Activity Overview

Recently, you've been thinking about identifying good data sources that would be useful for analysis. You also spent some time in a previous activity exploring a public dataset in BigQuery and writing some basic SQL queries. In addition to using public data in BigQuery, your future data career will involve importing data from other sources. In this activity, you will create a custom table and dataset that you'll load into a new table and query.

By the time you complete this activity, you will be able to load your own data into BigQuery for analysis. This will enable you to import your own data sources into BigQuery, which is a skill that will enable you to more effectively analyze data from different sources.

Step 1: Access the data source

To get started, download the baby names data zip file. This file contains about 7 MB of data about popular baby names from the U.S. Social Security Administration website.

Select the link to the baby names data zip file to download it.

Link to baby names data: names.zip

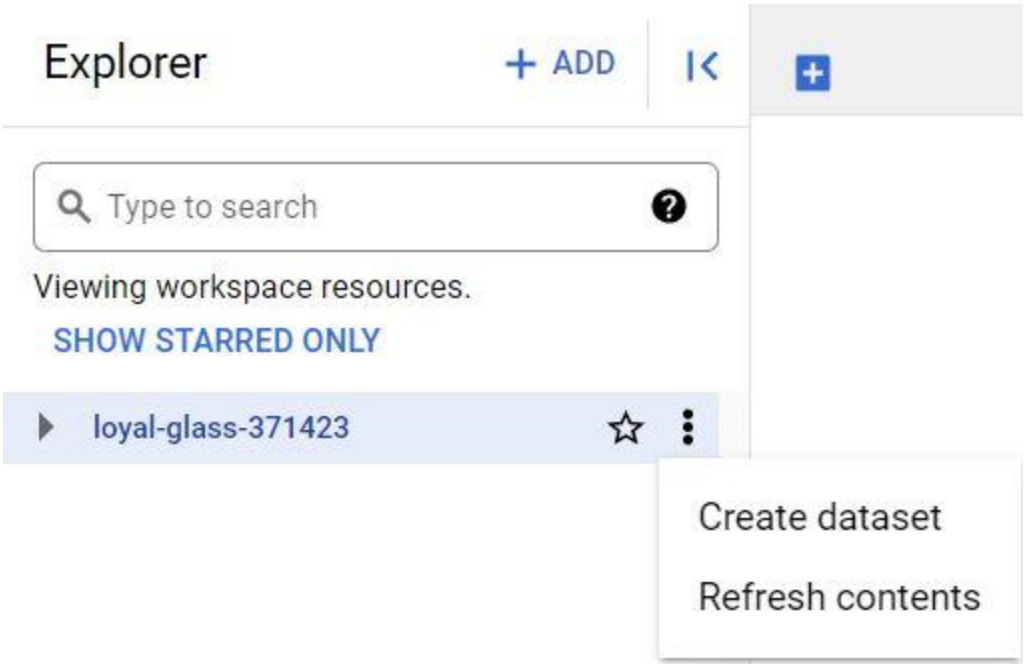
Step 2: Unzip the file

Unzip the file you downloaded onto your computer to access it on BigQuery. Once you have unzipped the file, find a .pdf file titled NationalReadMe that contains more information about the dataset. This dataset tracks the popularity of baby names for each year; you can find text files labeled by the year they contain. Open yob2014.txt to preview the data. You will notice that it's a .csv file with three columns. Remember where you saved this folder so you can reference it later.

