Google BigQuery

I. Load the data into a Google BigQuery endpoint for analysis

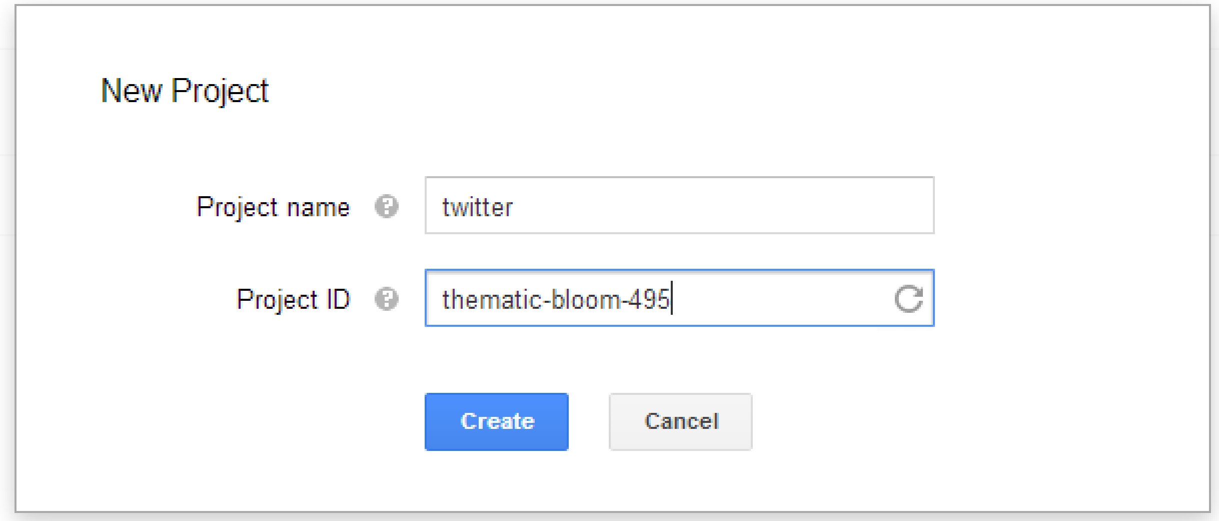
Download the sample files by copying and pasting these URLs into browser to download these two files and then upload to Google Cloud Storage

https://storage.googleapis.com/uwclddata230publichw2017/week07hw/seq0fix.csv

https://storage.googleapis.com/uwclddata230publichw2017/week07hw/seq1fix.csv

To load the data into Google BigQuery the first thing you need to do is upload the files into Google Cloud Storage

If you have not already done so you might have to create a New Project since the Cloud Services and billing are Project based. Here is an example to creating a project called twitter

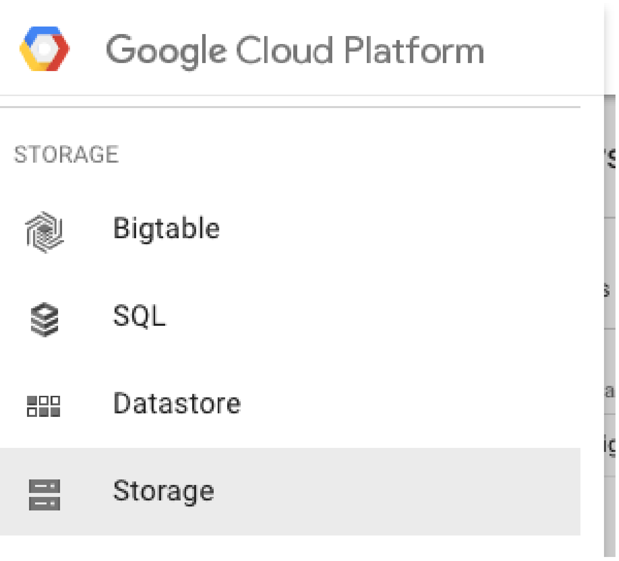


New Project

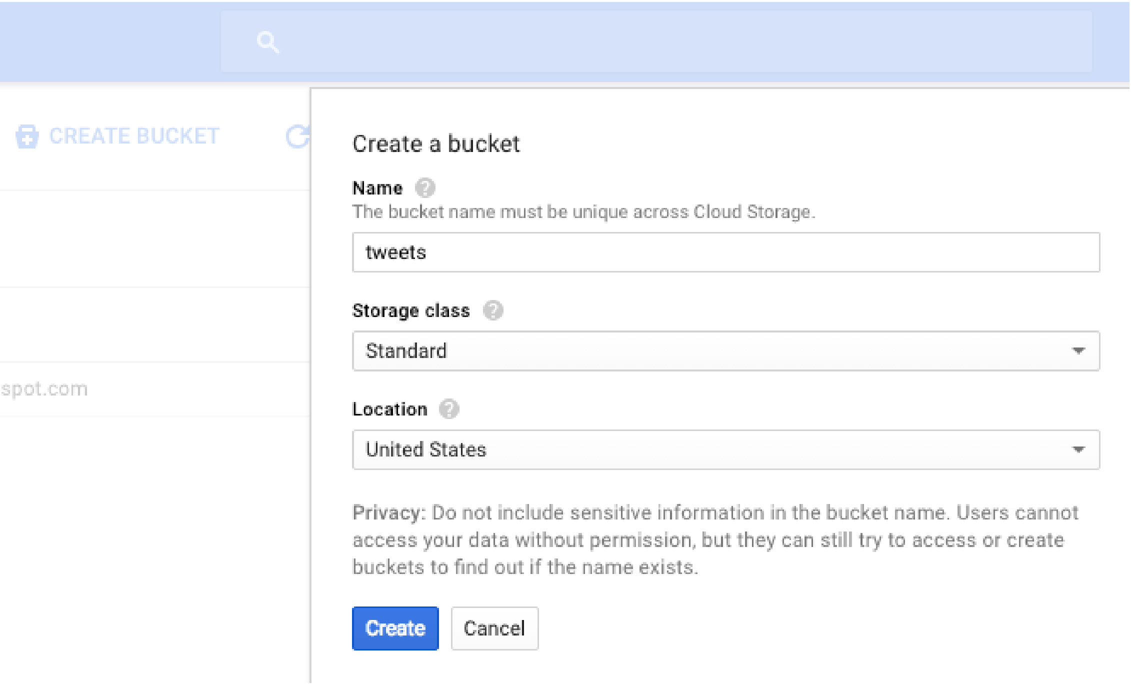
Project name: twitter

Project ID: thematic-bloom-495

Create a Storage Bucket



Click on **CREATE BUCKET** and name your bucket something like tweets and click **Create**



Create a bucket

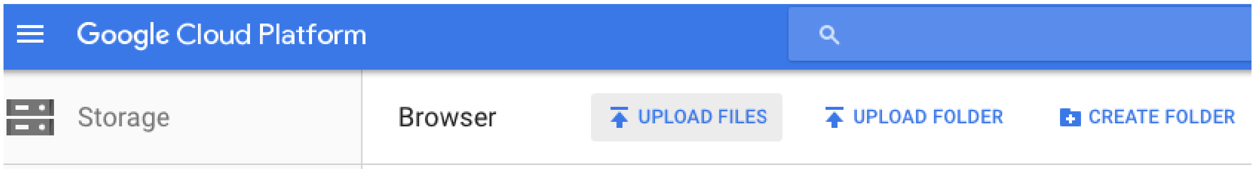
Name
The bucket name must be unique across Cloud Storage.
tweets

