## **Hands-On Activity: Processing time with SQL**

**TOTAL POINTS 2**

1.

Question 1



## Activity overview



In previous activities, you learned about and practiced SQL. In this activity, you’ll work with SQL queries of different sizes.

By the time you complete this activity, you’ll be familiar with the different sizes used to measure data storage. This will help you understand how data size affects the amount of time queries take to run and how valuable tools like SQL can be to data analysts.

## Understand how data is measured



Data is measured by the number of bits it takes to represent it. All information in a computer can be represented as a binary number consisting solely of 0’s and 1’s. Each 0 or 1 in a number is a bit. A bit is the smallest unit of storage in computers. Since computers work in binary (Base 2), this means that all the important numbers that differentiate between different data sizes will be powers of 2.

A byte is a collection of 8 bits. Take a moment to examine the table below to get a feel for the difference between data measurements and their relative sizes to one another.

| **Unit** | **Equivalent to** | **Abbreviation** | **Real-World Example** |
| --- | --- | --- | --- |
| Byte | 8 bits | B | 1 character in a string |
| Kilobyte | 1024 bytes | KB | A page of text (~4 kilobytes) |
| Megabyte | 1024 Kilobytes | MB | 1 song in MP3 format (~2-3 megabytes) |
| Gigabyte | 1024 Megabytes | GB | ~300 songs in MP3 format |
| Terabyte | 1024 Gigabytes | TB | ~500 hours of HD video |
| Petabyte | 1024 Terabytes | PB | 10 billion Facebook photos |
| Exabyte | 1024 Petabytes | EB | ~500 million hours of HD video |
| Zettabyte | 1024 Exabytes | ZB | All the data on the internet in 2019 (~4.5 ZB) |

The amount of data in the world is exploding and growing at an incredible pace every year. This growth is largely the result of the over 4.6 billion people around the world connected to the Internet. Now that smartphones and other Internet-connected devices have become common, they generate a staggering amount of new data. Many experts believe that the size of all the data on the Internet will swell to 175 ZB by the end of 2025!

The size of the dataset you’re working with usually determines which tool, spreadsheets or SQL, is best suited for the task. Spreadsheets often start to have performance issues as dataset sizes increase beyond a few megabytes. SQL databases are much better at working with larger datasets that have billions of rows with sizes measured in gigabytes. The dataset’s size still matters here--larger datasets will take longer for queries to complete, depending on the query’s content and the number of rows SQL has to process to complete the query.

## Query a large dataset



You’ll now discover for yourself how these runtimes change with dataset size by running some queries on a huge dataset—Wikipedia!

1. Log in to [BigQuery Sandbox](https://cloud.google.com/bigquery/docs/sandbox). If you have a free trial version of BigQuery, you can use that instead. On the BigQuery page, click the Go to BigQuery button.

* Note: BigQuery Sandbox frequently updates its user interface. The latest changes may not be reflected in the screenshots presented in this activity, but the principles remain the same. Adapting to changes in software updates is an essential skill for data analysts, and it’s helpful for you to practice troubleshooting. You can also reach out to your community of learners on the discussion forum for help.

2. If you have never created a BigQuery project before, click CREATE PROJECT on the right side of the screen. If you have created a project before, you can use an existing one or create a new one by clicking the project dropdown in the blue header bar and selecting NEW PROJECT.

3. Name your project something that will help you identify it later. You can give it a unique project ID or use an auto-generated one. Don’t worry about selecting an organization if you don’t know what to put.

4. Now, you’ll see the Editor interface. In the middle of the screen is a window where you can type code, and to the left is the Explorer menu where you can search for datasets.

5. Copy and paste the following query into the editor and run it. The formatting is just cosmetic, so don’t worry if it changes when copied over. The query should take 10-15 seconds to run:

10

11

12

13

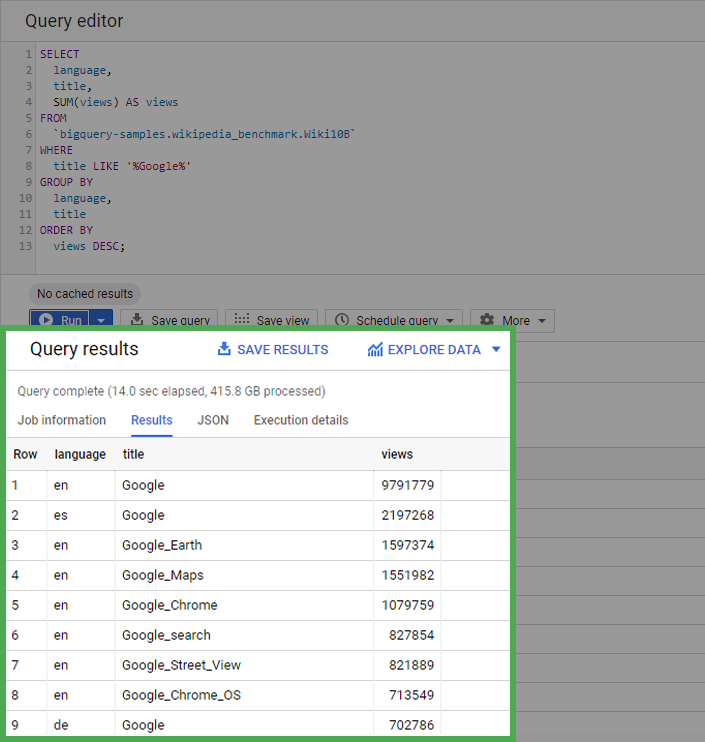
language,

title

ORDER BY

views DESC;

After the query finishes, your screen should appear like this:



This query returns a table that displays the total number of times each Wikipedia page with “Google” in the title has been viewed in each language. Note the information that BigQuery provides on the query you just ran. As you can infer from the dataset’s title in the query, this dataset is a sample consisting of 10 billion rows from the Wikipedia public dataset.

You’ll find that the query processes over 415 gigabytes of data when run—pretty impressive for 15 seconds! Note that if you run the query again, the runtime will be almost instant (as long as you haven’t changed the default caching settings). This is because BigQuery caches the query results to avoid extra work if the query needs to be rerun.

## Confirmation and reflection



In your last query, you processed 415.8 GB of data. How many rows were returned by the query?

**1 / 1 point**

214,710

305,710

225,038

198,768

**Correct**

The last query you ran returns 214,710 rows of data. At the bottom of the data preview, you can see how many rows you returned. Going forward, you can apply this knowledge of data size measurements to better understand how much data you will work with and what tool is best suited to each data analysis project.

2.

Question 2

In this activity, you compared the amount of time it takes to process different sizes of queries in SQL. In the text box below, write 2-3 sentences (40-60 words) in response to each of the following questions:

* How did working with SQL help you query a larger dataset?
* How long do you think it would take a team to query a dataset like this manually?
* How does the ability to query large datasets in reasonable amounts of time affect data analysts?

**1 / 1 point**

With the help of SQL, we can query a large dataset in a very short amount of time. which is not possible with a spreadsheet. It will take years for a team to query a dataset like this manually.

**Correct**

Congratulations on completing this hands-on activity! A good response would include how querying a dataset with billions of items isn’t feasible without tools such as relational databases and SQL.

Performing large queries by hand would take years and years of manual work. The ability to query large datasets is an extremely helpful tool for data analysts. You can gain insights from massive amounts of data to discover trends and opportunities that wouldn’t be possible to find without tools like SQL.