

Big Scale Analytics

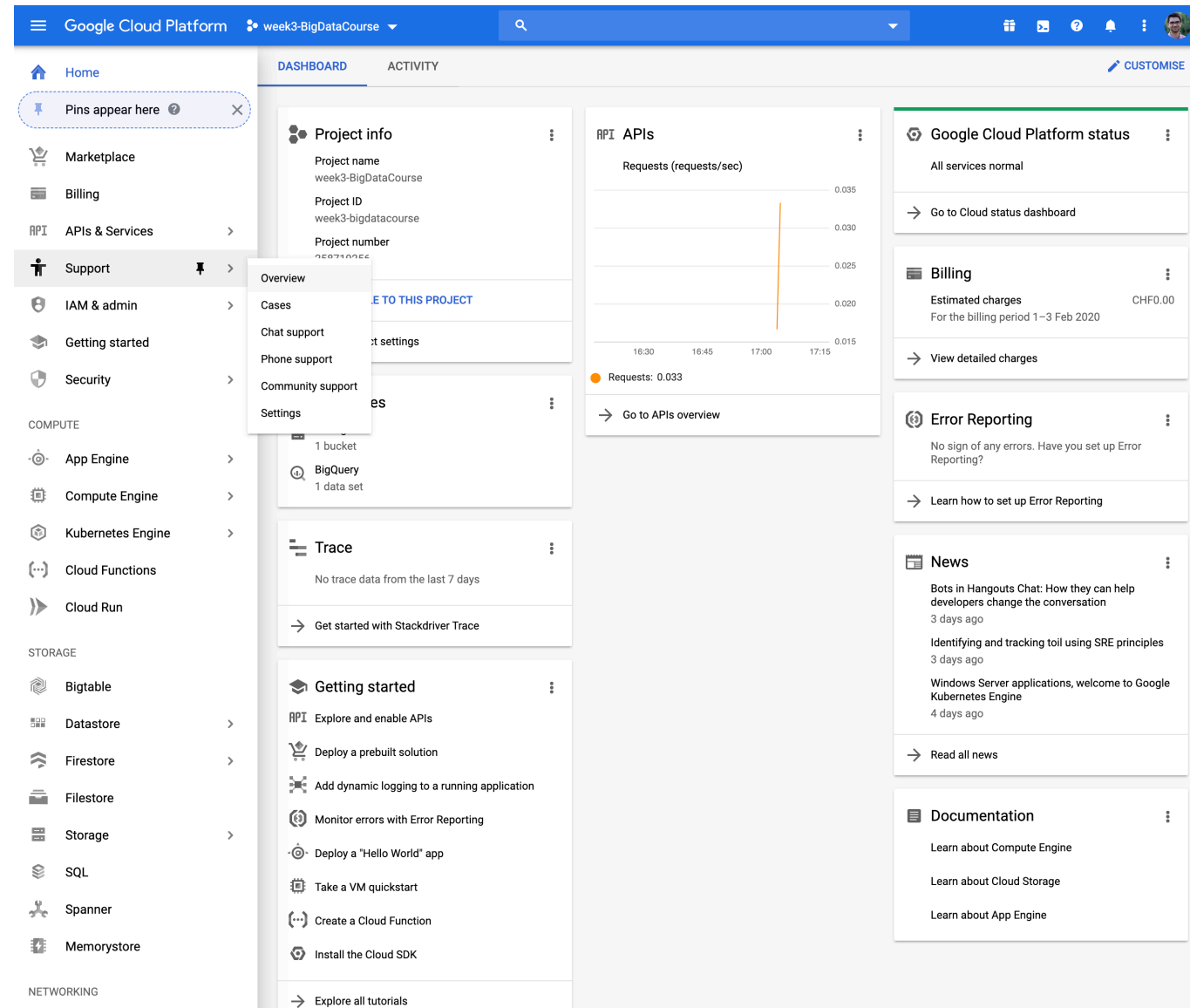
Week 4: How to use BigQuery from Google cloud

Create an account on google cloud

- Go to <https://cloud.google.com/> and create a new account. (you may use your existing Google account)
- Choose the free subscription plan
- You will get \$300 credit for one year
- You will be asked to enter your credit card info, but don't worry! You won't be charged unless you upgrade to a paid plan.