Storage class
Standard

Location
United States

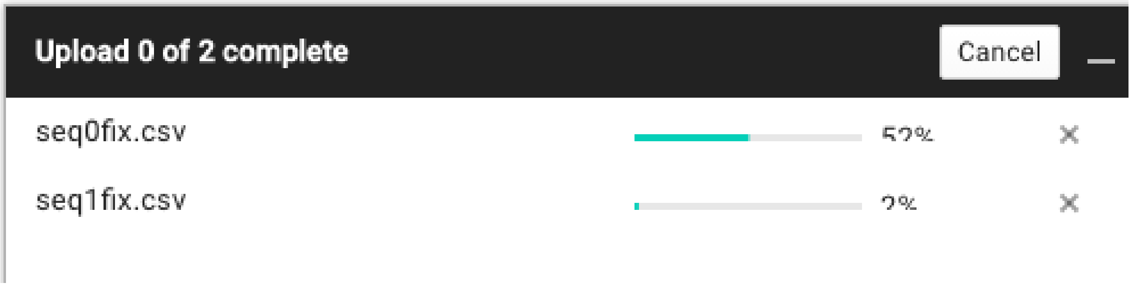
Privacy: Do not include sensitive information in the bucket name. Users cannot access your data without permission, but they can still try to access or create buckets to find out if the name exists.

As I mentioned earlier Google has a nice browse-based upload for Files or the contents of a Folder



Upload the 2 files or the folder with the 2 files

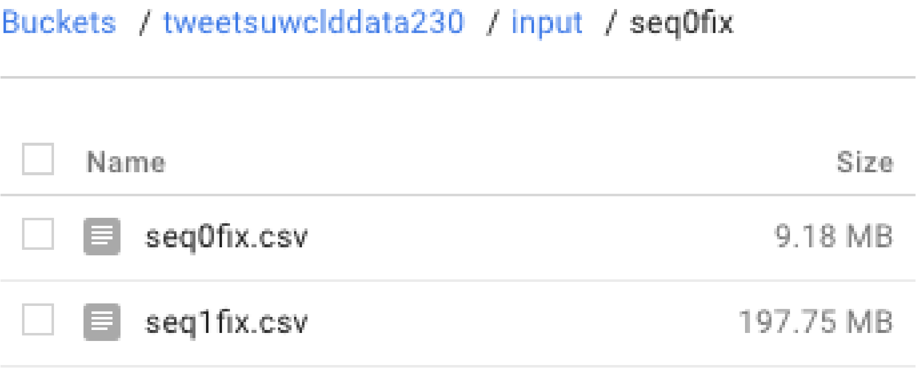
When you select the folder the files start to upload and provide you with a status



Upload 0 of 2 complete

seq0fix.csv	52%
seq1fix.csv	2%

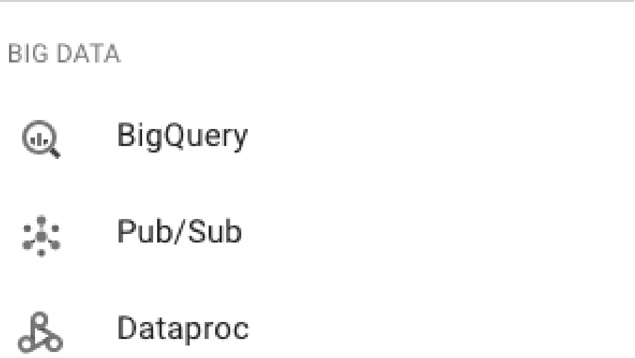
You only have seq0fix and seq1fix



Buckets / tweetsuwclddata230 / input / seq0fix

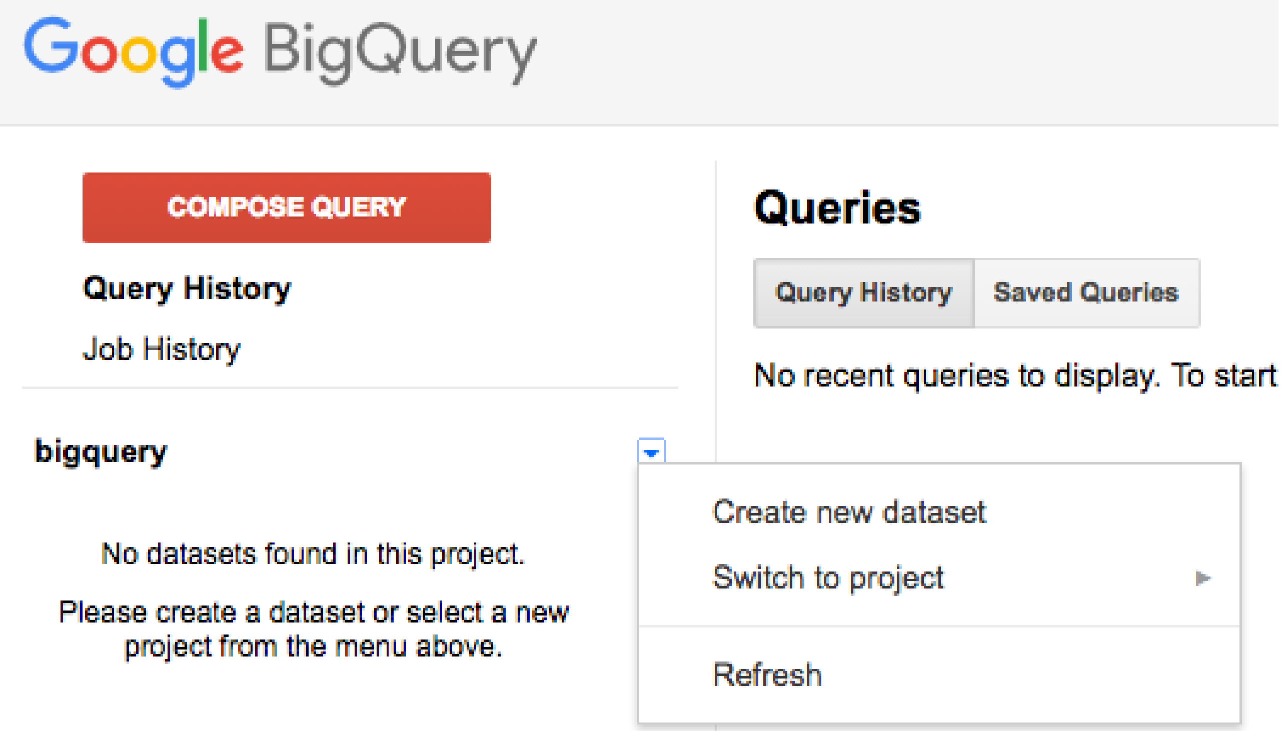
□ Name	Size
□ seq0fix.csv	9.18 MB
□ seq1fix.csv	197.75 MB

Click on BigQuery in the Big Data section in the left panel of the Google Developer Console



Once the files are uploaded to Cloud Storage you can create a BigQuery Dataset (mine is called **olympics**) and upload the data into a BigQuery Table (mine is called **summer2012**)