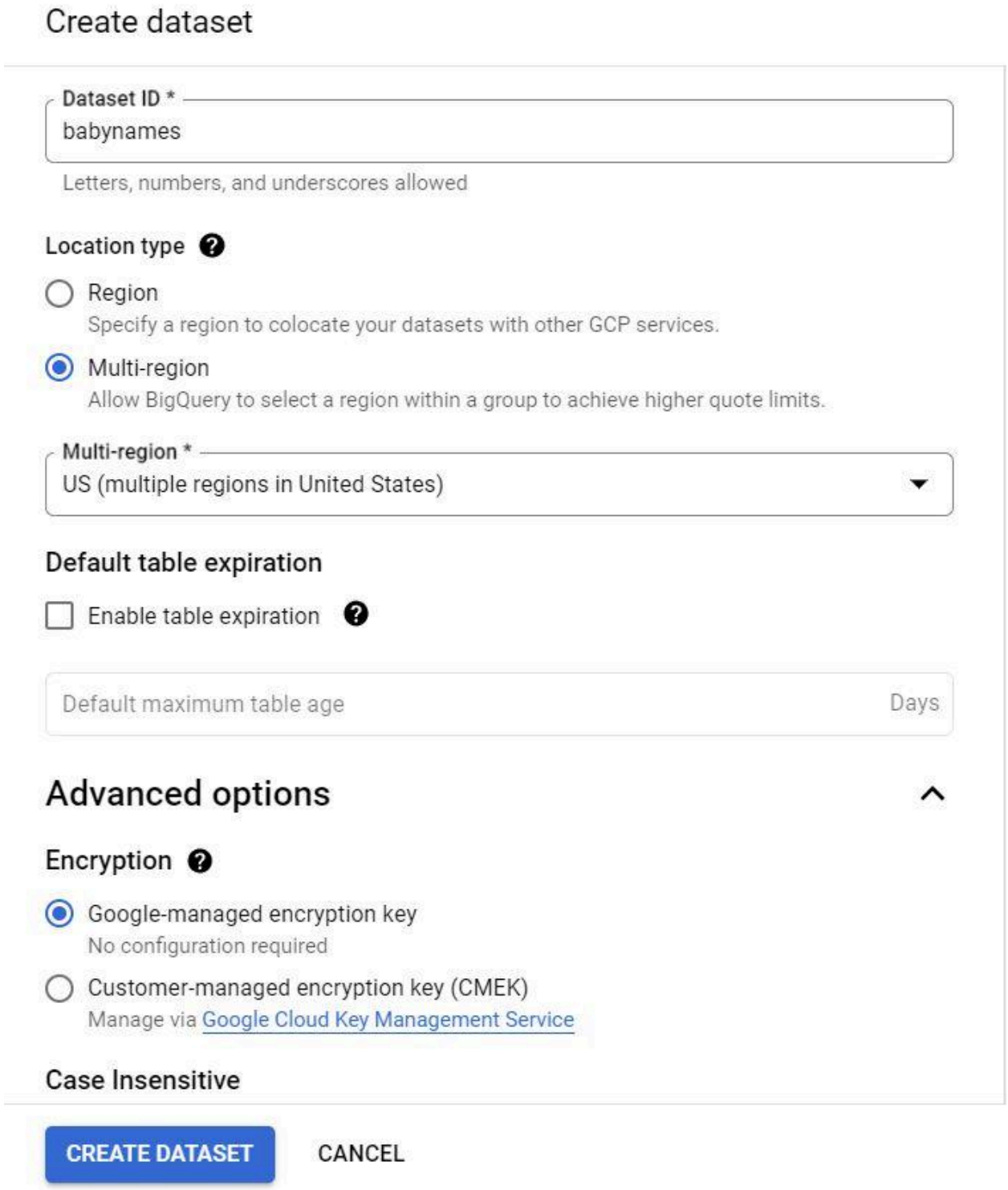
Step 3: Create a dataset

Before uploading your .txt file and creating a table to query, you will need to create a dataset to upload your data into and store your tables.

1. From the BigQuery console, go to the Explorer pane in your workspace and select the three dots next to your project to open a menu. From here, select Create dataset. Note that unless you have already specified your own project name, a unique name is assigned to your project by BigQuery, typically in the format of two words and a number, separated by hyphens (e.g. loyal-glass-371423 in the image below).