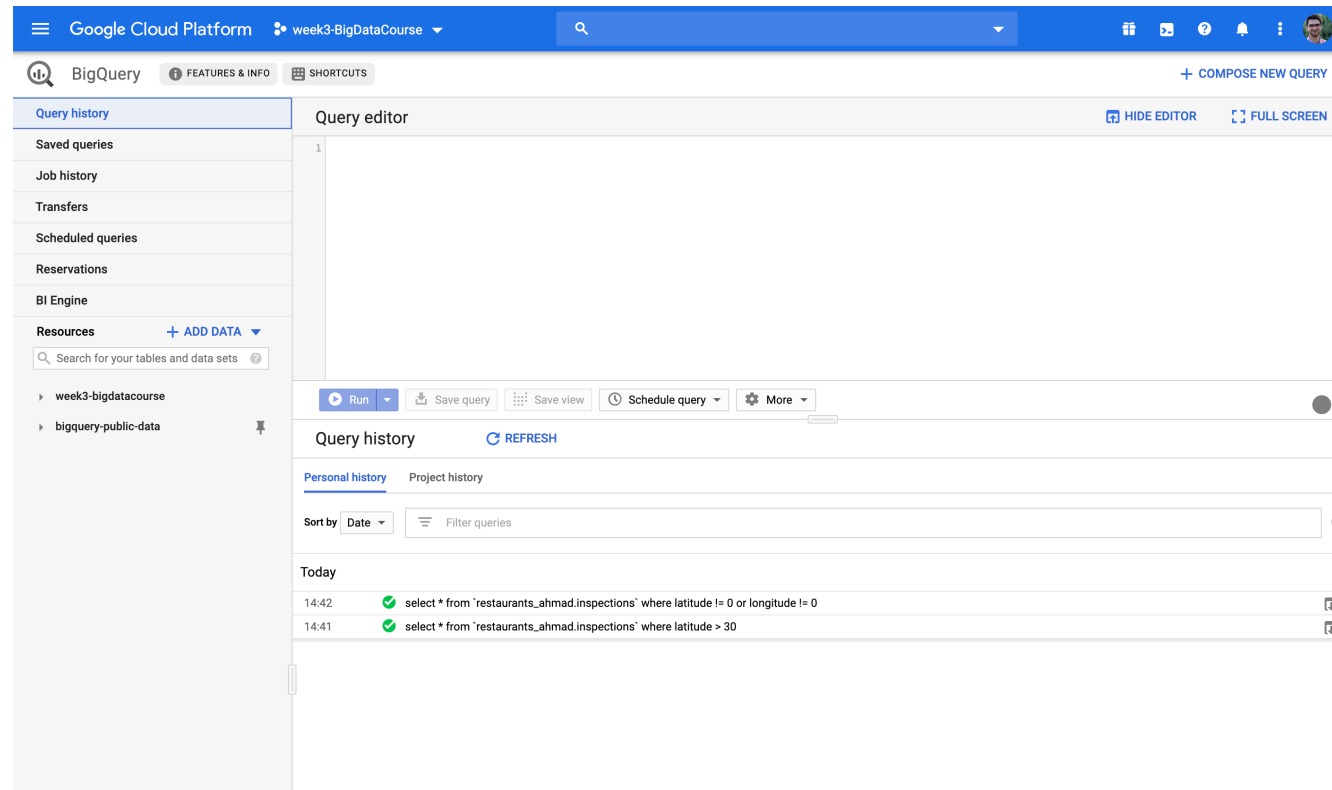
The web UI console

- After creating your account , you can see the console
- You can access several services from the panel on the left
- Try to play with this console a little bit!



