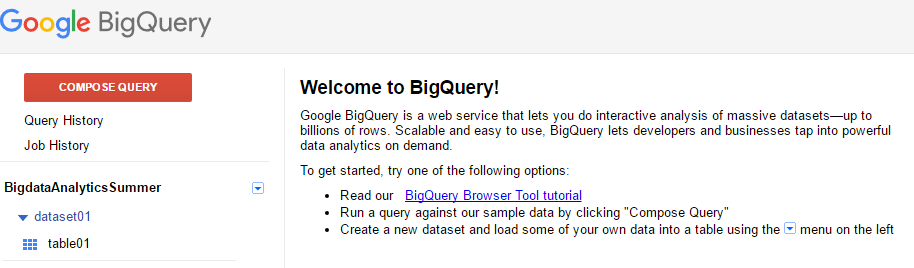
1. In the welcome page, click “**an existing project**” link.



1. Select the project you just create and disregard the credential.
2. Go back to **Google cloud Platform**, open the product and service menu and choose google **bigquery** to start the console.



**Following is a screenshot of BigQuery Web Interface.**



On the top of left-hand side menu bar there’s a big red button “**COMPOSE QUERY**”, by clicking which you can write queries in the popped out window.

In the middle under your API Project you can check all your datasets and tables.

On the bottom these are some sample large dataset provided by Google.

**Example 1: Find the 10 heaviest newborn babies born recently.**

In this example, we’ll use the Google sample dataset, Natality, which recorded over 137 million newborn babies since 1969.

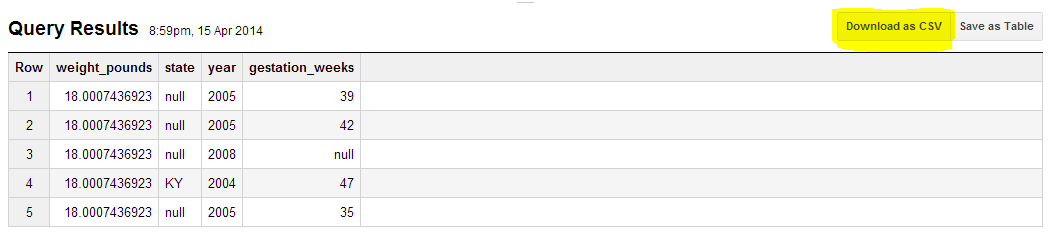
1. Click the red **Compose Query** button．
2. In the “New Query” window, type the following SQL query:

**SELECT weight\_pounds, state, year, gestation\_weeks FROM publicdata:samples.natality**

**ORDER BY weight\_pounds DESC LIMIT 10;**



Click “**Run Query**” button.

1. Check your result below. Keep in mind that you can always download your result into your local disk.
2. Download your result as “**sample 1.csv**”

**Example 2: List the word count of each work of Shakespeare**

In this example, we’ll use another Google sample dataset, Shakespeare, which recorded all the words Shakespeare has used in his works. Click the red **Compose Query** button to start:

1. In the “**New Query**” window, type the following SQL query:

**SELECT corpus, sum(word\_count) AS wordcount FROM publicdata:samples.shakespeare GROUP BY corpus ORDER BY wordcount DESC;**



Click “**Run Query**” button.

Check your result below. 

1. Download your result as “**sample 2.csv**”

**Example 3 - Simple calculation: calculate the daily mean temperature of the world in 1990.**

In this example, we’ll use the**gsod** dataset, which is the one recorded the worldwide weather data collected by The US National Climatic Data Center since 1929. Every day over 9000 worldwide stations collect and transfer the weather data under the GSOD (GLOBAL SURFACE SUMMARY OF DAY)program. For more information please visit <http://www7.ncdc.noaa.gov/CDO/GSOD_DESC.txt>

In this example we will calculate the average temperature by adding all the mean temperature reported by different stations together and then divide it by the number of stations.