Create **olympics** Dataset

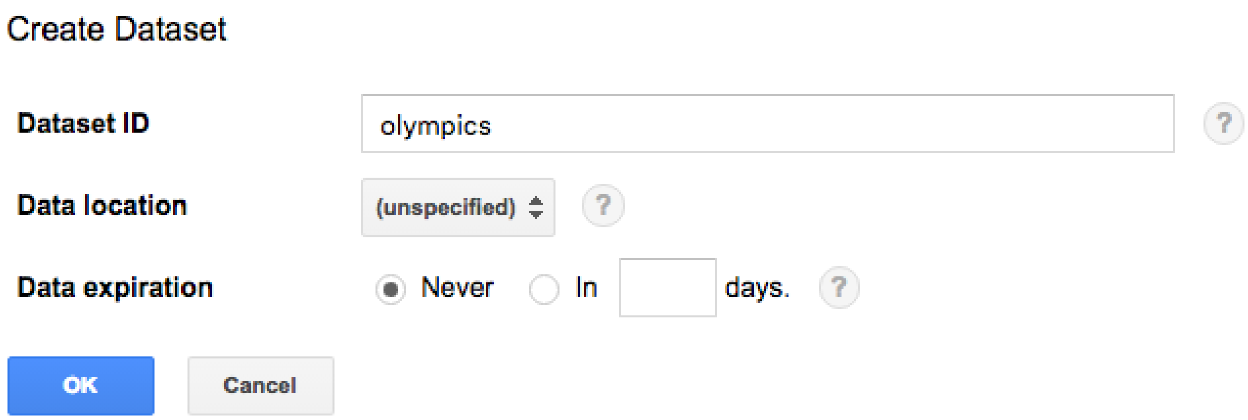


No recent queries to display. To start

No datasets found in this project.

Please create a dataset or select a new project from the menu above.

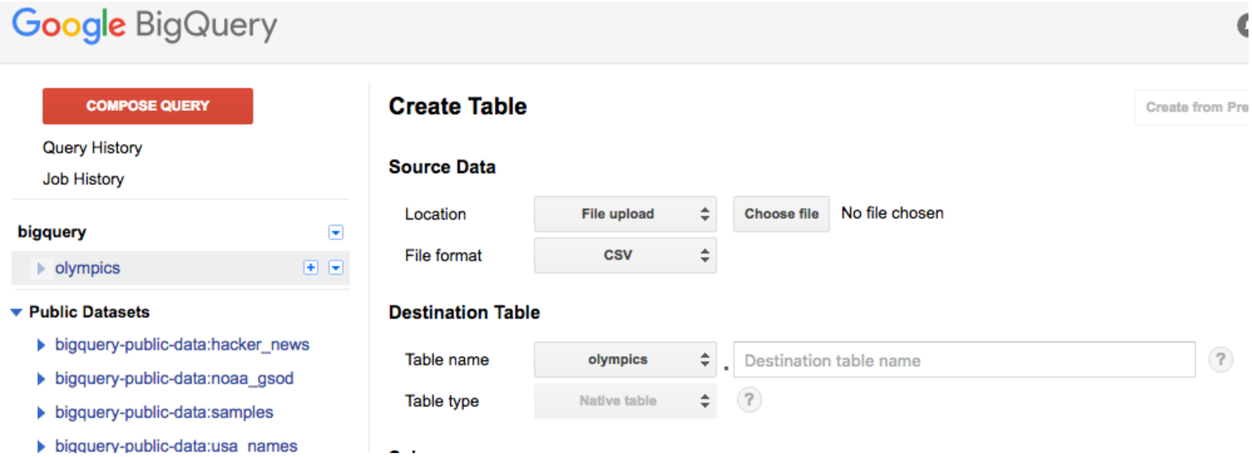
Enter Dataset Name and click the OK button



Create Dataset

Dataset ID	olympics
Data location	(unspecified)
Data expiration	● Never ○ In ___ days.

Create the **summer2012** Table


Google BigQuery
COMPOSE QUERY
Query History
Job History
bigquery
olympics
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
Create from Pre
Choose file
?

Create Table

Source Data

Location: File upload — No file chosen

File format: CSV

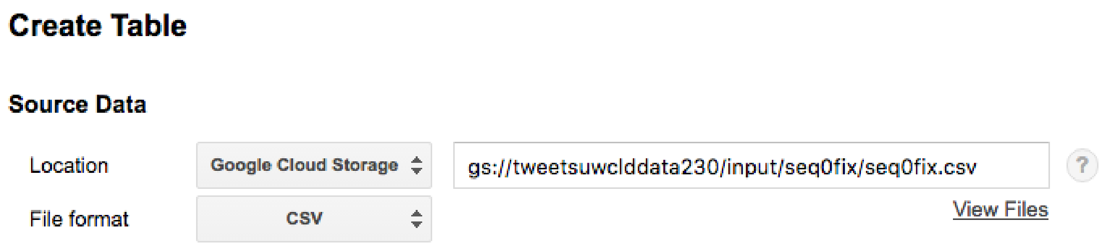
Destination Table

Table name: olympics . Destination table name

Table type: Native table

Pick Google Cloud Storage as the location and provide a path the file. Mine is gs://tweetsuwclddata230/input/seq0fix/seq0fix.csv

gs://<bucket>/<folderIfYouHaveOne>/seq0fix.csv)



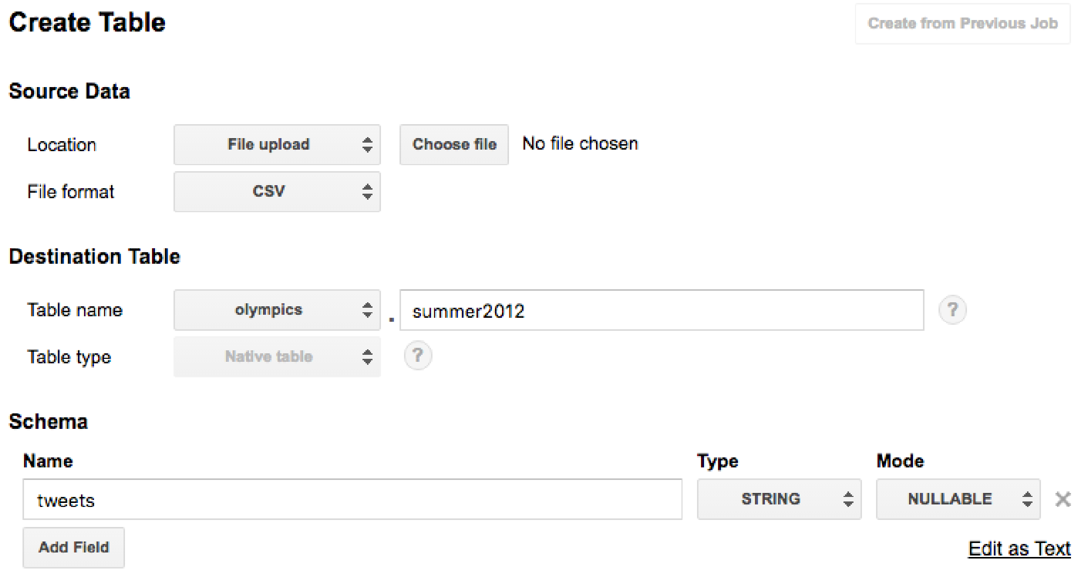
Create Table

Source Data

Location: Google Cloud Storage — gs://tweetsuwclddata230/input/seq0fix/seq0fix.csv

File format: CSV

Enter Table name as **summer2012** and Schema Name as **tweets** and then click Edit as Text