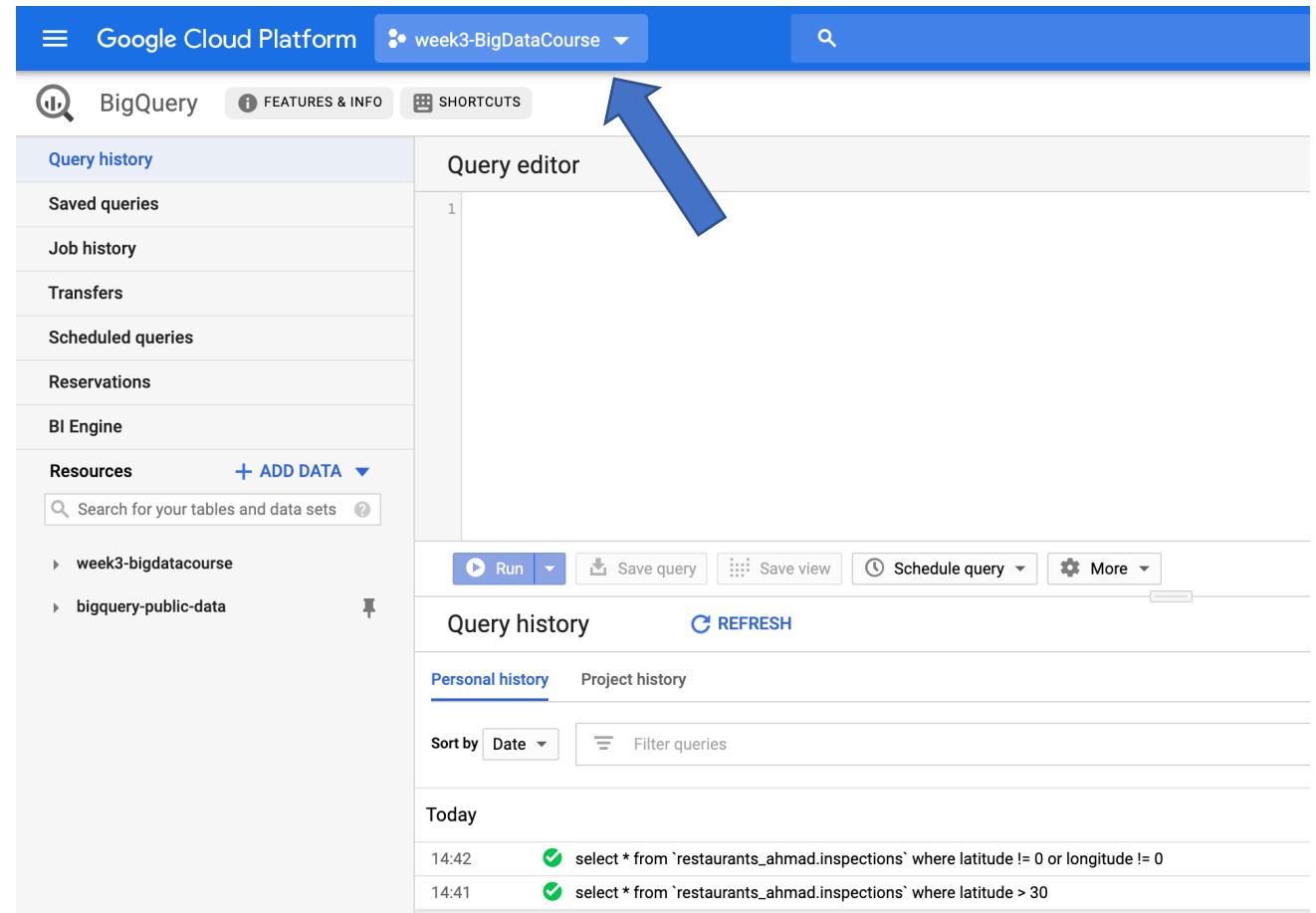
Big Query

- From the panel on the left click on BigQuery. You will see a page like this:



Create you own project in BigQuery

- On the top left you can see the project name which is set by default. But you can make your own project. Just click on it and then in the window that pops up select New Project.