1. In the “**New Query**” window, type the following SQL query:

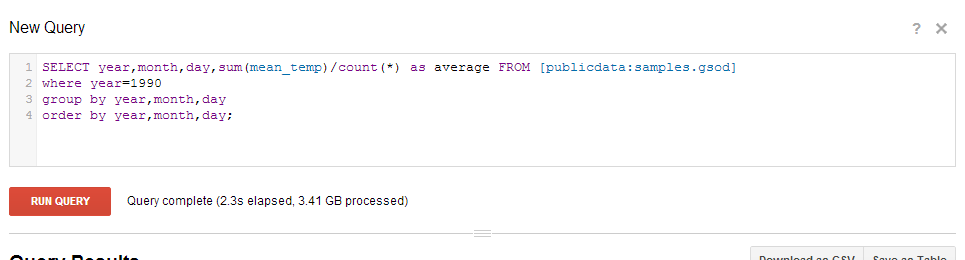
**SELECT year,month,day,sum(mean\_temp)/count(\*) as average FROM publicdata:samples.gsod**

**where year=1990**

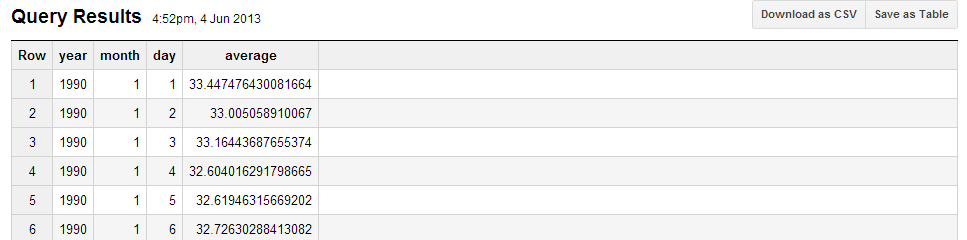
**group by year,month,day**

**order by year,month,day;**

1. Click “**Run Query**” button.



1. Check your result below.



1. Download your result as “sample 3.csv”

**Example 4: Nested Select**

We can filter the result of example 3 by running a nested select, which basically means select something from the result of another query. In this case we will find the days with highest average degree in that year.

1. In the “**New Query**” window, type the following SQL query:

**select month,day, average from**

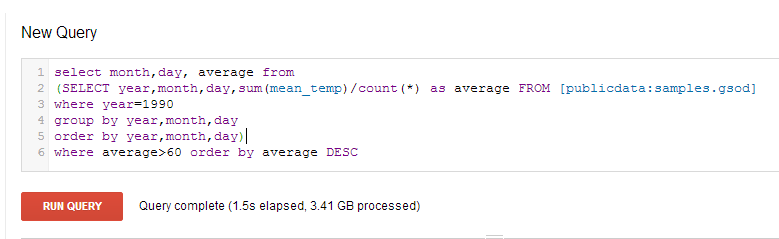
**(SELECT year,month,day,sum(mean\_temp)/count(\*) as average FROM [publicdata:samples.gsod]**

**where year=1990**

**group by year,month,day**

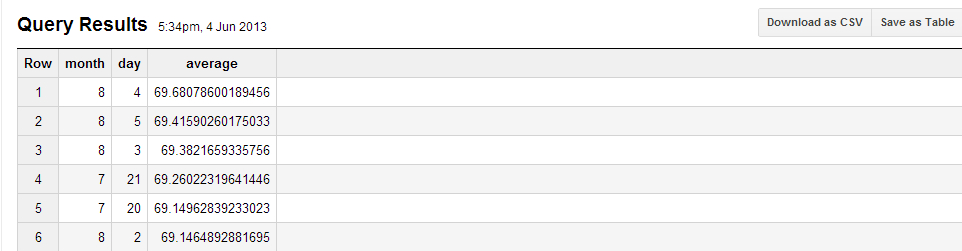
**order by year,month,day)**

**where average>60 order by average DESC;**



Click “**Run Query**” button.

1. Check your result below.



1. Download your result as “**sample 4.csv**”

**Example 5: upload and analyze your own dataset**

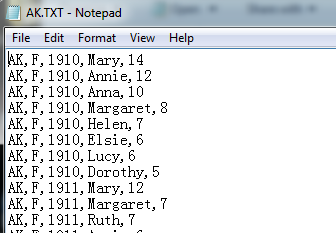
In this example we will first upload a data table and then perform some basic analysis.