Create Table

Source Data

Location: File upload — No file chosen

File format: CSV

Destination Table

Table name: olympics . summer2012

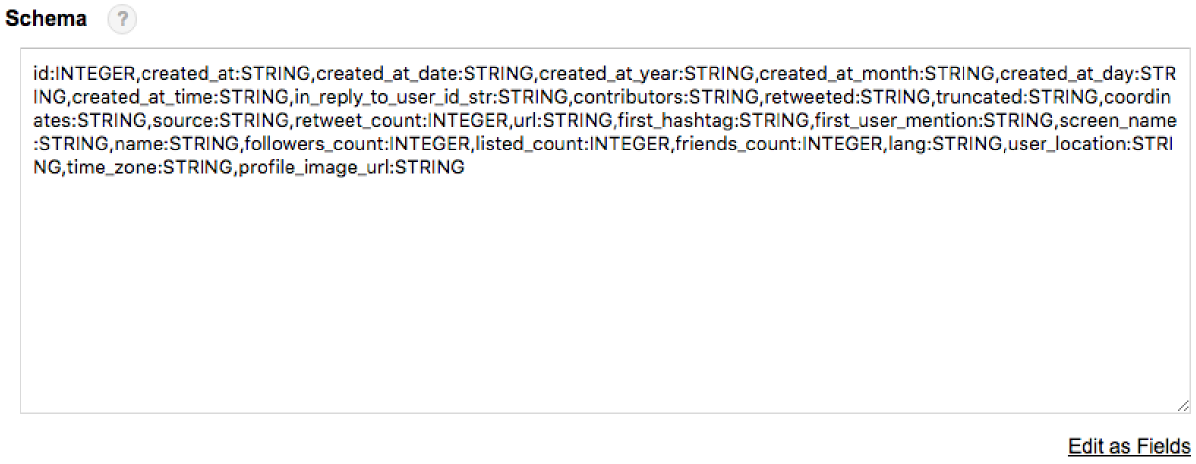
Table type: Native table

Schema

Name	Type	Mode
tweets	STRING	NULLABLE

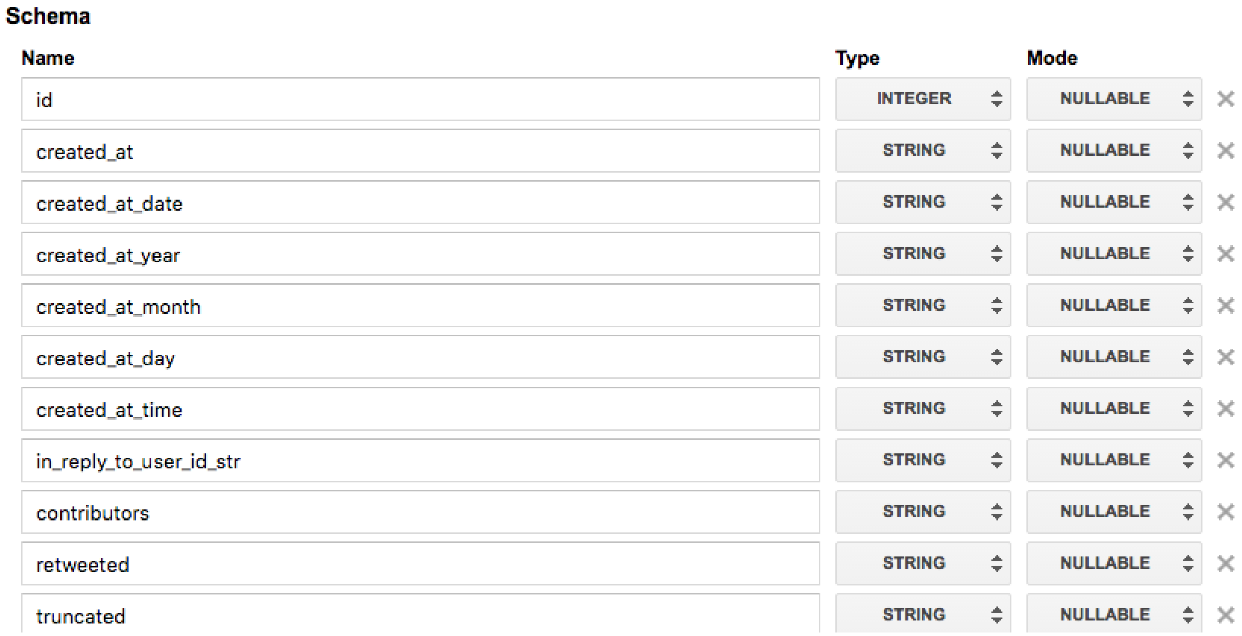
Next it asks for the Schema which is as follows:

id:INTEGER,created_at:STRING,created_at_date:STRING,created_at_year:STRING,created_at_month:STRING,created_at_day:STRING,created_at_time:STRING,in_reply_to_user_id_str:STRING,contributors:STRING,retweeted:STRING,truncated:STRING,coordinates:STRING,source:STRING,retweet_count:INTEGER,url:STRING,first_hashtag:STRING,first_user_mention:STRING,screen_name:STRING,name:STRING,followers_count:INTEGER,listed_count:INTEGER,friends_count:INTEGER,lang:STRING,user_location:STRING,time_zone:STRING,profile_image_url:STRING



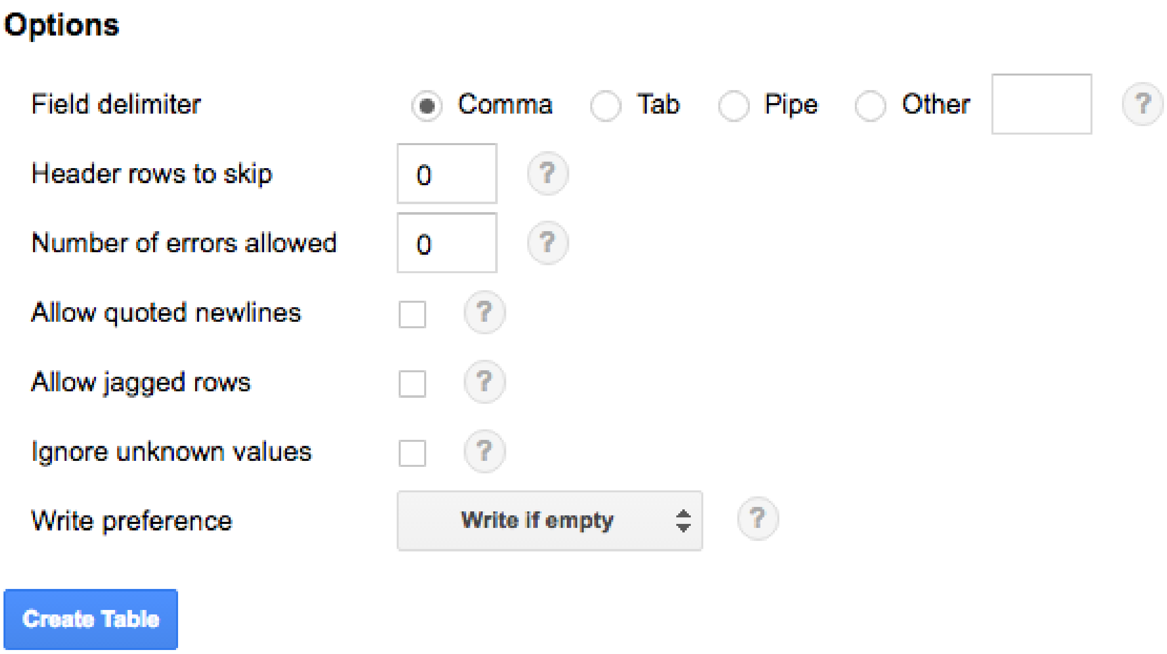
id:INTEGER,created_at:STRING,created_at_date:STRING,created_at_year:STRING,created_at_month:STRING,created_at_day:STRING,created_at_time:STRING,in_reply_to_user_id_str:STRING,contributors:STRING,retweeted:STRING,truncated:STRING,coordinates:STRING,source:STRING,retweet_count:INTEGER,url:STRING,first_hashtag:STRING,first_user_mention:STRING,screen_name:STRING,name:STRING,followers_count:INTEGER,listed_count:INTEGER,friends_count:INTEGER,lang:STRING,user_location:STRING,time_zone:STRING,profile_image_url:STRING