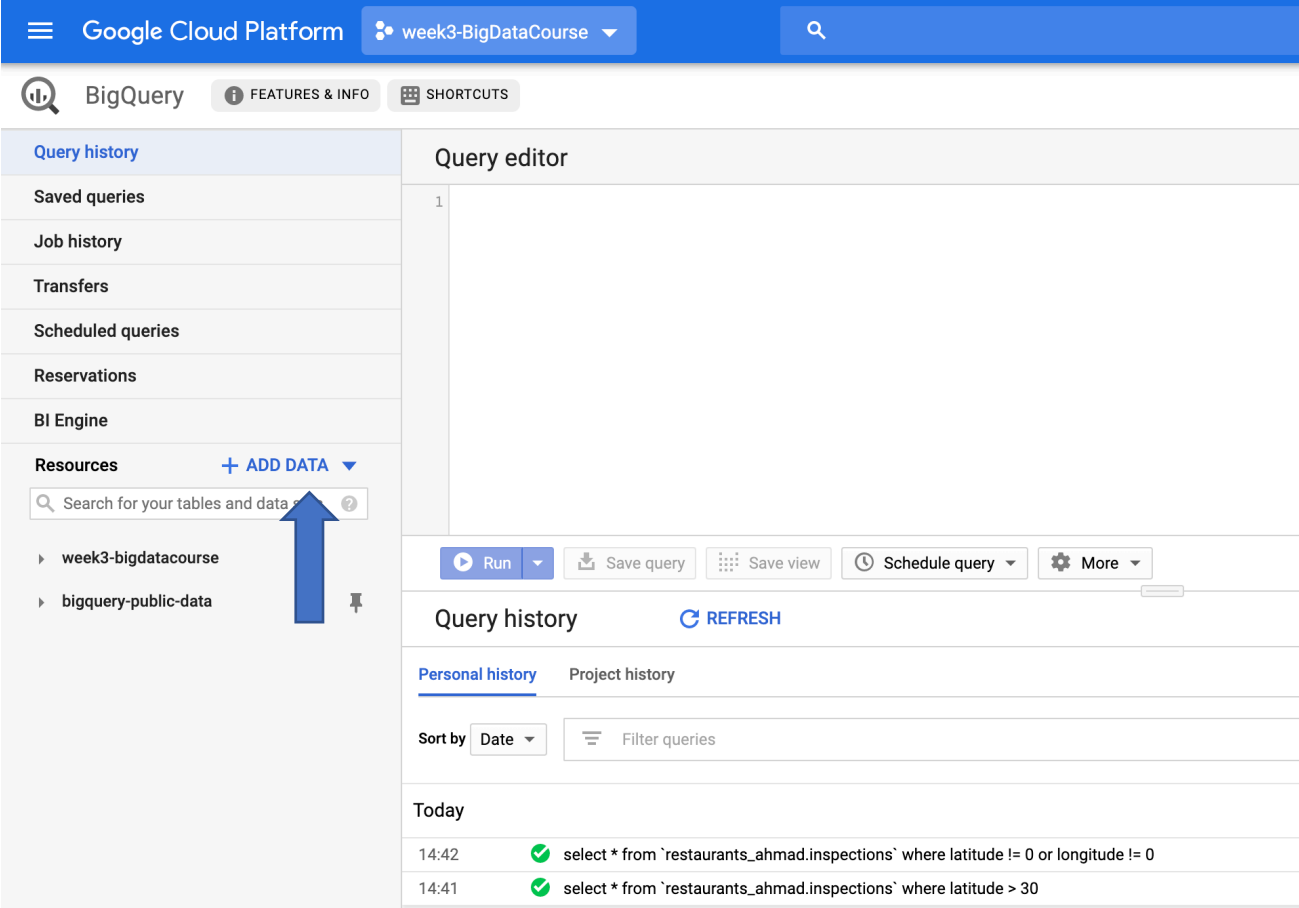


The screenshot shows the Google Cloud Platform BigQuery interface. At the top, the Google Cloud Platform logo and the project name 'week3-BigDataCourse' are displayed. A blue arrow points to the project name. Below the header, the BigQuery logo and 'FEATURES & INFO' and 'SHORTCUTS' tabs are visible. The left sidebar contains a list of navigation items: Query history, Saved queries, Job history, Transfers, Scheduled queries, Reservations, BI Engine, and Resources. The Resources section is expanded, showing a search bar and a list of projects: 'week3-bigdatacourse' and 'bigquery-public-data'. The main area is the Query editor, which is currently empty. Below the Query editor, there are buttons for Run, Save query, Save view, Schedule query, and More. The bottom section shows the Query history, with tabs for Personal history and Project history. The Personal history tab is selected, showing a list of queries with their execution times and results.

Time	Status	Query
14:42	✓	select * from `restaurants_ahmad.inspections` where latitude != 0 or longitude != 0
14:41	✓	select * from `restaurants_ahmad.inspections` where latitude > 30