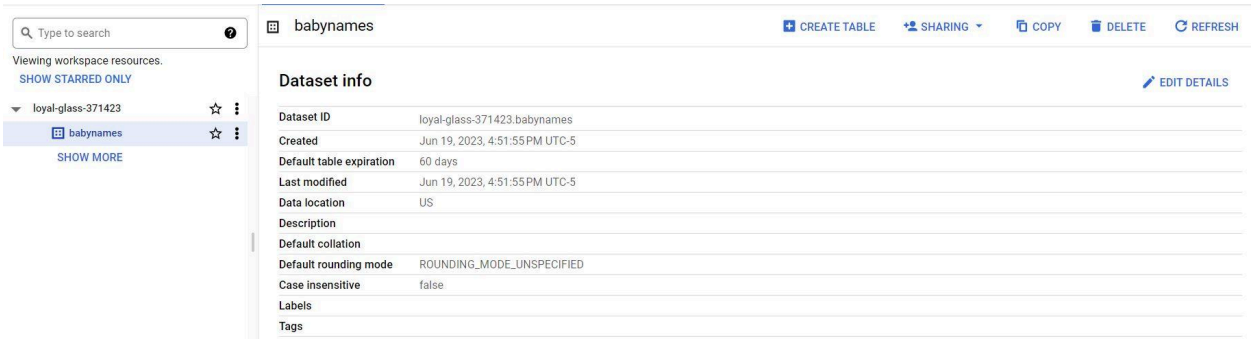
2. This will open the Create dataset menu. This is where you will fill out information about the dataset. Input the Dataset ID as babynames and set the Data location to Multi-region (US). Once you have finished filling out this information, select the blue CREATE DATASET button at the bottom of the menu.



Step 4: Create a table

Now that you have a custom database stored in your project space, this is where you will add your table.

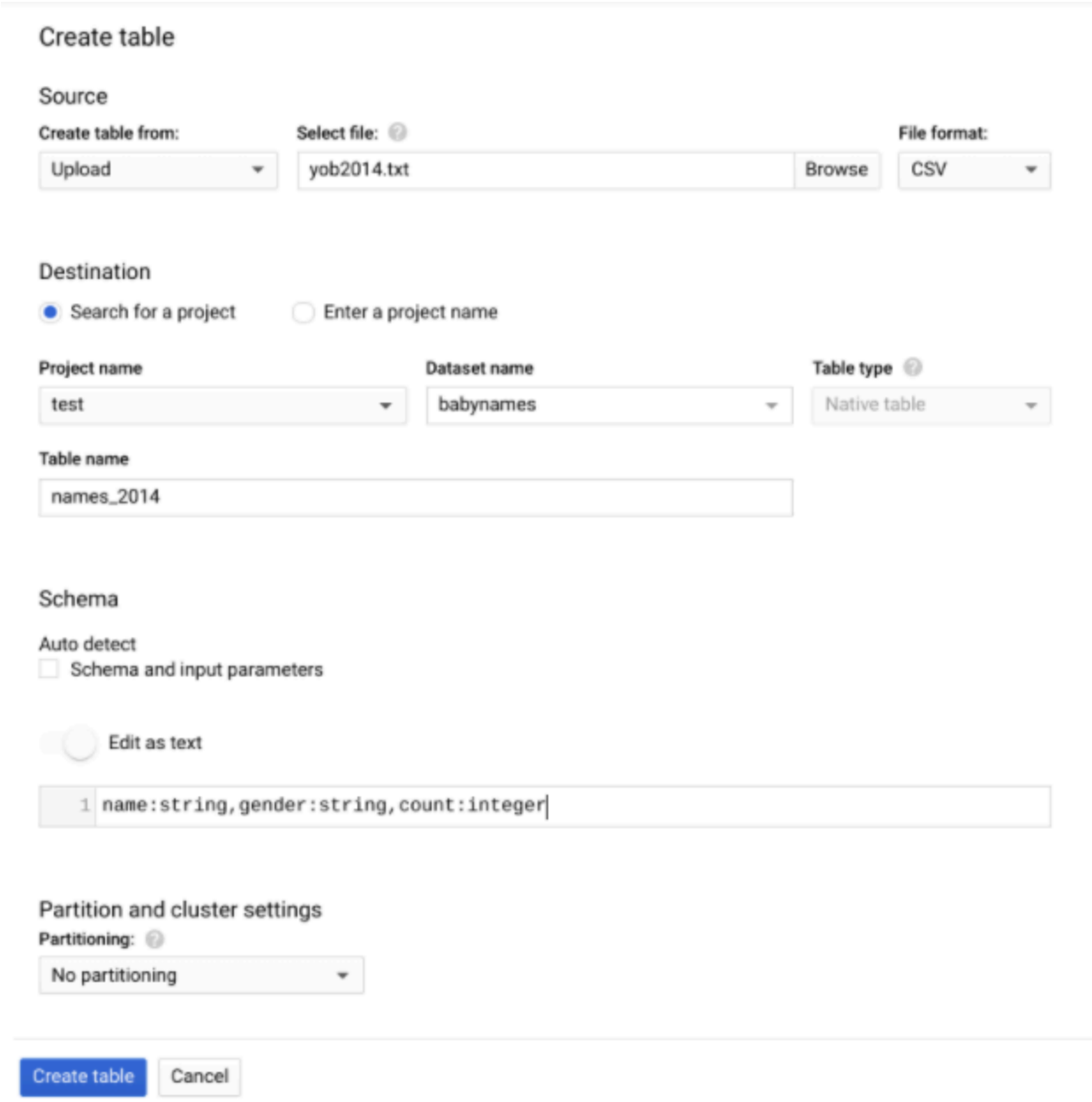
1. Select the newly created babynames database. Check the tabs in your Dataset info window and select the first blue + CREATE TABLE button. This will open another menu in your console.