Click on Edit as fields to see what that looks like



Name	Type	Mode
id	INTEGER	NULLABLE
created_at	STRING	NULLABLE
created_at_date	STRING	NULLABLE
created_at_year	STRING	NULLABLE
created_at_month	STRING	NULLABLE
created_at_day	STRING	NULLABLE
created_at_time	STRING	NULLABLE
in_reply_to_user_id_str	STRING	NULLABLE
contributors	STRING	NULLABLE
retweeted	STRING	NULLABLE
truncated	STRING	NULLABLE

Check your Options and click Create Table



Options

Field delimiter ● Comma ○ Tab ○ Pipe ○ Other

Header rows to skip 0

Number of errors allowed 0

Allow quoted newlines □

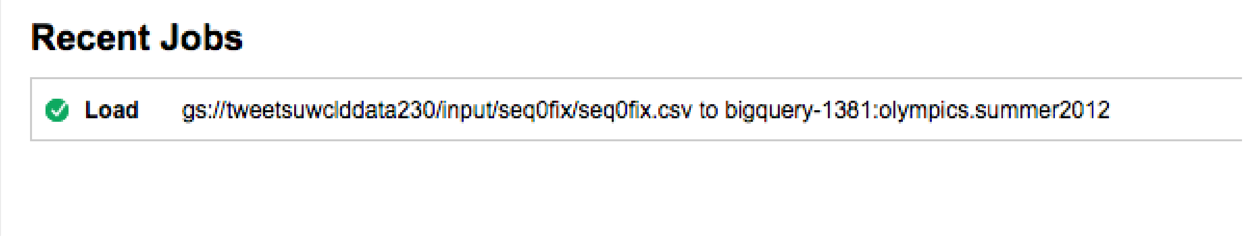
Allow jagged rows □

Ignore unknown values □

Write preference Write if empty

This is a CSV file so choose Comma as the Field delimiter and enter the Number of errors allowed. I choose 100 usually.

You will see that the job is running the and table is loading



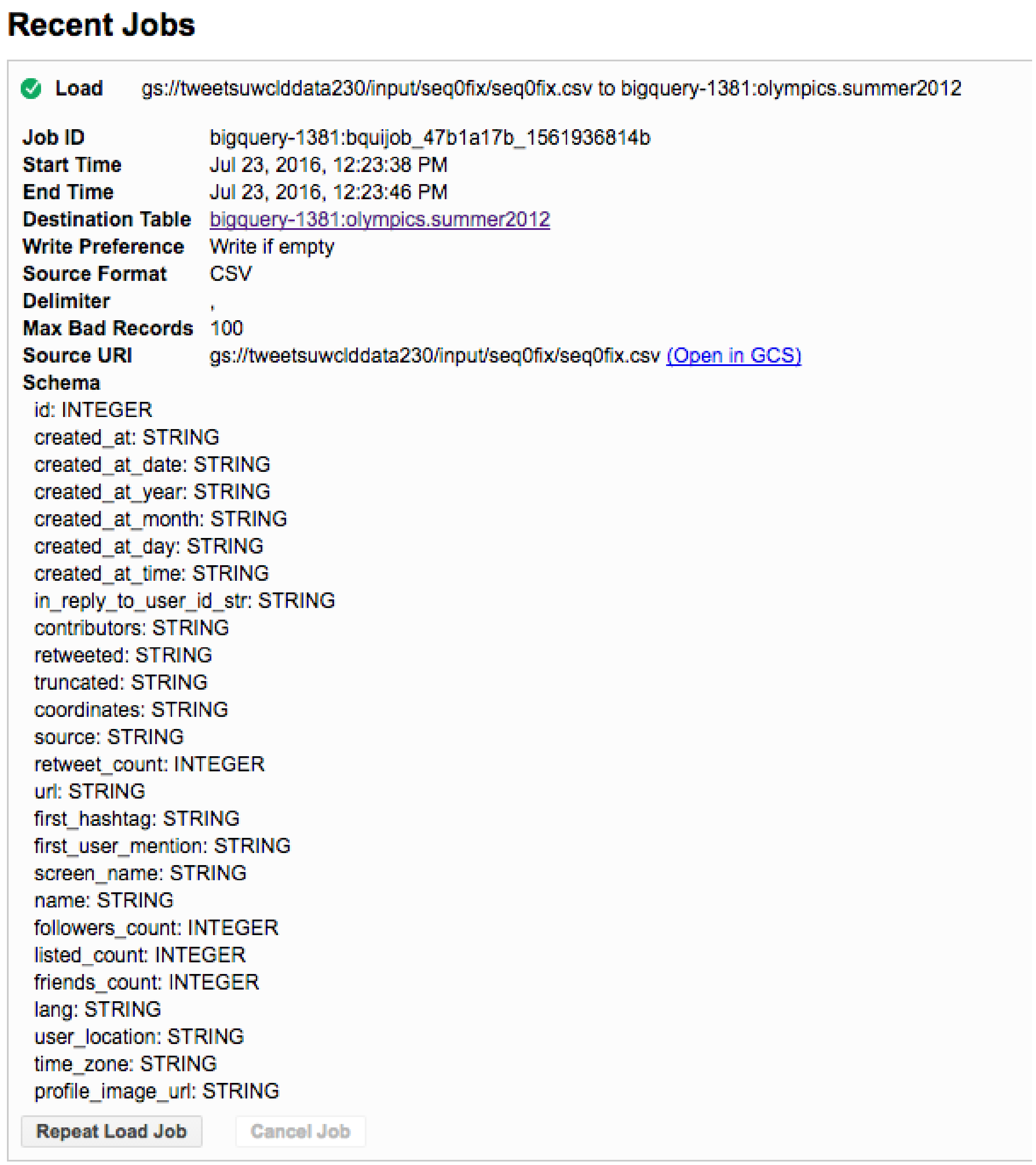
Recent Jobs

✓ Load gs://tweetsuwclddata230/input/seq0fix/seq0fix.csv to bigquery-1381:olympics.summer2012

Once the load completes you will be able to review the Job History. If you have errors or exceed the error threshold it will provide guidance on how to solve the issue.

Click on the Repeat load job and load the second file in the same way as you did above.