Public Datasets in Google cloud

- There are several public datasets available within Google cloud which you can use them
- To do so, just click on the ADD DATA button in the Resources tab on the left and then explore the datasets available.
- For instance you can search for “usa_names” or just “names” and explore the dataset of names selected for newborns in USA in different years



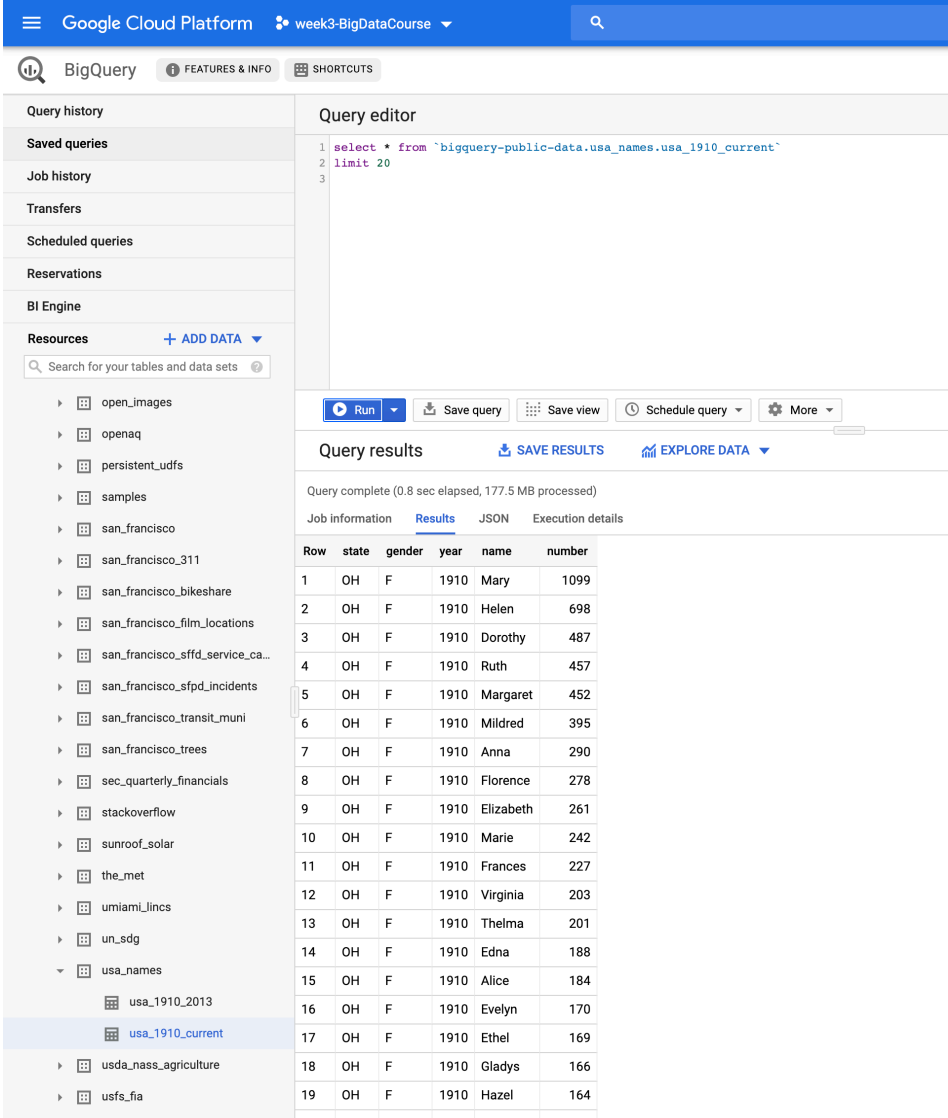
The screenshot displays the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'week3-BigDataCourse', and a search icon. The left sidebar contains a 'Query history' section with links to 'Saved queries', 'Job history', 'Transfers', 'Scheduled queries', 'Reservations', and 'BI Engine'. Below these is the 'Resources' section, which includes a search bar and a list of datasets: 'week3-bigdatacourse' and 'bigquery-public-data'. A blue arrow points to the '+ ADD DATA' button in the Resources section. The main area is the 'Query editor', which has a large text area for writing queries. Below the editor are buttons for 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. At the bottom, the 'Query history' section shows a list of recent queries with timestamps and SQL snippets.