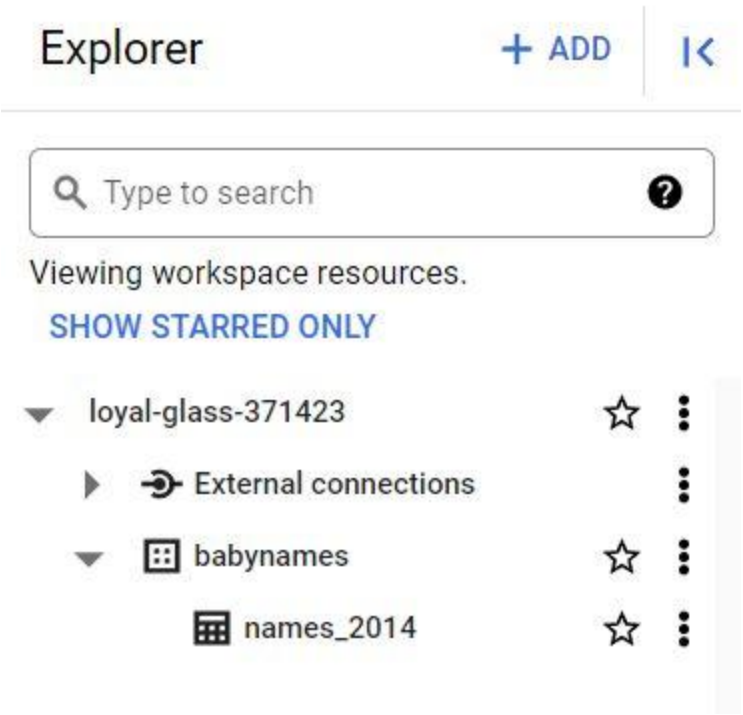
2. In the Source section, select the Upload option in Create table from. Then select Browse to open your files. Find and open the yob2014.txt file. Set the file format to .csv. In the Destination section, in the Table data box, name your table as names_2014. For Schema, select Edit as text and input the following code:

```
1   name:string,
2   gender:string,
3   count:integer
```

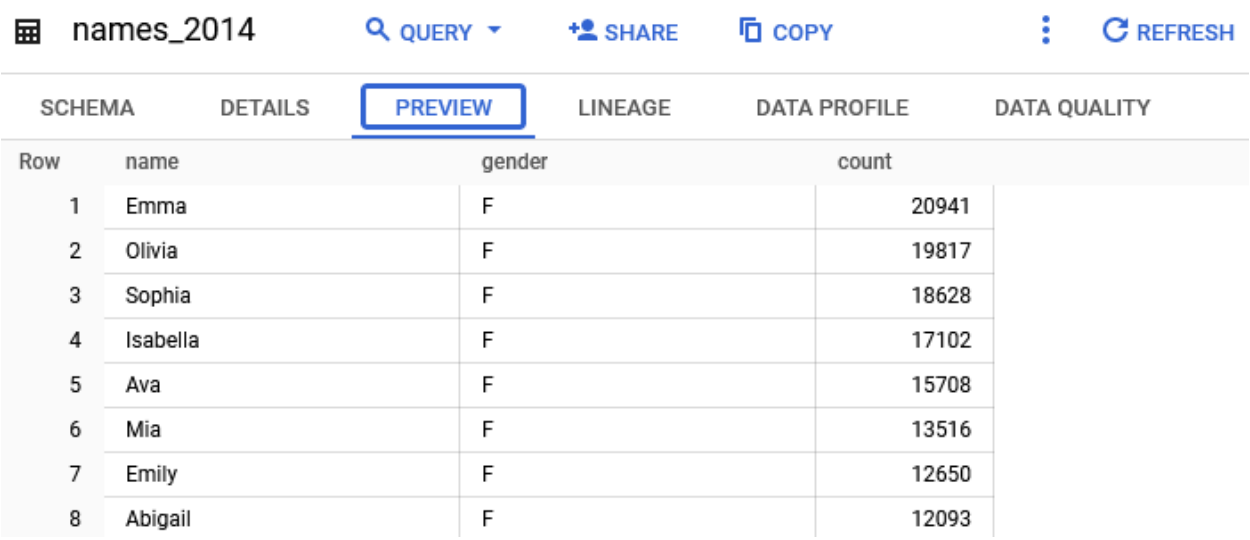
3. This will establish the data types of the three columns in the table. Leave the other parameters as they are, and select Create table.



Once you have created your table titled names_2014, it will appear in your explorer pane under the database **`babynames`** that you created earlier.



4. Select the table to open it in your workspace. Here, you can check the table schema. Then, go to the Preview tab to explore your data. The table should have three columns: **`name`**, **`gender`**, and **`count`**.



Row	name	gender	count
1	Emma	F	20941
2	Olivia	F	19817
3	Sophia	F	18628
4	Isabella	F	17102
5	Ava	F	15708
6	Mia	F	13516
7	Emily	F	12650
8	Abigail	F	12093

Step 5: Query your custom table

Now that your table is set up, you're ready to start writing queries and answering questions about this data. For example, let's say you were interested in the top five baby names for boys in the United States in 2014.

Select COMPOSE NEW QUERY to start a new query for this table. Then, enter this code:

```sql
SELECT
  name,
  count
FROM
  your-project.babynames.names_2014
WHERE
  gender = 'M'
ORDER BY
  count DESC
LIMIT
  5
```

NOTE: Making sure that your **`FROM`** statement is correct is key to making this query run! The database needs the query to tell it the location of the table you just uploaded so that it can fetch the data. It's like giving the query a map to your table. That map will include your unique BigQuery project name (replace **`your-project`** in the code above with your unique project name), the database name (**`babynames`**), and the table name (**`names_2014`**). The location names for each of these elements are separated by periods. The final result will be something like this:
`loyal-glass-371423.babynames.names_2014`
Note that **`loyal-glass-371423`** is just an example of a project name. You must use your project's name in your **`FROM`** statement.
This query selects the **`name`** and **`count`** columns from the **`names_2014`** table. Using the **`WHERE`** clause, you are filtering for a specific gender for your results. Then, you're sorting how you want your results to appear with **`ORDER BY`**. Because you are ordering by the **`count`** in descending order, you will get names and the corresponding count from largest to smallest. Finally, **`LIMIT`** tells SQL to only return the top five most popular names and their counts.
Once you have input this in your console, select RUN to get your query results.

Reflection

After running the query on your new table, what was the third-most popular baby name for boys in 2014?

- Jacob
- William
- Noah
- Mason

Question 2

In this activity, you explored public data in BigQuery and used it to create a custom table. In the text box below, write 2-3 sentences (40-60 words) in response to each of the following questions:

- Why is being able to use data from different sources useful as a data analyst?
- How can you use BigQuery custom tables and datasets in your future analysis projects?

Congratulations! In this activity, you created a new dataset within your project, uploaded a .csv file to create a new table, and ran a SQL query. A good response would include that being able to evaluate and use different data sources allows you access more data.

As a data analyst, being able to evaluate data sources and use the appropriate tool to analyze them is important. For instance, you were able to use SQL to analyze a dataset that was previously stored on your computer as a .csv file.