Remember to change the file to seq1fix.csv



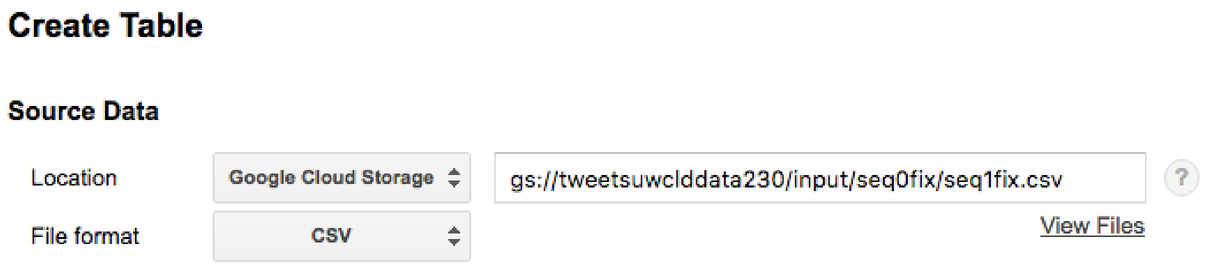
Recent Jobs

✅ **Load** gs://tweetsuwclddata230/input/seq0fix/seq0fix.csv to bigquery-1381:olympics.summer2012

Job ID	bigquery-1381:bquijob_47b1a17b_1561936814b
Start Time	Jul 23, 2016, 12:23:38 PM
End Time	Jul 23, 2016, 12:23:46 PM
Destination Table	bigquery-1381:olympics.summer2012
Write Preference	Write if empty
Source Format	CSV
Delimiter	,
Max Bad Records	100
Source URI	gs://tweetsuwclddata230/input/seq0fix/seq0fix.csv (Open in GCS)

Schema

id: INTEGER
created_at: STRING
created_at_date: STRING
created_at_year: STRING
created_at_month: STRING
created_at_day: STRING
created_at_time: STRING
in_reply_to_user_id_str: STRING
contributors: STRING
retweeted: STRING
truncated: STRING
coordinates: STRING
source: STRING
retweet_count: INTEGER
url: STRING
first_hashtag: STRING
first_user_mention: STRING
screen_name: STRING
name: STRING
followers_count: INTEGER
listed_count: INTEGER
friends_count: INTEGER
lang: STRING
user_location: STRING
time_zone: STRING
profile_image_url: STRING

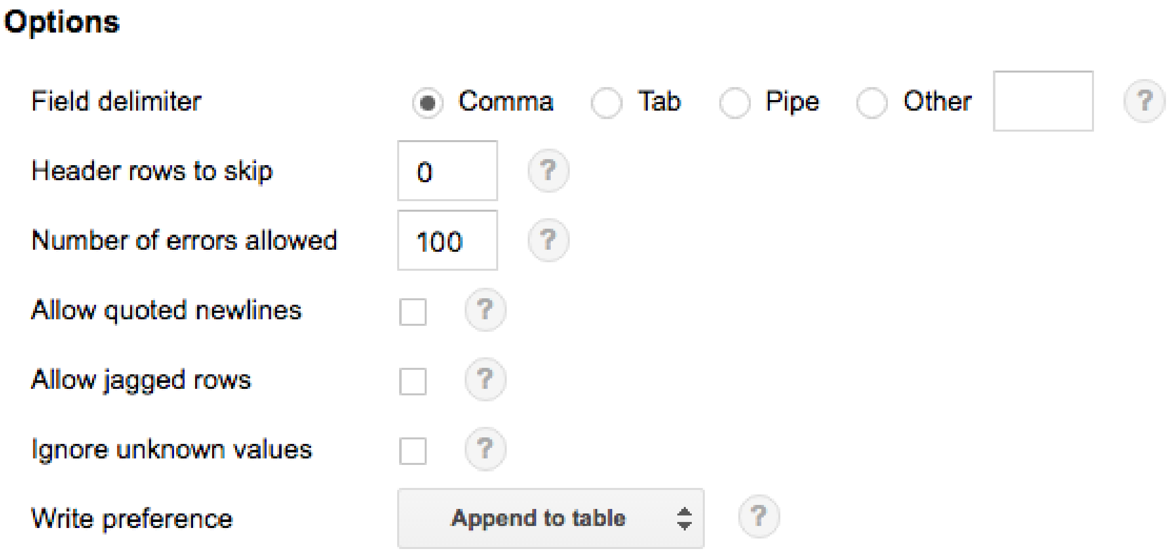


Create Table

Source Data

Location	Google Cloud Storage	gs://tweetsuwclddata230/input/seq0fix/seq1fix.csv
File format	CSV	

Also make sure you change the Write preference in Options to **Append to table**



Options

Field delimiter ● Comma ○ Tab ○ Pipe ○ Other

Header rows to skip 0

Number of errors allowed 100

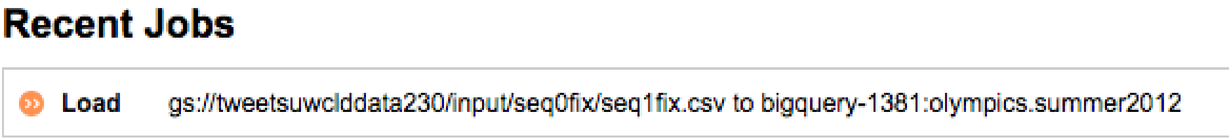
Allow quoted newlines □

Allow jagged rows □

Ignore unknown values □

Write preference Append to table

Click **Create Table**



Recent Jobs

Load gs://tweetsuwclddata230/input/seq0fix/seq1fix.csv to bigquery-1381:olympics.summer2012

If you click on the table summer2012 you can see the Schema

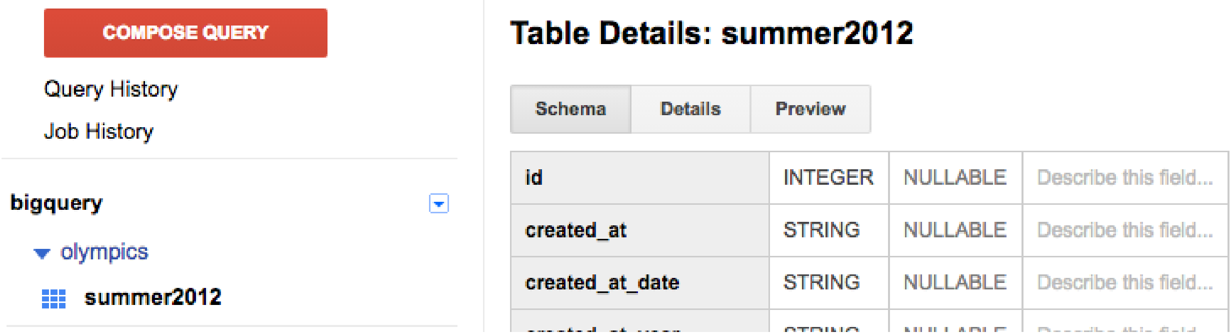


Table Details: summer2012

id	INTEGER	NULLABLE	Describe this field...
created_at	STRING	NULLABLE	Describe this field...
created_at_date	STRING	NULLABLE	Describe this field...
created_at_year	STRING	NULLABLE	Describe this field...