Data file used in this sample: **namesbystate.zip** (<http://www.ssa.gov/OACT/babynames/state/namesbystate.zip>)

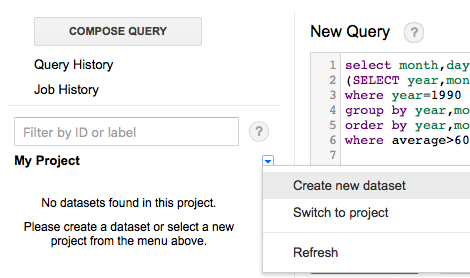
This is the US Government records lists of baby names given to children born each. Download the national data file from the government's Popular Baby Names page and unzip it. This zipped file consists of several comma-delimited value (CSV) files, each listing the baby names for a single year. The file names list the year described by that file.

**The format of each CSV file is state, gender, year, name, count** where gender is either 'M' or 'F', and count is the number of children given that name in that year.

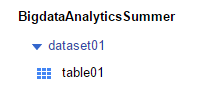
1. Download and unzip the data file. The unzipped file should be like this:



1. Go to your BigQuery web interface, expand the triangle button next to your project name, and click “create new dataset”. Type in whatever your dataset name is then click OK.



1. Now you should be able to see a newly created dataset under your project name.



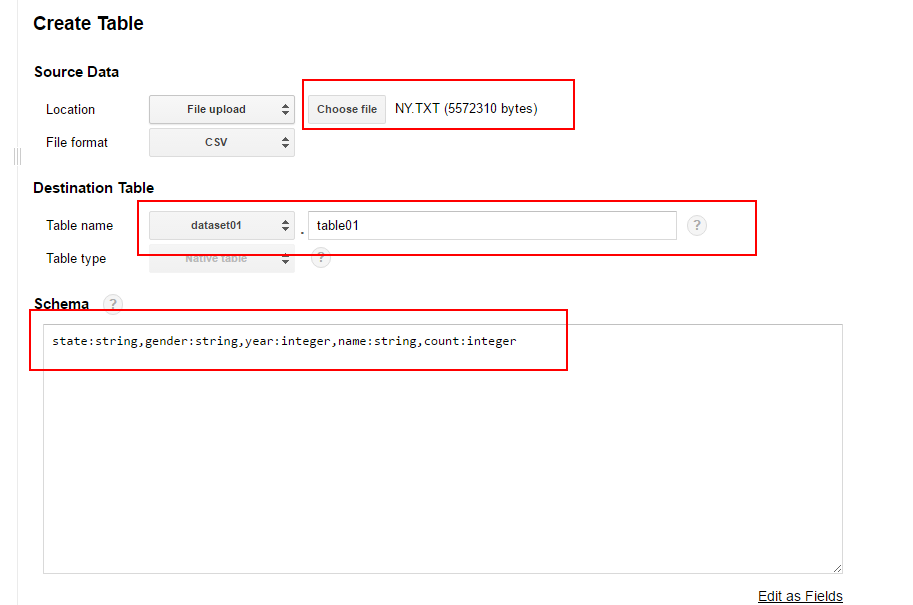
Pin your mouse to that name and you should see two little button

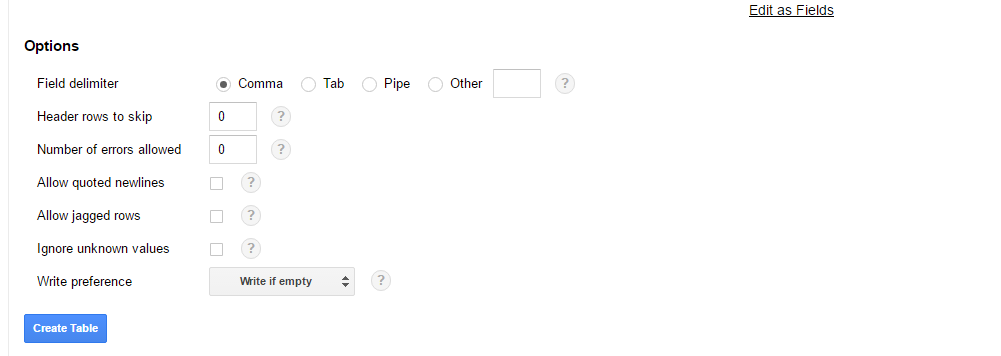


Click the plus mark “+”, an import window popped out. Give your table a name then click next.

Set your file type as csv; Click “choose file” button to locate the name data file we just downloaded (only choose one state, e.g. NY). Then click next.

Make sure your settings same as screenshot below





1. Define your data schema. As we mentioned before, the file has 5 columns：state, gender, year, name, count. Among them “state”, “gender” and “name” are string, while “year” and “count” are integer. So Define your schema as below:

state:string,gender:string,year:integer,name:string,count:integer

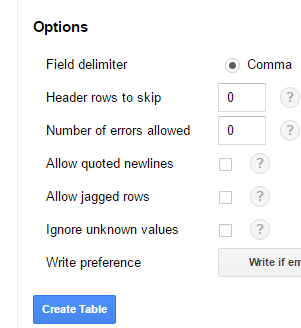
(no space).

**Remember to edit schema as text**

Here each [name:type] pair defines one column. The word before Colon is column name and can be any words reasonable. The word after colon defines column type. Most commonly used type include string (for text), integer (for integer number), float (for float number), Boolean (True or False).

Click “next”

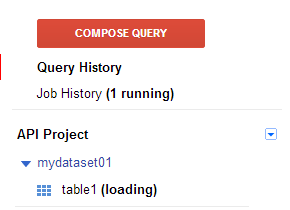
1. Set delimiter as comma; then click **submit.**



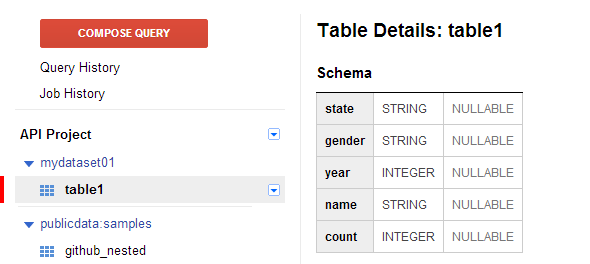
(Notice: always choose the right delimiter that fit your data file. If your file is delimited by tab, choose tab instead.

In this example our data file doesn’t have a header row. If your file has a header row, don’t forget to change “header rows to skip” option.)

Wait a few seconds for Google to load your file



Once loading is completed, you can click your table name to check details:



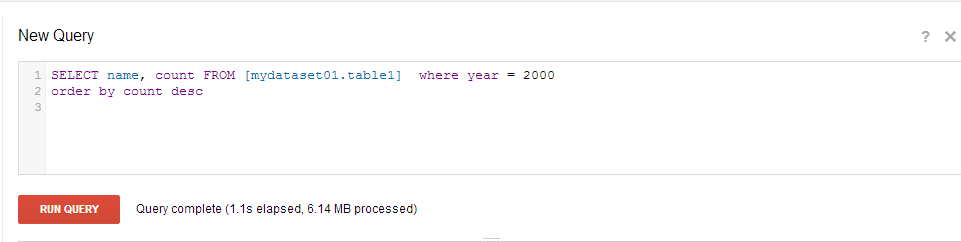
1. Run a simple query for analysis on your newly uploaded data table.

Find the most popular name in year 2000:

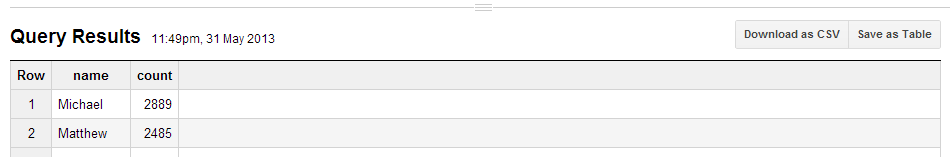
* Type the following sql query into query box then click run query

SELECT name, count FROM [dataset01.table01] where year = 2000

order by count desc;



* Check your result below



* Download your result as “**sample 5.csv**”