Time	Status	Query
14:42	✓	select * from `restaurants_ahmad.inspections` where latitude != 0 or longitude != 0
14:41	✓	select * from `restaurants_ahmad.inspections` where latitude > 30

Run sample queries using the query editor

- Run some sample queries on the `usa_1910_current` table using the query editor
- Pay attention to the way the table name is specified in the example query in the screenshot

Project_name.dataset.table.name



Google Cloud Platform week3-BigDataCourse

BigQuery FEATURES & INFO SHORTCUTS

Query history

Saved queries

Job history

Transfers

Scheduled queries

Reservations

BI Engine

Resources + ADD DATA

Search for your tables and data sets

Query editor

```
1 select * from `bigquery-public-data.usa_names.usa_1910_current`
2 limit 20
3
```

Run Save query Save view Schedule query More

Query results SAVE RESULTS EXPLORE DATA

Query complete (0.8 sec elapsed, 177.5 MB processed)

Job information Results JSON Execution details

Row	state	gender	year	name	number
1	OH	F	1910	Mary	1099
2	OH	F	1910	Helen	698
3	OH	F	1910	Dorothy	487
4	OH	F	1910	Ruth	457
5	OH	F	1910	Margaret	452
6	OH	F	1910	Mildred	395
7	OH	F	1910	Anna	290
8	OH	F	1910	Florence	278
9	OH	F	1910	Elizabeth	261
10	OH	F	1910	Marie	242
11	OH	F	1910	Frances	227
12	OH	F	1910	Virginia	203
13	OH	F	1910	Thelma	201
14	OH	F	1910	Edna	188
15	OH	F	1910	Alice	184
16	OH	F	1910	Evelyn	170
17	OH	F	1910	Ethel	169
18	OH	F	1910	Gladys	166
19	OH	F	1910	Hazel	164
20	OH	F	1910	Martha	164

Create your own dataset in BigQuery

- It is possible to upload data to google cloud and create your own data-set in BigQuery
- For that, first click on your project name on the right panel
- Then click in the CREAT DATASET button
- In the window that pops up, choose a proper name for your dataset
- It is better to choose locations in EU to have less latency when querying your data
- Notice that a dataset can contain several tables (eg, csv files that you upload)
- Check out the screen shot in the next slide

Create Dataset

Google Cloud Platform

week3-BigDataCourse

BigQuery

FEATURES & INFO

SHORTCUTS

+ COMPOSE NEW QUERY

Query history

Saved queries

Job history

Transfers

Scheduled queries

Reservations

BI Engine

Resources

+ ADD DATA

Search for your tables and data sets

week3-bigdatacourse

restaurants_ahmad

inspections

bigquery-public-data

austin_311

austin_bikeshare

Query editor

HIDE EDITOR

FULL SCREEN

1 select * from `bigquery-public-data.usa_names.usa_1910_current`

2 limit 20

3

Run

Save query

Save view

Schedule query

More

week3-bigdatacourse

CREATE DATASET

PIN PROJECT

This query will access 177.5 MB when run.

Data sets and tables available

Use the Resources tree to view your data or create a new dataset using the controls above.

Create table

Click on the dataset you just created. Then click on CREAT TABLE to upload a csv file.

The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'week3-BigDataCourse', a search bar, and various utility icons. Below the navigation bar, the 'BigQuery' section is active, with links to 'FEATURES & INFO' and 'SHORTCUTS'. A '+ COMPOSE NEW QUERY' button is also visible.

The main interface is divided into two main sections: a left sidebar and a central query editor. The sidebar contains a 'Query history' section and a 'Resources' section. Under 'Resources', there is a search bar and a list of datasets. The 'week3-bigdatacourse' dataset is expanded, showing a sub-dataset 'week3_datasets' which is highlighted. Below it, the 'bigquery-public-data' dataset is also expanded, showing sub-datasets 'austin_311', 'austin_bikeshare', and 'austin_crime'.

The central query editor displays a SQL query: `1 select * from `bigquery-public-data.usa_names.usa_1910_current`
2 limit 20
3`. Below the query editor, there are buttons for 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. A blue arrow points to the '+ CREATE TABLE' button, which is located below the 'Run' button. To the right of the 'CREATE TABLE' button, there are buttons for 'SHARE DATASET', 'COPY DATA SET', and 'DELETE DATASET'. A status message at the bottom right indicates 'This query will process 177.5 MB when run.' with a green checkmark.

At the bottom of the interface, there are sections for 'Description' and 'Labels'. The 'Description' section shows 'None' and the 'Labels' section shows 'None'.

Create table: World cup data

- You can upload different formats like csv, json, etc.
- The schema could be automatically inferred by BigQuery, however if there are problems then you would probably need to specify the schema
- As an example try to upload the Teams.csv and Players.csv from previous week and create two tables under the dataset you made
- Try to run some queries on these tables

Query editor

```
1 select * from `week3-bigdatacourse.week3_datasets.Teams`  
2 where ranking < 30  
3 order by ranking asc  
4
```

Access BigQuery tables from Python

In Colab:

- You need to use the google.cloud package (already available in colab)
- You would also need to give your credentials. It is easy to do this in colab. Run the following code:

```
from google.colab import auth
auth.authenticate_user()
print('Authenticated')
```

- You will be redirected to a page (and possibly you need to sign in to your google account). There is a key there, just copy it and paste it in colab in the box that appeared by running the above code.
- You're all set to connect to BigQuery client

Access BigQuery tables from Python

In your local machine:

- Follow the instruction [here](#) to install the client library and setting up the authentication.
- Edit the `bash_profile` file in mac (or `.bash_rc` file in linux) and add the line below to it. Instead of `[PATH]`, enter the path to the json file you just downloaded (which contain your credentials)

```
export GOOGLE_APPLICATION_CREDENTIALS="[PATH]"
```

- Save your changes and do `source .bash_profile` to apply the changes.
- Make sure to run your python script or jupyter notebook from the same shell that you create the above environment variable