As well as the Details

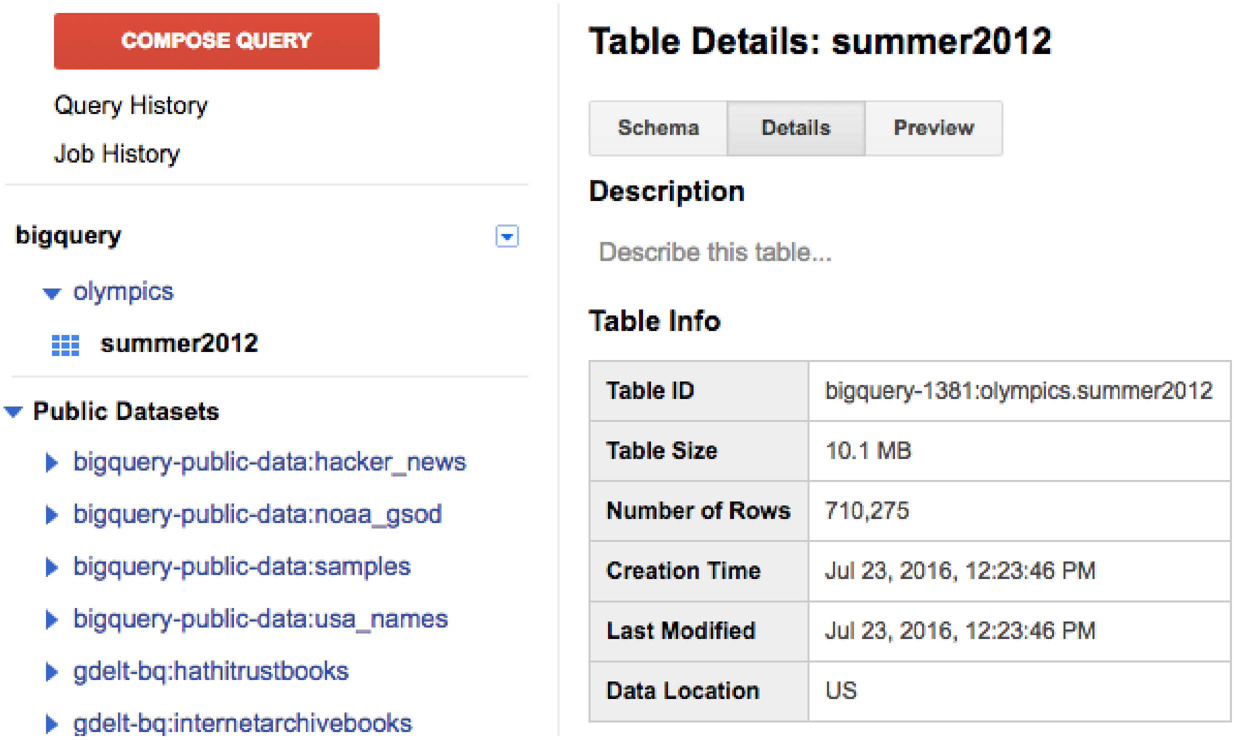

COMPOSE QUERY
Query History
Job History
bigquery
olympics
summer2012
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
Schema
Details
Preview

Table Details: summer2012

Description

Describe this table...

Table Info

Table ID	bigquery-1381:olympics.summer2012
Table Size	10.1 MB
Number of Rows	710,275
Creation Time	Jul 23, 2016, 12:23:46 PM
Last Modified	Jul 23, 2016, 12:23:46 PM
Data Location	US

and a Preview of the table

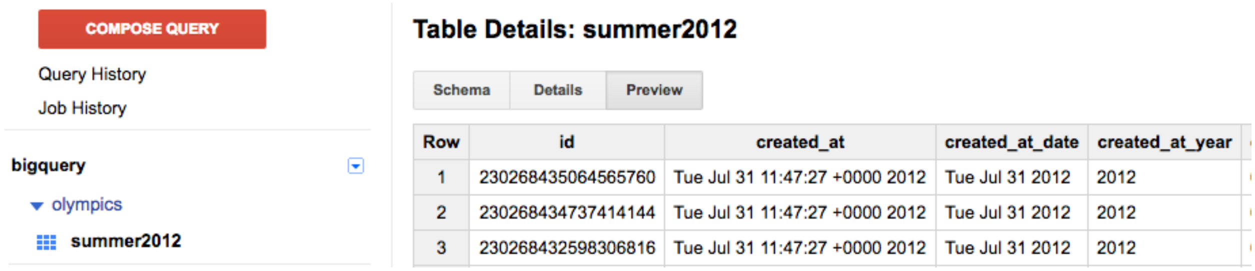


Table Details: summer2012

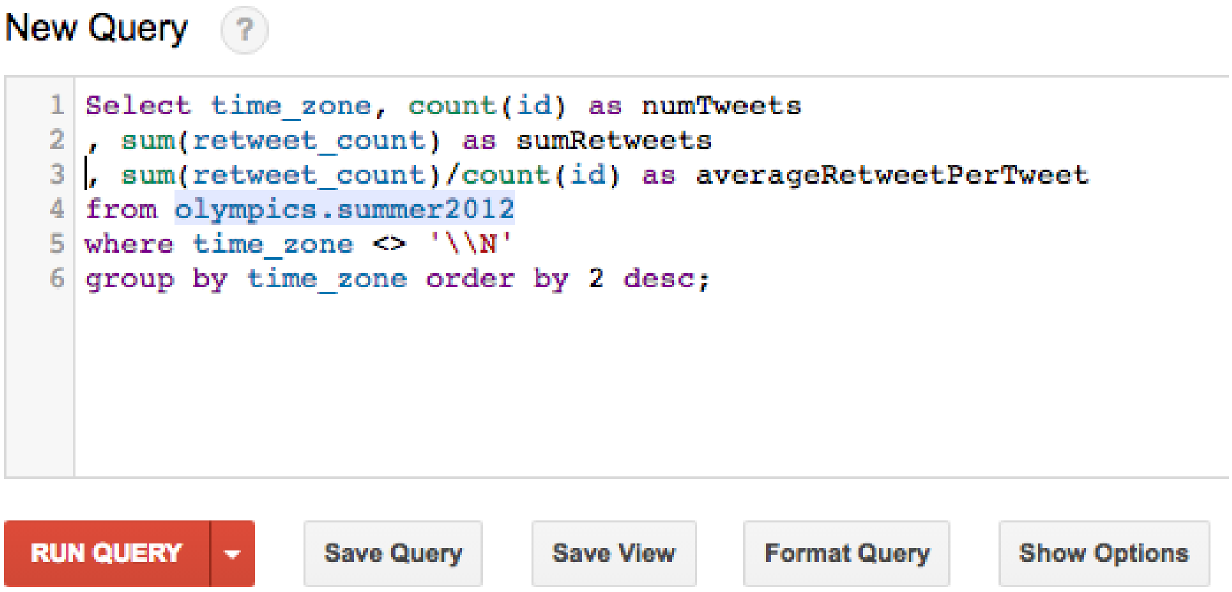
Row	id	created_at	created_at_date	created_at_year
1	230268435064565760	Tue Jul 31 11:47:27 +0000 2012	Tue Jul 31 2012	2012
2	230268434737414144	Tue Jul 31 11:47:27 +0000 2012	Tue Jul 31 2012	2012
3	230268432598306816	Tue Jul 31 11:47:27 +0000 2012	Tue Jul 31 2012	2012

To run a query, click on the **COMPOSE QUERY** button

Enter a query like:

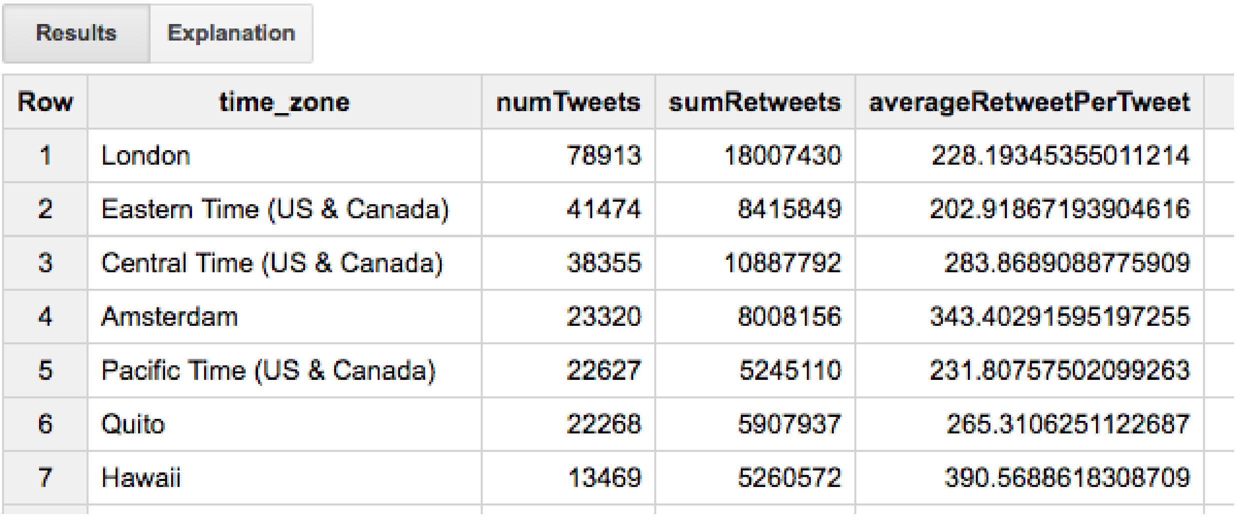
```
Select time_zone, count(id) as numTweets, sum(retweet_count) as sumRetweets,
sum(retweet_count)/count(id) as averageRetweetPerTweet
from olympics.summer2012
where time_zone <> '\\N'
group by time_zone order by 2 desc;
```

Then click the Run Query button



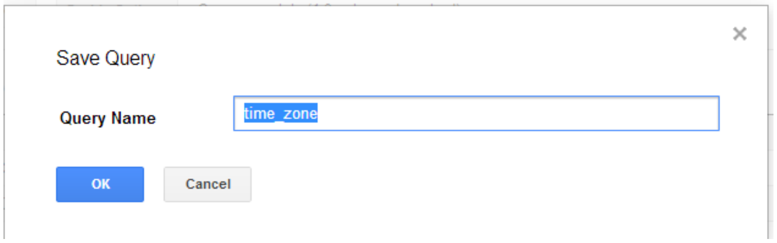
```
Select time_zone, count(id) as numTweets
, sum(retweet_count) as sumRetweets
, sum(retweet_count)/count(id) as averageRetweetPerTweet
from olympics.summer2012
where time_zone <> '\\N'
group by time_zone order by 2 desc;
```

Which will return records.



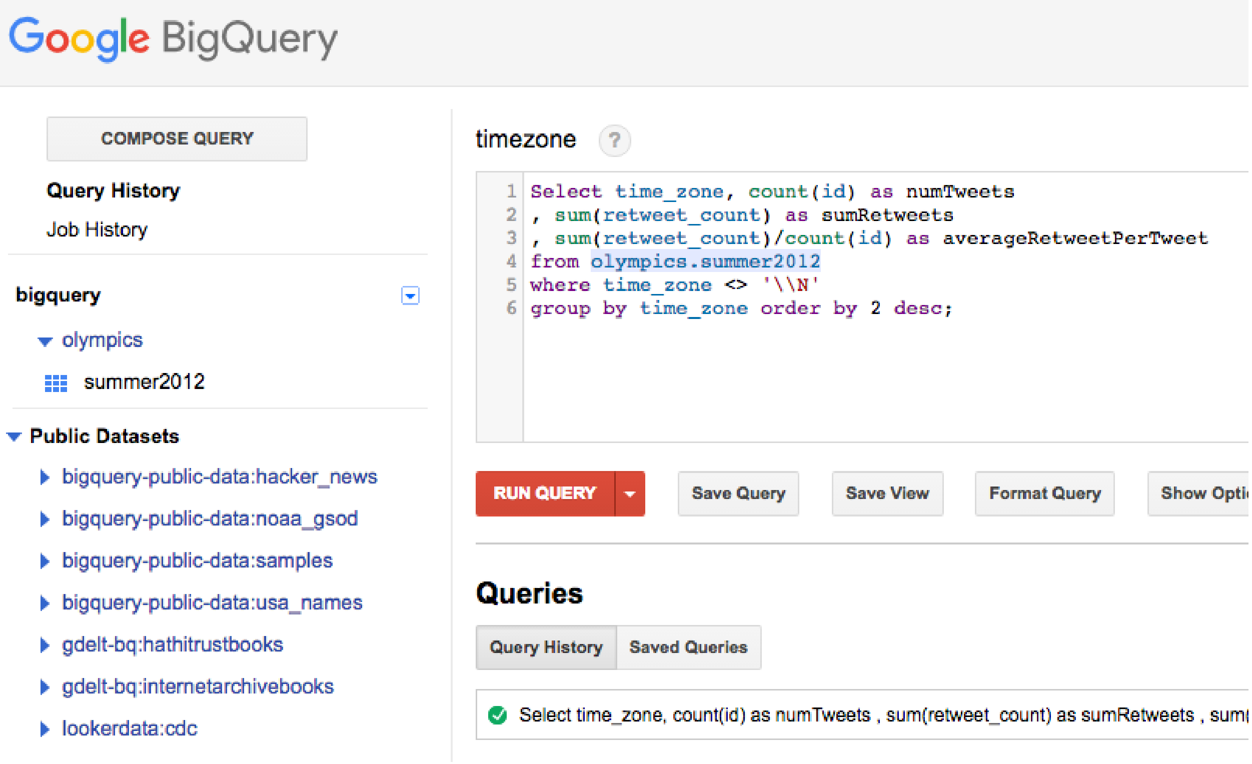
Row	time_zone	numTweets	sumRetweets	averageRetweetPerTweet
1	London	78913	18007430	228.19345355011214
2	Eastern Time (US & Canada)	41474	8415849	202.91867193904616
3	Central Time (US & Canada)	38355	10887792	283.8689088775909
4	Amsterdam	23320	8008156	343.40291595197255
5	Pacific Time (US & Canada)	22627	5245110	231.80757502099263
6	Quito	22268	5907937	265.3106251122687
7	Hawaii	13469	5260572	390.5688618308709

If you like the Query you can save it by clicking the Save Query button, name it and click OK



Query Name: time_zone

You can retrieve the saved query and other recent queries by clicking on Query History in the Left navigation panel

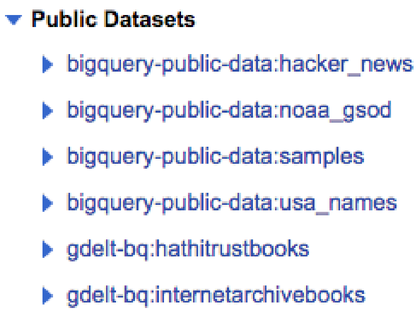


```
Select time_zone, count(id) as numTweets
, sum(retweet_count) as sumRetweets
, sum(retweet_count)/count(id) as averageRetweetPerTweet
from olympics.summer2012
where time_zone <> '\\N'
group by time_zone order by 2 desc;
```

Select time_zone, count(id) as numTweets , sum(retweet_count) as sumRetweets , sum(

Have fun writing some queries and look at homework exercises over the weekend.

Also feel free to try on the Public Datasets


Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks