**MODULE 3- DATABASE ESSENTIALS**

**WORK WITH DATABASES**

[**ALL ABOUT DATABASES**](https://www.coursera.org/learn/data-preparation/lecture/szLl7/all-about-databases)

Next step, we're going to learn all about databases. As a refresher, a **database is a collection of data stored in a computer system, but storage is just the beginning**. You'll discover how databases make it possible to find the exact piece of information you need for your analysis. You'll also learn how to **sort data** in order to zoom in on what you need to generate insightful reports and much more. Then we'll go even deeper, and I mean really, really deep. I'm talking about **metadata**. You've probably heard someone say, wow that's so meta. **Usually they're talking about something referencing back to itself or being completely self aware**. For example if a character in a book knows she's in a book, that's meta. If you make a documentary about making documentaries, that's also meta. And here at Google, I constantly analyze how I analyze data. That's definitely meta.

I do that to give my work a quality check to make sure my methods are fair. And to be certain that I'm paying attention to any biases that might affect the outcome. As an analyst, you should do this too. Sometimes we get a little too close to our data. So stepping back and asking ourselves if our processes make sense is key. But let's back up just a bit and define metadata. **Metadata is data about data**. Like I said: deep.

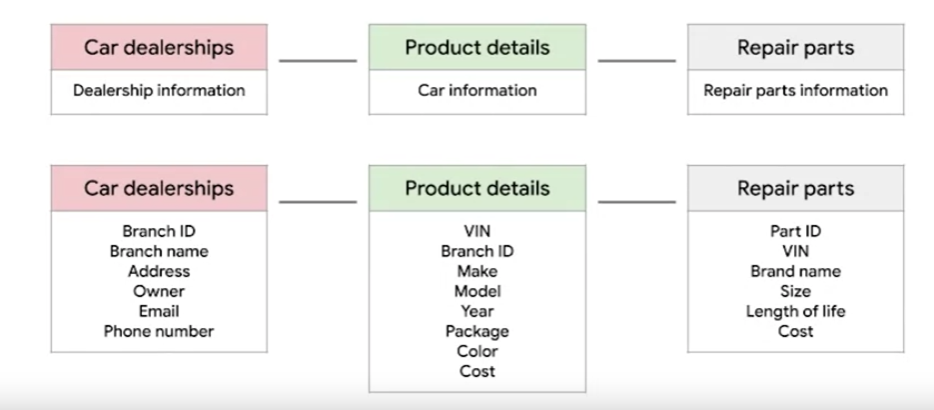
**Metadata is extremely important when working with databases**. Think of it like a reference guide. Without the guide all you have is a bunch of data with no context explaining what it means. Metadata tells you where the data comes from, when and how it was created, and what it's all about.

Up next, you'll learn how to take data from a database or another source and bring it into a spreadsheet. You'll do this either by importing it directly or by using SQL to generate the request. And once you have data in a spreadsheet, the possibilities are endless. Everything we're about to cover is a very important part of the Prepare phase of the data analysis process. It's how data analysts figure out which kind of data is going to be most helpful to them. If you have the right data, you're much more likely to be able to solve your business problems successfully.

[**DATABASE FEATURES AND COMPONENTS**](https://www.coursera.org/learn/data-preparation/lecture/JCWIr/database-features-and-components)

**Databases** are essential tools for data analysts. I use them constantly. Just about all of the data I access is stored within databases. Databases store and organize data, making it much easier for data analysts to manage and access information. They help us get insights faster, make data-driven decisions, and solve problems. You've already heard a bit about what databases are and how they're used by data analysts.

Now let's learn more about database features and components. Here's a simple database structure.



It contains tables with information from a car manufacturer. The top level includes car dealerships, product details, and repair parts. Then if you drill down to the next level by selecting one of those tables, you'll find more specific details about each item. This is called a **relational database**.

A relational database is a **database that contains a series of related tables that can be connected via their relationships**. For two tables to have a relationship, one or more of the same fields must exist inside both tables.

For example, here, **branch ID exists in this table and this one. If a field exists within both tables, we can use it to connect the tables together. The branch ID field is the key to connecting these tables.**

There are two types of keys.

A **primary key** **is an identifier that references a column in which each value is unique**. **You can think of it as a unique identifier for each row in a table**. For our dealership table with information about the different dealership branches, branch ID is the primary key. Similarly, for the product details table about each car, **VIN is our primary key**. As an analyst you may need to create tables. If you do decide to include **a primary key, it should be unique, meaning no two rows can have the same primary key. Also, it cannot be null or blank.**

**There are also foreign keys**. A foreign key **is a field within a table that's a primary key in another table**. In other words, **a foreign key is how one table can be connected to another**. Because our repair parts table contains information about each car part, the primary key is part ID. Each row in our repair parts table represents one unique part. All the other keys in this table, such as the VIN, are the foreign keys that allow the repair parts table to be connected to the other tables. **A table can only have one primary key but it can have multiple foreign keys.**

**As a general summary**, **a primary key is used to ensure data in a specific column is unique**. It uniquely identifies a record in a relational database table.

**Only one primary key is allowed in a table and they cannot contain null or blank values**.

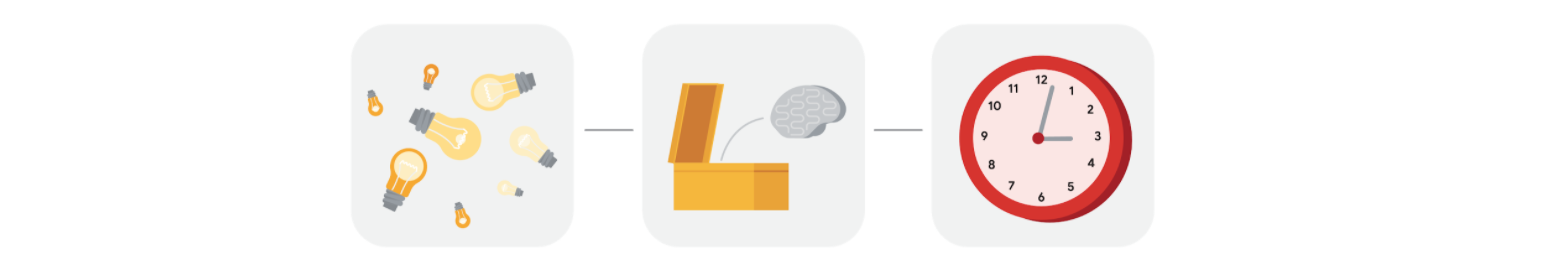
**A foreign key is a column or group of columns in a relational database table that provides a link between the data and two tables**. It refers to the field in a table that's the primary key of another table.

Lastly, **it's important to note that more than one foreign key is allowed to exist in a table.**

Feel free to rewatch this video to be sure you understand primary and foreign keys clearly. And coming up, you'll begin practicing how to access and analyze data from actual databases. That will be a great opportunity to improve your understanding of primary and foreign keys, database organization and how you might use databases in your future analytics career.

[**MAXIMIZE DATABASES IN DATA ANALYTICS**](https://www.coursera.org/learn/data-preparation/supplement/uXqEX/maximize-databases-in-data-analytics)

**Databases enable analysts to manipulate, store, and process data.** This helps them search through data a lot more efficiently to get the best insights.



## **Relational databases**

A **relational database** is a database that contains a series of tables that can be connected to form relationships. Basically, they allow data analysts to organize and link data based on what the data has in common.

In a non-relational table, you will find all of the possible variables you might be interested in analyzing all grouped together. This can make it really hard to sort through. **This is one reason why relational databases are so common in data analysis:** they simplify a lot of analysis processes and make data easier to find and use across an entire database.

**Normalization** is a process of organizing data in a relational database. For example, creating tables and establishing relationships between those tables. **It is applied to eliminate data redundancy, increase data integrity, and reduce complexity in a database.**

## **The key to relational databases**

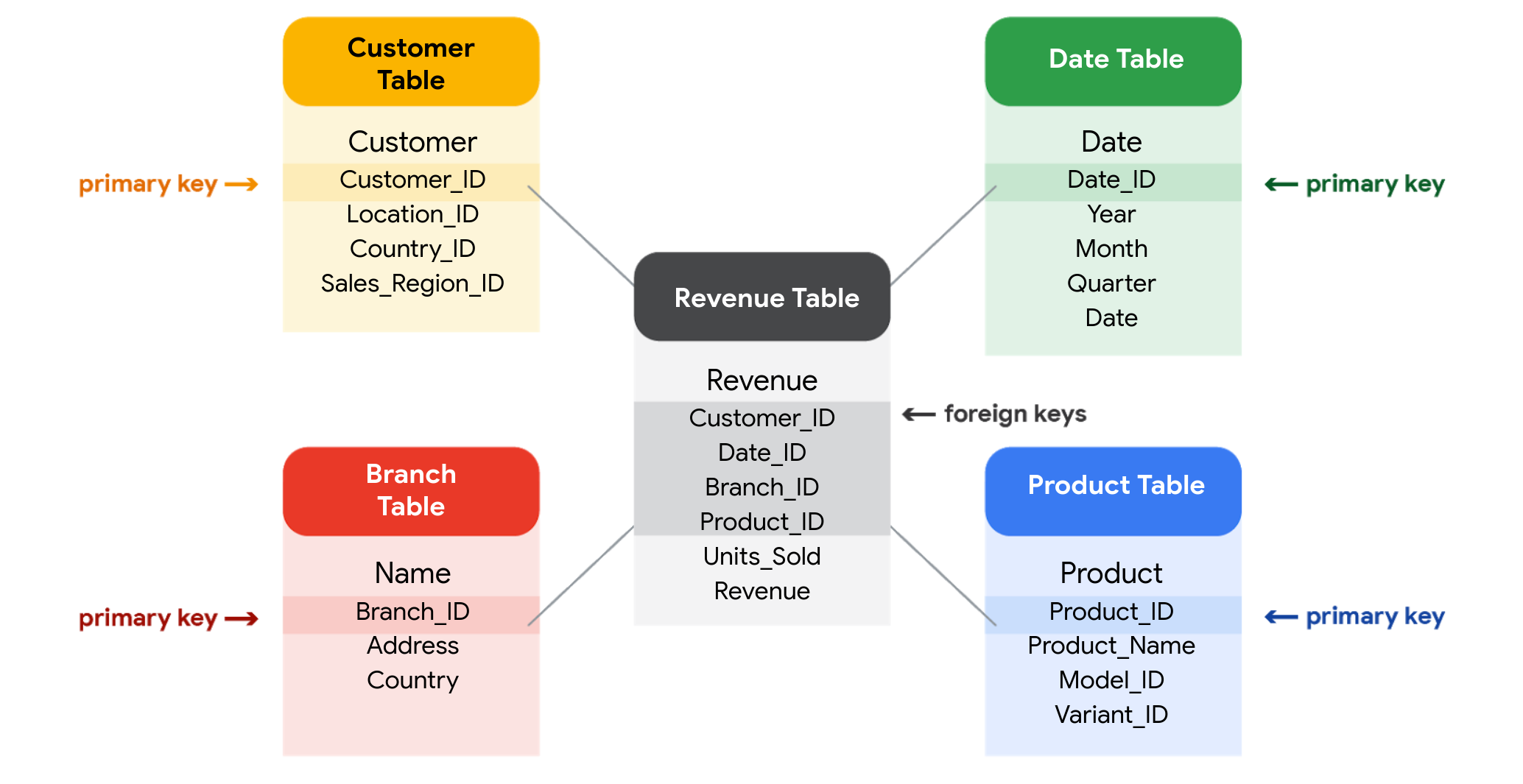
Tables in a relational database are connected by the fields they have in common. You might remember learning about primary and foreign keys before. As a quick refresher, a **primary key** is an identifier that references a column in which each value is unique. In other words, it's a column of a table that is used to uniquely identify each record within that table. The value assigned to the primary key in a particular row must be unique within the entire table. For example, if customer\_id is the primary key for the customer table, no two customers will ever have the same customer\_id.

By contrast, a **foreign key** is a field within a table that is a primary key in another table. A table can have only one primary key, but it can have multiple foreign keys.

These keys are what create the relationships between tables in a relational database, which helps organize and connect data across multiple tables in the database.

**Some tables don't require a primary key. For example, a revenue table can have multiple foreign keys and not have a primary key.**

**A primary key may also be constructed using multiple columns of a table**. This type of primary key is called a **composite key**. For example, if customer\_id and location\_id are two columns of a composite key for a customer table, the values assigned to those fields in any given row must be unique within the entire table.



## **SQL? You’re speaking my language**

As you've been learning, **Structured Query Language** (SQL) **is a type of query language that enables data analysts to communicate with a database**. So, a data analyst will use SQL to create a query to view the specific data that they want from within a larger dataset. In a relational database, data analysts can write queries to get data from the related tables. SQL is a powerful tool for working with databases—which is why you are going to learn more about it coming up!

[**INSPECT A DATASET: A GUIDED, HANDS-ON TOUR**](https://www.coursera.org/learn/data-preparation/supplement/0FIHG/inspect-a-dataset-a-guided-hands-on-tour)

As a data analyst, you'll use data to answer questions and solve problems. When you analyze data and draw conclusions, you are generating insights that can influence business decisions, drive positive change, and help your stakeholders meet their goals.

Before you begin an analysis, it’s important to inspect your data to determine if it contains the specific information you need to answer your stakeholders’ questions. In any given dataset, it may be the case that:

* The data is not there (you have sandwich data, but you need pizza data)
* The data is insufficient (you have pizza data for June 1-7, but you need data for the entire month of June)
* The data is incorrect (your pizza data lists the cost of a slice as $250, which makes you question the validity of the dataset)

Inspecting your dataset will help you pinpoint what questions are answerable and what data is still missing. You may be able to recover this data from an external source or at least recommend to your stakeholders that another data source be used.

In this reading, imagine you’re a data analyst inspecting spreadsheet data to determine if it’s possible to answer your stakeholders’ questions.

## 

## **The scenario**

You are a data analyst working for an ice cream company. Management is interested in improving the company's ice cream sales.

The company has been collecting data about its sales—but not a lot. The available data is from an internal data source and is based on sales for 2019. You’ve been asked to review the data and provide some insight into the company’s ice cream sales. Ideally, management would like answers to the following questions:

1. What is the most popular flavor of ice cream?
2. How does temperature affect sales?
3. How do weekends and holidays affect sales?
4. How does profitability differ for new versus returning customers?

## **Download the data**

You can download the data to follow along with this reading. To use the template for the sales data, click the link below and select “Use Template.”

Link to template: [Ice Cream Sales](https://docs.google.com/spreadsheets/d/1NgiKb8wCnJbUTuUkDUiNRpx9NhwncEmoKuPvgfYfOIY/template/preview?resourcekey=0-X3e7NzehG2Y74MIBhOaqeQ#gid=653912415)

OR

If you don’t have a Google account, you can download the spreadsheets directly from the attachments below:

[SalesByTemp](https://d3c33hcgiwev3.cloudfront.net/jmigEulNR7yooBLpTYe8Cw_9ecaf818f1a74b7987fe6a7d9af3c1f1_SalesByTemp.xlsx?Expires=1711238400&Signature=Z4CqaxIyXW5OavHref7vuTvgXBArgO4iY7pKaRzBzBJT0Qwot26dgbLMdZhRf-bdFxtwxOff4anZanztuLxUQrOgucQrdp8Lag562PlsjOa5bjXiJn7-7Vw71LvvlxJ-e1xaXCRP~J-6atG8VYGnQdKL07aGmX4ZEouL7sJRiiA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/jmigEulNR7yooBLpTYe8Cw_9ecaf818f1a74b7987fe6a7d9af3c1f1_SalesByTemp.xlsx?Expires=1711238400&Signature=Z4CqaxIyXW5OavHref7vuTvgXBArgO4iY7pKaRzBzBJT0Qwot26dgbLMdZhRf-bdFxtwxOff4anZanztuLxUQrOgucQrdp8Lag562PlsjOa5bjXiJn7-7Vw71LvvlxJ-e1xaXCRP~J-6atG8VYGnQdKL07aGmX4ZEouL7sJRiiA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[SalesByDay](https://d3c33hcgiwev3.cloudfront.net/B3ofmLtERPq6H5i7RFT6Pg_1ca5eec9c08941518e2c16034a2e65f1_SalesByDay.xlsx?Expires=1711238400&Signature=SCHwjdKSsCIlFBag5TNW2wwnrgVGifOcVN0XsXMngSJ1BvFWkjsTwDidWXWxCWO330pr6C3qQWfxGkqT6kUbVJBu4qhk-1DRySfjDuRzZEkZXNxcZWd7b6r-VFh3DjfTb5LIjMvNmuWZDKx1U-uO12HRX2iHyJOZKRjlxdVxM8U_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/B3ofmLtERPq6H5i7RFT6Pg_1ca5eec9c08941518e2c16034a2e65f1_SalesByDay.xlsx?Expires=1711238400&Signature=SCHwjdKSsCIlFBag5TNW2wwnrgVGifOcVN0XsXMngSJ1BvFWkjsTwDidWXWxCWO330pr6C3qQWfxGkqT6kUbVJBu4qhk-1DRySfjDuRzZEkZXNxcZWd7b6r-VFh3DjfTb5LIjMvNmuWZDKx1U-uO12HRX2iHyJOZKRjlxdVxM8U_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[SalesByFlavor](https://d3c33hcgiwev3.cloudfront.net/DHN9hYWCSDCzfYWFgvgwgg_b0e0d35f6a4f4bde9c84ecd0dd69c0f1_SalesByFlavor.xlsx?Expires=1711238400&Signature=WmDSTTHuZBq8KHICDznaZ5hFi-wh5GQ6zpRG-6NIpK1fE5chbFe2xDFOrq4f9G9emjDsLeO6Vq8QjKk~3rT9OcCzBC1u3wqHpIXA5yL43wMfVKaBs-RC1A7JnZEEO6D5O6xAOLtdFoFDMPeTPv7qqLm2UfamWxHwtegU0ND1JLM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/DHN9hYWCSDCzfYWFgvgwgg_b0e0d35f6a4f4bde9c84ecd0dd69c0f1_SalesByFlavor.xlsx?Expires=1711238400&Signature=WmDSTTHuZBq8KHICDznaZ5hFi-wh5GQ6zpRG-6NIpK1fE5chbFe2xDFOrq4f9G9emjDsLeO6Vq8QjKk~3rT9OcCzBC1u3wqHpIXA5yL43wMfVKaBs-RC1A7JnZEEO6D5O6xAOLtdFoFDMPeTPv7qqLm2UfamWxHwtegU0ND1JLM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

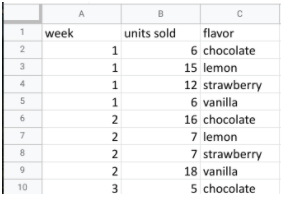
[](https://d3c33hcgiwev3.cloudfront.net/DHN9hYWCSDCzfYWFgvgwgg_b0e0d35f6a4f4bde9c84ecd0dd69c0f1_SalesByFlavor.xlsx?Expires=1711238400&Signature=WmDSTTHuZBq8KHICDznaZ5hFi-wh5GQ6zpRG-6NIpK1fE5chbFe2xDFOrq4f9G9emjDsLeO6Vq8QjKk~3rT9OcCzBC1u3wqHpIXA5yL43wMfVKaBs-RC1A7JnZEEO6D5O6xAOLtdFoFDMPeTPv7qqLm2UfamWxHwtegU0ND1JLM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

## **Inspect the data**

### **Question 1: What is the most popular flavor of ice cream?**

To discover the most popular flavor, you first need to define what is meant by "popular." Is the most popular flavor the one that generated the most revenue in 2019? Or is it the flavor that had the largest number of units sold in 2019? Sometimes your measurement choices are limited by what data you have—you can review your spreadsheet to find out if either of these definitions of “popular” make sense based on the available data.

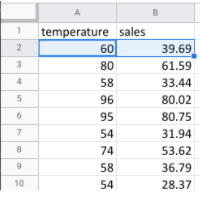
Click the **flavors** tab on your spreadsheet to view the relevant data. The **flavors** sheet has three columns and 209 rows of data. The column headers are **week**, **units sold***,* and **flavor**. This dataset did not come with a data description, so you have to figure out the significance of the columns on your own. Based on the data, you deduce that these columns provide information about the number of units sold for each ice cream flavor, by week, in 2019



In this case, you can discover what the most popular flavor is by using units sold as your measure. In particular, you can use the **units sold** column to calculate the total number of units sold during the year for each flavor*.* Unfortunately, the dataset does not provide the annual sales amount by flavor. In this case, your next step would be to ask your stakeholders if the annual sales per flavor data is available from another source. If not, you can add a statement about the current data’s limitations to your analysis.

### **Question 2: How does temperature affect sales?**

To explore your second question, you click the **temperatures** tab and check out the data. The **temperature** sheet has two columns and 366 rows of data. The column headers are **temperature** and **sales**. The data may show total 2019 sales per temperature (for instance, the first entry might sum up $39.69 in sales for three separate days that each had a high of 60 degrees). Or, the data may show a snapshot of sales and temperature for each day in 2019 (for instance, the first entry might refer to a single day with a high of 60 degrees and $39.69 in sales).

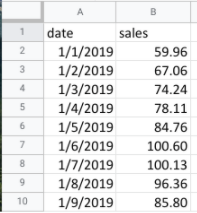


So, which is it? It’s probably a daily snapshot because there are 365 entries for temperature, and multiple rows with the same temperature and different sales values. This implies that each entry is for a single day and not a summary of multiple days. However, without more information, you can’t be certain. Plus, you don’t know if the current data is listed in consecutive order by date or in a different order. Your next step would be to contact the owner of the dataset for clarification.

If it turns out that temperature does affect sales, you’ll be able to offer your stakeholders an insight such as the following: “When daily highs are above X degrees, average ice cream sales increase by Y amount. So the business should plan on increasing inventory during these times to maximize sales.”

### **Question 3: How do weekends and holidays affect sales?**

Next, you click on the **sales** tab to view the data about dates of sale. The **sales** sheet has two columns and 366 rows of data. The column headers are **date** and **sales**. This data is most likely total daily sales in 2019, as sales are recorded for each date in 2019.



You can use this data to determine whether a specific date falls on a weekend or holiday and add a column to your sheet that reflects this information. Then, you can find out whether sales on the weekends and holidays are greater than sales on other days. This will be useful to know for inventory planning and marketing purposes.

### 

### **Question 4: How does profitability differ for new customers versus returning customers?**

Your dataset does not contain sales data related to new customers. Without this data, you won’t be able to answer your final question. However, it may be the case that the company collects customer data and stores it in a different data table.

If so, your next step would be to find out how to access the company’s customer data. You can then join the revenue sales data to the customer data table to categorize each sale as from a new or returning customer and analyze the difference in profitability between the two sets of customers. This information will help your stakeholders develop marketing campaigns for specific types of customers to increase brand loyalty and overall profitability.

## **Key takeaways**

When working on analytics projects, you won’t always have all the necessary or relevant data at your disposal. In many of these cases, you can turn to other data sources to fill in the gaps.

Despite the limitations of your dataset, it’s still possible to offer your stakeholders some valuable insights. For next steps, your best plan of action will be to take the initiative to ask questions, identify other relevant datasets, or do some research on your own. No matter what data you’re working with, carefully inspecting your data makes a big impact on the overall quality of your analysis.

[**TEST YOUR KNOWLEDGE ON WORKING WITH DATABASES**](https://www.coursera.org/learn/data-preparation/quiz/rz8nP/test-your-knowledge-on-working-with-databases)

**MANAGE DATA WITH METADATA**

[**DEMYSTIFY METADATA**](https://www.coursera.org/learn/data-preparation/lecture/vsFl7/demystify-metadata)

Now that you understand the different ways to organize data in a database, let's talk about **how** you can **describe** that **data**. In this video, we'll start exploring **metadata**, which is a very **important aspect** of **database management**.

**Metadata is an abstract concept**, though. Let's kick things off with a simple, everyday example. Did you know that every time a photo is taken with a smartphone, data is automatically collected and stored within that photo? Take a look. Choose any photo on your computer. Here's a cute shot of my friend's dogs, Rudy and Matilda. On your photo, right-click on "Get Info" or "Properties."

This will give you the **photo's metadata**, which may tell you the **type of file it is; the date and time it was taken; the geolocation, or where it was taken; what kind of device was used to take the photo; and much more**.

Another example. Every time you send or receive an email, metadata is sent right along with that message. You can find it by clicking on "Show Original" or "View Message Details." An email message's metadata includes its subject, who it's from, who it's to, and the date and time it was sent. The metadata even knows how quickly it was delivered after the sender pressed, "Send."

Metadata is information that's used to describe the data that's contained in something, like a photo or an email.

Keep in mind that metadata is not the data itself. Instead, it's data about the data. In data analytics, metadata **helps data analysts interpret the contents of the data within a database**. That's why metadata is so **important** when **working with databases**. It tells an analyst what the data is all about. That makes it possible to put the data to work solving problems and making data-driven decisions.

As a data analyst, there are **three common types of metadata** that you'll come across: **administrative**, **descriptive**, and **structural**.

**3 Common types of metadata: A D S**

**Administrative** metadata which is metadata that indicates the technical source of a digital asset.

When we looked at the metadata inside the photo, that was administrative metadata. It shows you the type of file it was, the date and time it was taken, and much more.

**Descriptive** metadata is metadata that describes a piece of data and can be used to identify it at a later point in time. For instance, the descriptive metadata of a book in a library would include the code you see on its spine, known as a unique International Standard Book Number, also called the ISBN. It would also include the book's author and title.

**Structural** metadata, which is metadata that describes how many locations contain a certain piece of data, and indicates **how** a piece of data is organized and whether it's part of one or more than one data collection. Let's head back to the library. An example of structural data would be how the pages of a book are put together to create different chapters. It's important to note that structural metadata also keeps track of the relationship between two things. For example, it can show us that the digital document of a book manuscript was actually the original version of a now printed book.

**MORE EXAMPLES OF THE THREE TYPES OF METADATA:**

**Administrative metadata**

* File type of a digital document
* Date a digital document was created
* Date a digital document was last modified
* Author of a digital document
* Copyright information for a digital document

**Descriptive metadata**

* Title of a book
* Author of a book
* Date a book was published
* ISBN of a book
* Keywords associated with a book

**Structural metadata**

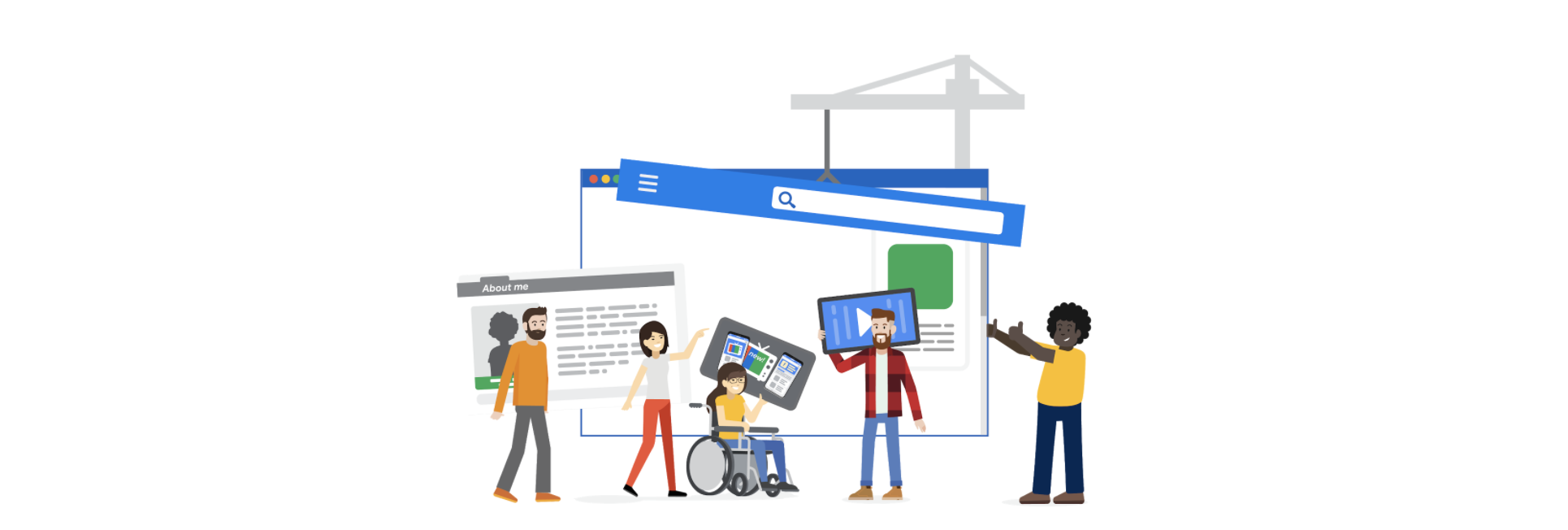
* Table of contents of a book
* Chapters of a book
* Sections of a book
* Relationships between different parts of a book

**Here's one final thought to help you understand metadata.** If you're on your way to the library to pick out a book, you could research a book's title, author, length, and number of chapters. That's all metadata, and it can tell you a lot about the book, but you have to actually read the book to know what it's all about.

Likewise, you can read about data analytics, but you have to take this course to earn the Google Data Analytics certificate. Keep moving forward to gain that new perspective.

[**METADATA IS AS IMPORTANT AS THE DATA ITSELF**](https://www.coursera.org/learn/data-preparation/supplement/mdF9p/metadata-is-as-important-as-the-data-itself)

Data analytics, by design, is a field that thrives on collecting and organizing data. In this reading, you’ll learn about metadata and the type of information it can provide. In addition, you’ll explore examples of metadata.



Explore a data file by opening any file on your computer or a document in your home or workplace. What is it? Where did it come from? Is it useful? How do you know? This is where metadata comes in to provide a deeper understanding of the data. To put it simply, **metadata** is data about data. In database management, metadata provides information about other data and helps data analysts interpret the contents of the data within a database.

Regardless of whether you’re working with a large or small quantity of data, metadata is the mark of a knowledgeable analytics team. Metadata helps people communicate about data across the business and makes it easier to reuse data. In essence, metadata **tells the who, what, when, where, which, why, and how of data**.

## **Elements of metadata**

Before examining metadata examples, it’s important to understand what type of information metadata typically provides:

* **File or document type:** What type of file or document are you examining?
* **Date, time, and creator:** When was it created? Who created it? When was it last modified?
* **Title and description:** What is the name of the item you are examining? What type of content does it contain?
* **Geolocation:** If you’re examining a photo, where was it taken?
* **Tags and categories:** What is the general overview of the item that you have? Is it indexed or described in a specific way?
* **Who last modified it and when:** Were any changes made to the file? If yes, when were the most recent modifications made?
* **Who can access or update it:** If you’re examining a dataset, is it public? Are special permissions needed to customize or modify it?

## **Examples of metadata**

In today’s digital world, metadata is everywhere! Here are some examples—with accompanying images—of where you might find metadata.

### **Photos**

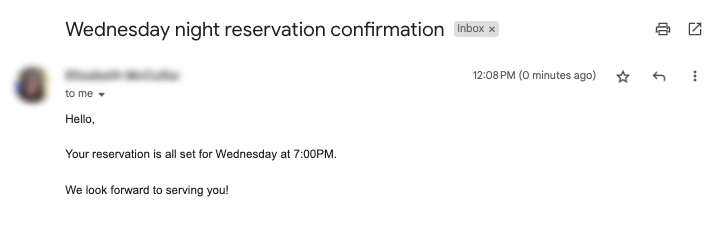
Whenever a photo is captured with a camera, metadata such as filename, date, time, geolocation, and the type of device on which it was taken are gathered and saved with it. The metadata of the following photo is displayed as a pop-up alongside the photo.



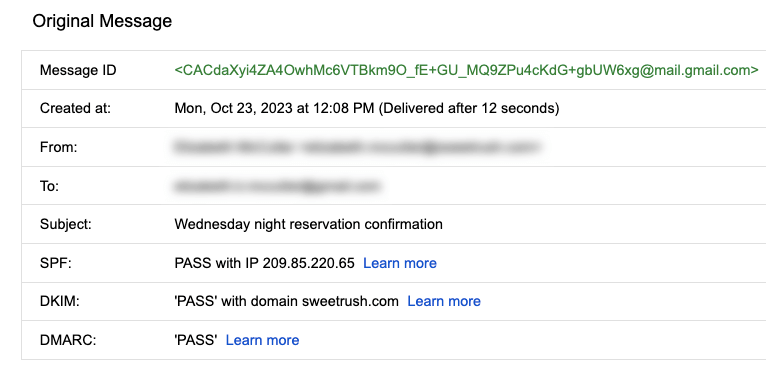
Image with accompanying Information pop-up that displays its description, the date and time the image was taken, its size, the device on which it was taken, and an option to add the geolocation of the image.

### **Emails**

When an email is sent or received, it contains metadata such as subject line, sender, recipient, date sent, and time sent.



Emails also contain hidden metadata that includes server names, IP addresses, HTML format, and software details. This image includes hidden email metadata such as the message ID and when the email was created.



Hidden metadata from an email that includes Message ID, creation date, the recipient, the sender, the subject line, the SPF, DKIM, and DMARC.

### **Spreadsheets and electronically created documents**

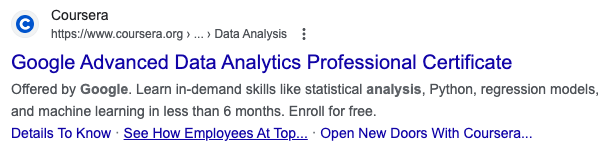
Spreadsheets and documents are already filled with a considerable amount of data, so it’s no surprise that they also include metadata such as title, author, creation date, number of pages, and user comments. Additionally, spreadsheet metadata includes tab names, tables, and columns. In the following example, the image demonstrates the metadata for an electronically created Google Sheet:



The metadata of a Google Sheet including the title, type, size, storage used, owner, the last person who modified the document, the last person who opened the document, when it was created, download permissions, and an option to include a description.

### **Websites**

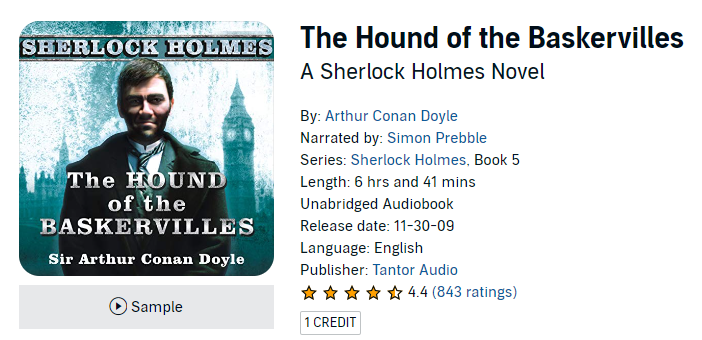
Every web page has a number of standard metadata fields such as tags and categories, the site creator’s name, web page title and description, and time of creation. Results of search engine queries that you might make on a daily basis are metadata!



A search engine result that includes Meta title as Google Advanced Data Analytics Professional Certificate. Under the website hyperlink is the Meta description that says: Offered by Google. Learn in-demand skills like statistical analysis, Python, regression models, and machine learning in less than 6 months. Enroll for free.

### **Books and audiobooks**

Non-digital items can have metadata, too! Every book has standard metadata that will inform you of its title, author’s name, a table of contents, publisher information, copyright description, index, and a brief description of the book’s contents. Audiobook metadata also includes this data, as well as metadata specific to the audiobook such as narrator and recording length.



Audiobook’s metadata including the title of the audiobook, author, narrator, its length, the release date, language the audiobook is read in, its published, and its rating score.

## **Key takeaways**

Metadata can be found in photos, emails, spreadsheets, websites, and much more! In your daily life, you use metadata to stay organized. As a data analyst, you’ll **use metadata to understand** the **content** and **context** of your **data**, as well as **how it’s structured**. Metadata provides data analysts with information about a data’s file type, title, geolocation, who created it, who last modified it, and who has access to it. As a data analyst, it’s important to keep accurate records of metadata to ensure that you are able to find, use, preserve, and reuse data in the future. Remember, it will be your responsibility to manage and make use of data in its entirety; metadata is as important as the data itself.

[**METADATA AND METADATA REPOSITORIES**](https://www.coursera.org/learn/data-preparation/supplement/noiNb/metadata-and-metadata-repositories)

As you’re learning, metadata is data about data. It clearly describes how and when data was collected and how it’s organized. Metadata puts data into context and makes the data more understandable. This helps data analysts use data to solve problems and make informed business decisions.

In this reading, you’ll learn more about the **benefits of metadata, metadata repositories, and metadata of external databases.**

## **The benefits of metadata**

### **Reliability**

Data analysts use reliable and high-quality data to identify the root causes of any problems that might occur during analysis and to improve their results. If the data being used to solve a problem or to make a data-driven decision is unreliable, there’s a good chance the results will be unreliable as well.

Metadata helps data analysts confirm their data is reliable by making sure it is:

* Accurate
* Precise
* Relevant
* Timely

It does this by helping analysts ensure that they’re working with the right data and that the data is described correctly. For example, a data analyst completing a project with data from 2022 can use metadata to easily determine if they should use data from a particular file.

### **Consistency**

Data analysts thrive on consistency and aim for uniformity in their data and databases, and metadata helps make this possible. For example, to use survey data from two different sources, data analysts use metadata to make sure the same collection methods were applied in the survey so that both datasets can be compared reliably.

When a database is consistent, it’s easier to discover relationships between the data inside the database and data that exists elsewhere.

**When data is uniform, it is**:

* **Organized**: Data analysts can easily find tables and files, monitor the creation and alteration of assets, and store metadata.
* **Classified**: Data analysts can categorize data when it follows a consistent format, which is beneficial in cleaning and processing data.
* **Stored**: Consistent and uniform data can be efficiently stored in various data repositories. This streamlines storage management tasks such as managing a database.
* **Accessed**: Users, applications, and systems can efficiently locate and use data.

Together, these benefits empower data analysts to effectively analyze and interpret their data.

## 

## **Metadata repositories**

Metadata repositories help data analysts ensure their data is reliable and consistent.

Metadata repositories are **specialized databases specifically created to store and manage metadata**. They can be kept in a physical location or a virtual environment—like data that exists in the cloud.

Metadata repositories describe where the metadata came from and store that data in an accessible form with a common structure. This provides data analysts with quick and easy access to the data. If data analysts didn’t use a metadata repository, they would have to select each file to look up its information and compare the data manually, which would waste a lot of time and effort.

Data analysts also use metadata repositories to bring together multiple sources for data analysis. Metadata repositories do this by describing the state and location of the data, the structure of the tables inside the data, and who accessed the user logs.

## **Metadata of external databases**

Data analysts **use both second-party and third-party data** to gain valuable insights and make strategic, data-driven decisions.

Second-party data is data that’s collected by a group directly from the group’s audience and then sold.

Third-party data is provided by outside sources that didn’t collect it directly. The providers of this data are not its original collectors and do not have a direct relationship with any individuals to whom the data belongs. The outside providers get the data from websites or other programs that pull it from the various platforms where it was originally generated.

Data analysts should understand the metadata of external databases to confirm that it is consistent and reliable. In some cases, they should also contact the owner of the third-party data to confirm that it is accessible and available for purchase. Confirming that the data is reliable and that the proper permissions to use it have been obtained are best practices when using data that comes from another organization.

**Key takeaways**

Metadata helps data analysts make data-driven decisions more quickly and efficiently. It also ensures that data and databases are reliable and consistent.

**Metadata repositories are used to store metadata—including data from second-party and third-party companies.** These repositories describe the state and location of the metadata, the structure of the tables inside it, and who has accessed the repository. Data analysts use metadata repositories to ensure that they use the right data appropriately.

[**MANAGE DATA WITH METADATA**](https://www.coursera.org/learn/data-preparation/lecture/C7QQB/manage-data-with-metadata)

Metadata and metadata repositories are very powerful tools in the data analyst toolbox. As we discussed previously, data analysts use them to create a single source of truth, keep data consistent and uniform, and ensure that the data we work with is accurate, precise, relevant, and timely.

These tools also make it easier to access and use data by standardizing our processes. In this video, we'll explore more components of metadata and learn how metadata analysts work to keep things organized.

We know that the amount of data out there continues to grow, but lots of businesses just aren't using their data. Sometimes, they don't know what they have, sometimes they can't find it or sometimes a business just doesn't trust it. Especially in bigger companies, data can span numerous different processes and systems. And pulling together data from so many places can be a big challenge. For example, let's say a company starts out with a traditional data storage system in its offices. But then, as the amount of data it owns continues to expand, cloud storage is needed too. Plus, this company could also be accessing and using second or third party data from a partner organization. Each of these systems has its own rules and requirements, so each organizes the data in a completely different way, adding even more complexity. It's no wonder so many organizations struggle to find the right data at the right moment.

On the other hand, metadata is stored in a single, central location and it gives the company **standardized information about all of its data**.

**This is done in two ways**.

**First**, metadata includes information about where each system is located and where the data sets are located within those systems.

**Second**, the metadata describes how all of the data is connected between the various systems.

Another important aspect of metadata is something called data governance.

**Data governance** is a process to ensure the formal management of a company’s data assets. This gives an organization better control of their data and helps a company manage issues related to data security and privacy, integrity, usability, and internal and external data flows.

It's **important** to note that data governance is about more than just standardizing terminology and procedures. It's about the **roles and responsibilities of the people who work with the metadata every day**. These are metadata specialists, and they organize and maintain company data, ensuring that it's of the highest possible quality. These people create basic metadata identification and discovery information, describe the way different data sets work together, and explain the many different types of data resources. Metadata specialists also create very important standards that everyone follows and the models used to organize the data. There's one thing they all have in common. Whether they work at a tech company, a nonprofit association, or a financial institution, **metadata analysts are great team players**.

They're passionate about making data accessible by sharing with colleagues and other stakeholders. If you're looking for a role that encourages you to explore all the data that the digital world has to offer, following the **path to becoming a metadata analyst** may be the **right choice for you**.

But either way, businesses of all kinds face market trends and competition, and they need to understand why one process works while another doesn't. Data analytics allows them to answer key questions and keep improving.

[**MEGAN: FUN WITH METADATA**](https://www.coursera.org/learn/data-preparation/lecture/bHqPA/megan-fun-with-metadata)

My name is Megan, and I am an agency measurement lead here at Google. Basically, I help to demystify measurement and analytics for advertising agencies. So people that are tasked with executing media plans for advertisers but also people that are interested in measuring the impact that media is having for their clients. So I've been doing this for about 17 years now and have seen a lot of evolution in the space from data availability, from different modeling techniques becoming more advanced but also more accessible, and it's just been a really cool journey to see how it's evolved, how analytics has become more mainstream, and how people are getting more excited about it. **Metadata is basically the key to your larger data set**. **It helps describe what's in the rows and the columns of the data that you'll be working with**. Metadata is kind of a shorthand or a CliffsNotes version of a much more complex set of information. It can be helpful in just kind of helping you get a handle on what's in a single data set that you may have access to.

It's an important part of the discovery process of any analytics project as you're working with either a client or a vendor to understand the resources that you'll have to address a problem and what might be missing. **It just gives you the keys to unlock that data in a really simple and straightforward way and is a great communication tool**.

When I was working for an advertiser, one of the things that we were trying to do was build something called a data lake. So essentially, this is bringing together all of the sources of data that you might want to use in an analysis into one place, which can be really, really tricky. One of the benefits of metadata was figuring out where we had sources that may overlap, where we had data sources that had things in common. And what the unique pieces of information were that we were getting from each of those data sets. So as we thought about tackling this really huge and important project, we were able to use metadata to quickly and easily get to the basic constructs that we were trying to tackle. When you're working with people who maybe don't have analytics as their day job, getting that "aha" moment, helping them understand how measurement and analytics are tools that can help them achieve their goals, is really important. And just getting to that idea that you made something that was previously inaccessible a little bit more accessible for that team and something they feel comfortable putting into practice is really important and really kind of a great way to come out of a partnership.

[**TEST YOUR KNOWLEDGE ON METADATA**](https://www.coursera.org/learn/data-preparation/quiz/gIn3V/test-your-knowledge-on-metadata)

**ACCESS DIFFERENT DATA SOURCES**

[**SO MANY PLACES TO FIND DATA**](https://www.coursera.org/learn/data-preparation/lecture/jj7B1/so-many-places-to-find-data)

We'll discuss the different places data analysts go to connect with data. There's all kinds of data out there and it's important to know how to access it.

Earlier, you learned that there are **two basic types of data** used by data analysts: **internal and externa**l.

**Internal data** is data that lives within a company's own systems. It's typically also generated from within the company. You may also hear internal data described as primary data.

Gathering internal data can be complicated. Depending on your data analytics project, you might need data from lots of different sources and departments, including sales, marketing, customer relationship management, finance, human resources, and even the data archives. But the effort is worth it. Internal data has plenty of advantages for a business. It provides information that's relevant to problems you're trying to solve, and it's free to access because the company already owns it. With internal data, analysts can work on all data projects without ever looking beyond their own walls. But sometimes internal data doesn't give you the full picture. In those cases, data analysts can turn to external data and apply that information to their analysis.

**External data** is data that lives and is generated outside an organization. It can come from a variety of places, including other businesses, government sources, the media, professional associations, schools, and more. External data is sometimes called **secondary data**.

For instance, as health care analysts, we often partner with other healthcare organizations or nonprofits and use their data to create deeper analyses and add some more industry- level perspective.

In an earlier video, you learned that **openness** has created a lot of data for analysts to use, largely through open data initiatives. As a reminder, **openness or open data** refers to the free access, usage and sharing of data.

For example, the United States government makes hundreds of thousands of data sets available to the public on Data.gov. These data sets contain information on weather patterns, educational progress, crime rates, transportation, and much more.

**There are lots of reasons for these open data initiatives.**

One is to make government activities more transparent, like letting the public see where money is spent. It also helps educate citizens about voting and local issues.

Open data also improves public service by giving people ways to be a part of public planning or provide feedback to the government.

Finally, open data leads to innovation and economic growth by helping people and companies better understand their markets. Google actually hosts lots of public databases with information on science, transportation, economics, climate, and more.

As an example, a bike sharing company could use traffic data from within our public transportation database to see where the roads are busiest, then choose those locations for their bikes in order to reduce cars on the road and give people another transportation option. Now you're familiar with internal and external data and how you can access both.

[**STEP-BY-STEP: IMPORT DATA FROM SPREADSHEETS AND DATABASES**](https://www.coursera.org/learn/data-preparation/supplement/esVz6/step-by-step-import-data-from-spreadsheets-and-databases)

This reading outlines the steps the instructor performs in the following video, [Importing data from spreadsheets and databases](https://www.coursera.org/learn/data-preparation/lecture/KCphN/importing-data-from-spreadsheets-and-databases). The video teaches you how to import a **.csv(Comma-separated values)** file into a Google Sheet so you can analyze the data.

Keep this guide open as you watch the video. It can serve as a reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

## 

## **What you’ll need**

To follow along with the examples in this video, open a blank spreadsheet.



### **Example 1: Use the menu to import a .csv file**

Sometimes, you’ll need to import the data from a .csv file into a Google Sheet. A .csv file saves data in a table format. Follow the steps below to bring the data from a .csv file into a new spreadsheet.

1. In the menu, select **File** then **Import**. The **Import file window** will pop up.
2. Select **Upload** then **Browse** to select the .csv file to import.
3. Next select the **Import location**. You can:
   1. Create a new spreadsheet
   2. Insert the .csv data as a new sheet
   3. Replace spreadsheet
   4. Replace current sheet
   5. Append (add) the data to the current spreadsheet
   6. Replace the data starting with a specific cell.
4. Next, select the **Separator type**. Google Sheets defaults to automatically detecting separator, or delimiter, type. To manually set the delimiter type, select the dropdown menu under Separator type and choose the separator.
5. Next, determine if you would like the text to be imported with or without formatting. In the box next to **Convert text to numbers, dates, and formulas**, keep the checkmark if you want text data to be formatted.
6. Select **Import data**. The data in the .csv file will be loaded into your sheet, and you can begin using it.

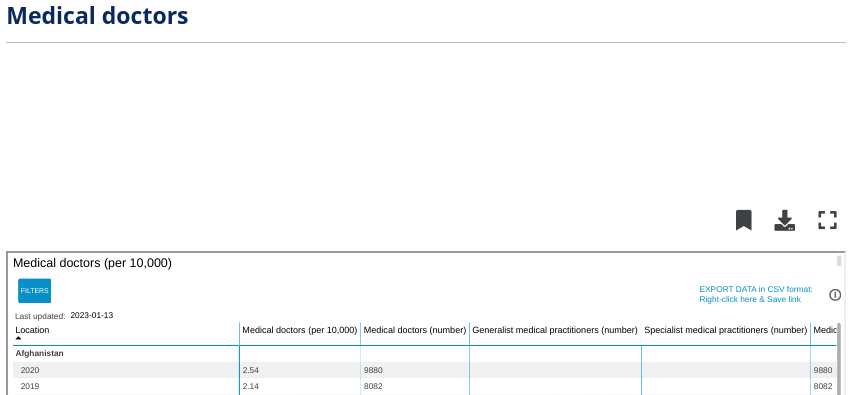
### **Example 2: Download data from the Global Health Observatory**

You can download public data from the internet, such as data from the World Health Organization and the Global Health Observatory.

**Note:** The Global Health Observatory’s website has been updated since this video was filmed. Follow these instructions to download the .csv file the instructor uses.

1. Navigate to the [Global Health Observatory workforce statistics database](https://www.who.int/data/gho/data/themes/topics/health-workforce).
2. Scroll to navigate to the **Medical doctors** table.
3. Then, scroll over **EXPORT DATA in .csv format** in the table. Right-click, then select **Save link as...**
4. The .csv file will download to your computer as **data.csv**.

**Note:** If you already have a .csv file named **data.csv**, your computer will add a number to the file name.



Medical doctors (number), General medical practitioners (number), and Specialist medical practitioners (number). The EXPORT DATA in CSV format is also in the screenshot

### **Example 3: Use the menu to import a .csv file**

Follow the steps below to bring the .csv file you downloaded, **data.csv**, into a new spreadsheet.

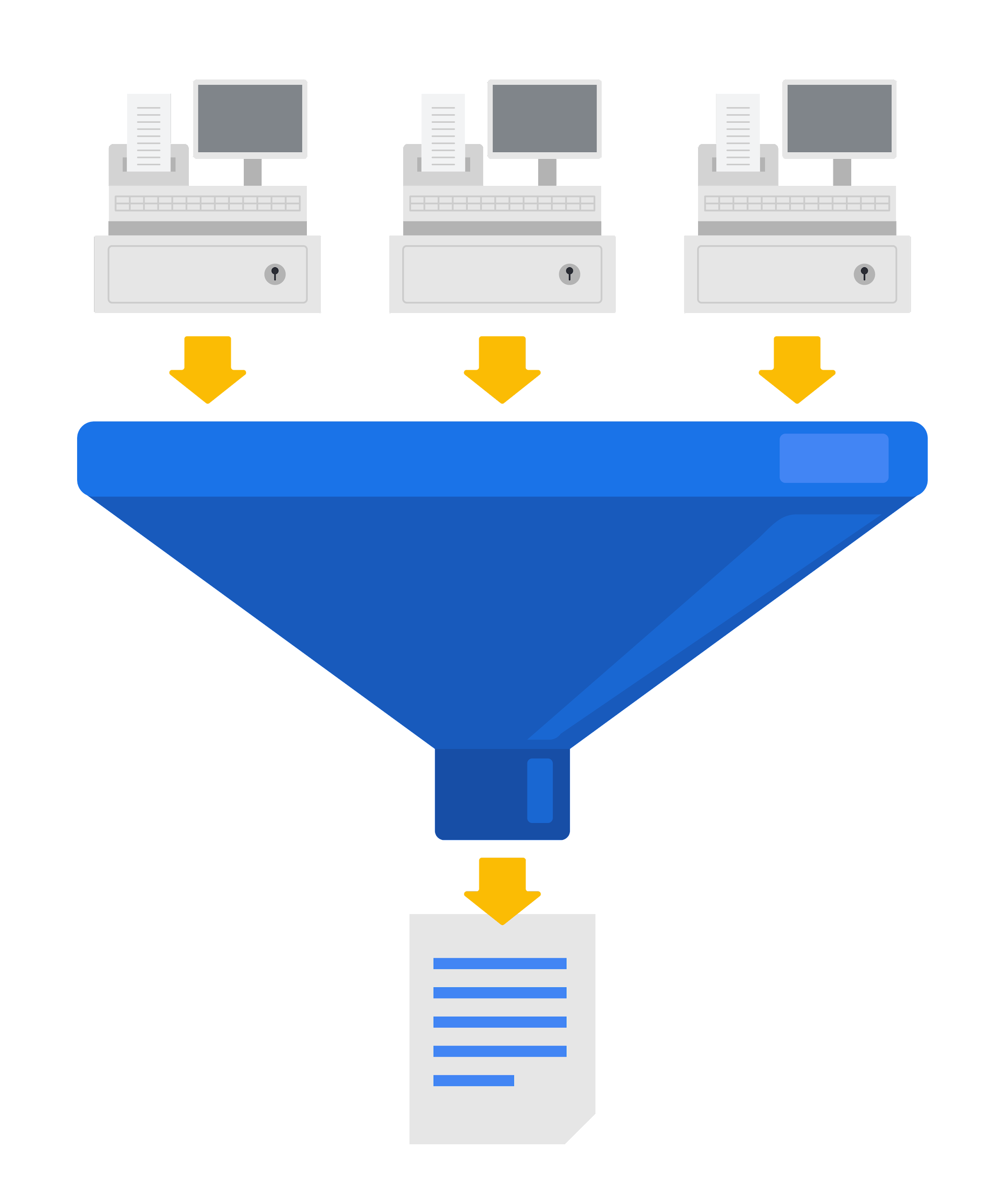
1. Open a blank spreadsheet.
2. In the menu, select **File**, then **Import**. The **Import file window** will pop up.
3. Select **Upload** then **Browse** to select the .csv file to import.
4. Next, select the **Import location**. You can:
   1. Create a new spreadsheet
   2. Insert the .csv data as a new sheet
   3. Replace spreadsheet
   4. Replace current sheet
   5. Append (add) the data to the current spreadsheet
   6. Replace the data starting with a specific cell.
5. Next, select the **Separator type**.
6. Next, determine if you would like the text to be imported with or without formatting. In the box next to **Convert text to numbers, dates, and formulas**, keep the checkmark if you want text data to be formatted.
7. Select **Import data**.
8. The data in the .csv file will be loaded into your sheet, and you can review and title it.

[**IMPORT DATA FROM SPREADSHEETS AND DATABASES**](https://www.coursera.org/learn/data-preparation/lecture/KCphN/import-data-from-spreadsheets-and-databases)

[**IMPORT DATA DYNAMICALLY**](https://www.coursera.org/learn/data-preparation/supplement/2slzg/import-data-dynamically)

As you’ve learned, you can import data from some data sources, like .csv files into a Google spreadsheet from the **File** menu. Keep in mind that, when you use this method, data that is updated in the .csv will not automatically be updated in the Google Sheet. Instead, it will need to be manually—and continually—updated in the Google Sheet. In some situations, such as when you want to be able to keep track of changes you’ve made, this method is ideal. In other situations, you might need to keep the data the same in both places, and using data that doesn’t update automatically can be time-consuming and tedious. Further, trying to maintain the same dataset in multiple places can cause errors later on.

Fortunately, there are tools to help you automate data imports so you don’t need to continually update the data in your current spreadsheet. Take a small general store as an example. The store has three cash registers handled by three clerks. At the end of each day, the owner wants to determine the total sales and the amount of cash in each register. Each clerk is responsible for counting their money and entering their sales total into a spreadsheet. The owner has the spreadsheets set up to import each clerks’ data into another spreadsheet, where it automates and calculates the total sales for all three registers. Without this automation, each clerk would have to take turns entering their data into the owner’s spreadsheet. This is an example of a dynamic method of importing data, which saves the owner and clerks time and energy. When data is dynamic, it is interactive and automatically changes and updates over time.



In the following sections you’ll learn how to import data into Google Sheets dynamically.

## 

## **IMPORT functions in Google Sheets**

### **The IMPORTRANGE function**

In Google Sheets, the **IMPORTRANGE** function can import all or part of a dataset from another Google Sheet.

To use this function, you need two pieces of information:

1. The URL of the Google Sheet from which you’ll import data.
2. The name of the sheet and the range of cells you want to import into your Google Sheet.

Once you have this information, open the Google Sheet into which you want to import data and select the cell into which the first cell of data should be copied. Enter **=** to indicate you will enter a function, then complete the **IMPORTRANGE** function with the URL and range you identified in the following manner: **=IMPORTRANGE("URL", "sheet\_name!cell\_range")**. Note that an exclamation point separates the sheet name and the cell range in the second part of this function.

An example of this function is:

**=IMPORTRANGE("https://docs.google.com/thisisatestabc123", "sheet1!A1:F13")**

**Note:** This URL is for syntax purposes only. It is not meant to be entered into your own spreadsheet.

Once you’ve completed the function, a box will pop up to prompt you to allow access to the Google Sheet from which you’re importing data. You must allow access to the spreadsheet containing the data the first time you import it into Google Sheets. Replace it with a spreadsheet’s URL that you have created so you can control access by selecting the Allow access button.

Refer to the Google Help Center's [IMPORTRANGE](https://support.google.com/docs/answer/3093340?hl=en&ref_topic=9199554) page for more information about the syntax. You’ll also learn more about this later in the program.

### **The IMPORTHTML function**

Importing HTML tables is a basic method to extract data from public web pages. This process is often called “scraping.” [Web scraping made easy](https://www.thedataschool.co.uk/anna-prosvetova/web-scraping-made-easy-import-html-tables-or-lists-using-google-sheets-and-excel) introduces how to do this with Google Sheets or Microsoft Excel.

In Google Sheets, you can use the **IMPORTHTML** function to import the data from an HTML table (or list) on a web page. This function is similar to the **IMPORTRANGE** function. Refer to the Google Help Center's [IMPORTHTML](https://support.google.com/docs/answer/3093339?hl=en) page for more information about the syntax.

### **The IMPORTDATA function**

Sometimes data displayed on the web is in the form of a comma- or tab-delimited file.

You can use the **IMPORTDATA** function in a Google Sheet to import data into a Google Sheet. This function is similar to the **IMPORTRANGE** function. Refer to Google Help Center's [IMPORTDATA](https://support.google.com/docs/answer/3093335?hl=en) page for more information and the syntax.

[**EXPLORE PUBLIC DATASETS**](https://www.coursera.org/learn/data-preparation/supplement/8yrhM/explore-public-datasets)

**Open data** helps create a lot of **public datasets** that you can access to make data-driven decisions. Here are some resources you can use to start searching for public datasets on your own:

* The [Google Cloud Public Datasets](https://cloud.google.com/public-datasets) allow data analysts access to high-demand public datasets, and make it easy to uncover insights in the cloud.
* The [Dataset Search](https://datasetsearch.research.google.com/) can help you find available datasets online with keyword searches.
* [Kaggle](https://www.kaggle.com/datasets?utm_medium=paid&utm_source=google.com+search&utm_campaign=datasets&gclid=CjwKCAiAt9z-BRBCEiwA_bWv-L6PpACh6RzmrJjQjmNGCCE7kky1FCtc6Jf1qld-4NwDMYL0WsUyxBoCdwAQAvD_BwE) has an Open Data search function that can help you find datasets to practice with.
* Finally, [BigQuery](https://cloud.google.com/bigquery/public-data) hosts 150+ public datasets you can access and use.

### **Public health datasets**

1. [Global Health Observatory data](https://www.who.int/data/collections): You can search for datasets from this page or explore featured data collections from the World Health Organization.
2. [The Cancer Imaging Archive (TCIA) dataset](https://cloud.google.com/healthcare/docs/resources/public-datasets/tcia): Just like the earlier dataset, this data is hosted by the Google Cloud Public Datasets and can be uploaded to BigQuery.
3. [1000 Genomes](https://cloud.google.com/life-sciences/docs/resources/public-datasets/1000-genomes): This is another dataset from the Google Cloud Public resources that can be uploaded to BigQuery.

### **Public climate datasets**

1. [National Climatic Data Center](https://www.ncei.noaa.gov/products): The NCDC Quick Links page has a selection of datasets you can explore.
2. [NOAA Public Dataset Gallery](https://www.climate.gov/maps-data/datasets): The NOAA Public Dataset Gallery contains a searchable collection of public datasets.

### **Public social-political datasets**

1. [UNICEF State of the World’s Children](https://data.unicef.org/resources/dataset/sowc-2019-statistical-tables/): This dataset from UNICEF includes a collection of tables that can be downloaded.
2. [CPS Labor Force Statistics](https://www.bls.gov/cps/tables.htm): This page contains links to several available datasets that you can explore.
3. [The Stanford Open Policing Project](https://openpolicing.stanford.edu/): This dataset can be downloaded as a .csv file for your own use.

[**TEST YOUR KNOWLEDGE ON ACCESSING DATA SOURCES**](https://www.coursera.org/learn/data-preparation/quiz/2fOBQ/test-your-knowledge-on-accessing-data-sources)

a

**SORT AND FILTER DATA**

[**SORT AND FILTER TO FOCUS ON RELEVANT DATA**](https://www.coursera.org/learn/data-preparation/lecture/LUq19/sort-and-filter-to-focus-on-relevant-data)

In the past few videos, you've learned about both internal and external data.

Now I'll show you how to focus on only the data that's relevant to the problem you're trying to solve.

This is **useful** **if** you're working with a very **large complex spreadsheet**, which data analysts encounter all the time. Having lots of data can make it difficult to quickly find and analyze the information you need. No two analytics projects are the same.

Often data analysts process, view, and use data very differently, even if it comes from the exact same source.

Here's an example. Check out this spreadsheet that shows a company's sales reps and where they work. Different data analysts might want different information from the spreadsheet, and that's where sorting and filtering comes in. Sorting and filtering the data in a spreadsheet helps us customize the way data is presented. They can also organize data so analysts can zoom in on the pieces that matter. Think of it like a magnifying glass for our data. Let's begin with sorting. Sorting involves arranging data into a meaningful order to make it easier to understand, analyze, and visualize. Data can be sorted in ascending or descending order, and alphabetically or numerically. Sorting can be done across all of a spreadsheet or just in a single column or table. You can also sort by multiple variables. For instance, if our data set contains both city and state fields, we can sort first by city and then by state.

Anytime you're sorting data, it's always a good idea to freeze the header row first. To do this, we'll highlight the row. Then from the view menu, choose freeze and one row.

This locks the row in place. Now when we scroll down the sheet, the header row stays visible so we know the category of each column.

Looks good to me. Now let's sort the entire spreadsheet. We'll sort by city first. To do this, select the city column,

then use the drop-down arrow to sort the sheet. Select A to Z.

This will sort all the columns from A to Z by row, with the selected column being the primary sort criteria. The cities are now sorted alphabetically, and they're still grouped with the corresponding states, sales reps, and auto parts. The details across each row are automatically kept together when sorting a particular section, as you can see here. Multiple criteria sorting is another very useful data analysis tool. For instance, let's say we want to see a list of sales reps by the cities and states in which they work. First, we select the entire data set,

then choose data and sort range.

In the dialog box, make sure that "Data has header row" is highlighted.

That way row A, city, states, sales rep, and auto parts won't be part of the sort.

Then in the sort by drop-down menu, select state and the sort order A to Z. Now add another sort column. In the "then by" drop-down, select city and then sort order A to Z.

Finally, select Sort.

Now we can search the data to easily find a sales rep who works in a particular state and city. Sorting is useful when you want to look at everything in a spreadsheet in alphabetical or numerical order. But sometimes data analysts want to isolate a particular piece of information. To do this, they use a filter. Filtering means showing only the data that meets a specific criteria while hiding the rest. A filter simplifies a spreadsheet by only showing us the information we need. For example, we could add a filter to see only the sales reps who worked with a particular product. To do this, we first select Data and Create a filter. Choose the column with the data we need. In this case, Auto Parts. Filter buttons will appear in the corner of each column header. To filter our spreadsheet by auto part, click the button in the Auto Parts header. In this example, let's say we want to only see sales reps who worked with rims. Remove the check marks from the categories we don't want to see, which is everything except for rims.

Then select okay.

The filter temporarily hides anything that doesn't meet the condition. But note that, even though they aren't visible, they're still there. When it's time to view the entire area spreadsheet again, simply turn off the filter.

Sorting and filtering are very important tools in the data analyst's toolbox. In the next video, you'll discover even more ways to narrow in on the exact information you need for any data analysis project.

[**HANDS-ON ACTIVITY: CLEAN DATA IN SPREADSHEETS WITH SORTING AND FILTERING**](https://www.coursera.org/learn/data-preparation/quiz/JY3Zv/hands-on-activity-clean-data-in-spreadsheets-with-sorting-and-filtering)

[**SELF-REFLECTION: COMPARE DATABASES AND SPREADSHEETS**](https://www.coursera.org/learn/data-preparation/quiz/a1ShE/self-reflection-compare-databases-and-spreadsheets)

[**TEST YOUR KNOWLEDGE ON SORTING AND FILTERING**](https://www.coursera.org/learn/data-preparation/quiz/HKePv/test-your-knowledge-on-sorting-and-filtering)

**LARGE DATASETS IN SQL**

[**GET TO KNOW BIGQUERY, WITH SANDBOX AND BILLING OPTIONS**](https://www.coursera.org/learn/data-preparation/lecture/YCkys/get-to-know-bigquery-including-sandbox-and-billing-options)

Big Query can be used to view and analyze data from tons of sources. Now we're going to explore the different accounts that Big Query offers, so you know how to choose the right one for your needs and how you can access them.

Big Query's offer to you at no charge. There are paid options available, but we won't need them for the activities in this course.

Instead, we're going to talk about two account types, sandbox and free trial. **A sandbox account is available at no charge and anyone with a Google account can log in and use it**. There are a couple of **limitations** to **this account type**. For example, you get a **maximum of 12 projects at a time**. This means that if you want to make a 13th project, you'll have to delete one of your original 12. It also doesn't allow you to insert new records to a database or update the field values of existing records. These Data Manipulation Language or DML operations aren't supported in the sandbox.

However, you won't need to do this in course activities, You can read more about the limitations of a sandbox account in the Big Query documentation. This is the account type we'll use for most of our activities. Before that though, we should talk about the other way to use Big Query without charges, the Google Cloud Free trial.

The free trial gives you access to more of what Big Query has to offer with fewer overall limitations.

The free trial offers 300 dollars in credit for use in Google Cloud during the first 90 days. You won't get anywhere near that spending limit if you just use the Big Query console to practice SQL queries.

After you spend the 300 dollars credit or after 90 days, your free trial will expire and you will need to personally select to upgrade to a paid account to keep working in Google Cloud. **Your method of payment will not be automatically charged after your free trial ends**. The free trial does require that you set up a payment option with Google Cloud, but unless you **choose to opt in for an account upgrade, it won't charge you**.

However, it does require you to enter a payment type, so we understand if you don't feel comfortable with this option.

This is one reason the Big Query sandbox account exists, so you don't have to enter any payment information.

With either type of account, you can upgrade to a paid account at any time and retain all of your existing projects.

If you set up a free trial account, but choose not to upgrade to a paid account when your trial period ends, you can set up a free sandbox account at that time.

[**SET UP YOUR BIGQUERY ACCOUNT**](https://www.coursera.org/learn/data-preparation/supplement/DYOQK/set-up-your-bigquery-account)

As you’ve been learning, BigQuery is a database you can use to access, explore, and analyze data from many sources. Now, you’ll begin using BigQuery, which will help you gain SQL knowledge by typing out commands and troubleshooting errors. This reading will guide you through the process of setting up your very own BigQuery account.

**Note:** Working with BigQuery is not a requirement of this program. Additional resources for other SQL database platforms are also provided at the end of this reading if you choose to use them instead.

## **BigQuery account options**

BigQuery offers a variety of account tiers to cater to various user needs and has two free-of-charge entry points, a sandbox account and a free-of-charge trial account. These options allow you to explore the program before selecting the best choice to suit your needs. A sandbox account allows you to practice writing queries and to explore public datasets free of charge, but it has [quotas and limits](https://cloud.google.com/bigquery/quotas), as well as some additional [restrictions](https://cloud.google.com/bigquery/docs/sandbox#limits).

If you prefer to use BigQuery with the standard limits, you can set up a free-of-charge trial account instead. The free-of-charge trial is a trial period prior to paying for a subscription. In this instance, there is no automatic charge, but you will be asked for payment information when you create the account.

This reading provides instructions for setting up either account type. An effective first step is to begin with a sandbox account and switch to a free-of-charge trial account when needed to run the SQL presented upcoming courses.

### **Sandbox account**

The sandbox account is available at no cost, and anyone with a Google account can use it. However, it does have some limitations. For instance, you are limited to a maximum of 12 projects at a time. This means that, to create a 13th project, you'll need to delete one of your existing 12 projects. Additionally, the sandbox account doesn't support all operations you’ll do in this program. For example, there are limits on the amount of data you can process and you can’t insert new records into a database or update the values of existing records. However, a sandbox account is perfect for most program activities, including all of the activities in this course. Additionally, you can convert your sandbox account into a free-of-charge trial account at any time.

**Set up your sandbox account**

To set up a sandbox account:

1. Visit the [BigQuery sandbox documentation](https://cloud.google.com/bigquery/docs/sandbox#limits) page.
2. Log in to your preferred Google account by selecting the profile icon in the BigQuery menu bar.
3. Select the **Go to BigQuery** button on the documentation page.
4. You'll be prompted to select your country and read the terms of service agreement.
5. This will bring you to the **SQL Workspace**, where you'll be conducting upcoming activities. By default, BigQuery creates a project for you.

After you set up your account, the name of the project will be in the banner in your BigQuery console.

### 

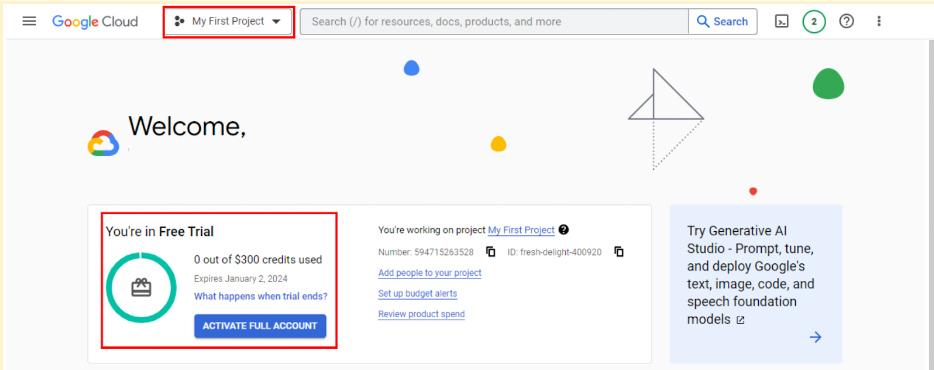
### **Free-of-charge trial**

If you wish to explore more of BigQuery's capabilities with fewer limitations, consider the Google Cloud Free Trial. It provides you with $300 in credit for Google Cloud usage during the first 90 days. If you're primarily using BigQuery for SQL queries, you're unlikely to come close to this spending limit. After you've used up the $300 credit or after 90 days, your free trial will expire, and you will only be able to use this account if you pay to do so. Google won't automatically charge your payment method when the trial ends. However, you'll need to set up a payment option with Google Cloud. This means that you’ll need to enter your financial information. Rest assured, it won't charge you unless you consciously opt to upgrade to a paid account. If you're uncomfortable providing payment information, don't worry; you can use the BigQuery sandbox account instead.

**Set up your free-of-charge trial**

1. Go to the [BigQuery](https://cloud.google.com/bigquery) page.
2. Select **Try BigQuery free**.
3. Log in using your Google email, or create an account free of charge if you don't have one. [Click here](https://cloud.google.com/bigquery?utm_source=google&utm_medium=cpc&utm_campaign=na-US-all-en-dr-bkws-all-all-trial-e-dr-1605212&utm_content=text-ad-none-any-DEV_c-CRE_665665924750-ADGP_Hybrid+%7C+BKWS+-+MIX+%7C+Txt_BigQuery-KWID_43700077225652770-kwd-274188433361&utm_term=KW_bigquery%20account-ST_bigquery+account&gclid=CjwKCAjwkNOpBhBEEiwAb3MvvYQXjIQ4TRnkITJoSXz7DFez4T-XKPG5IpfKmxUg2iHPEmiJBNQByhoCLVgQAvD_BwE&gclsrc=aw.ds) to create an account.
4. Select your country, a description of your organization or needs, and the checkbox to accept the terms of service, Then select **CONTINUE**.
5. Enter your billing information and select **START MY FREE TRIAL**.

After you set up your account, your first project, titled **My First Project** will be in the banner.



### **Transferring between BigQuery accounts**

With either a sandbox or free-of-charge trial account, you have the flexibility to upgrade to a paid account at any time. If you upgrade, all your existing projects will be retained and transferred to your new account. If you started with a free-of-charge trial, but choose not to upgrade when it ends, you can switch to a sandbox account. However, note that projects from your trial won't transfer to your sandbox. Essentially, creating a sandbox is like starting from scratch.

## **Get started with other databases (if not using BigQuery)**

It’s easiest to follow along with the course activities if you use BigQuery, but you may use other SQL platforms, if you prefer. If you decide to practice SQL queries on other database platforms, here are some resources to get started:

* [Getting Started with MySQL](https://dev.mysql.com/doc/mysql-getting-started/en/)
* [Getting Started with Microsoft SQL Server](https://docs.microsoft.com/en-us/sql/relational-databases/tutorial-getting-started-with-the-database-engine?view=sql-server-ver15)
* [Getting Started with PostgreSQL](https://www.postgresql.org/docs/10/tutorial-start.html)
* [Getting Started with SQLite](https://www.sqlite.org/quickstart.html)

## **Key takeaways**

BigQuery offers multiple account options. Keep the following in mind when you choose an account type:

* **Account tiers:** BigQuery provides various account tiers to cater to a wide range of user requirements. Whether you're starting with a sandbox account or exploring a paid account with the free-of-charge trial option, BigQuery offers flexibility to choose the option that aligns best with your needs and budget.
* **Sandbox limitations:** While a sandbox account is a great starting point, it comes with some limitations, such as a cap on the number of projects and restrictions on data manipulation operations like inserting or updating records, which you will encounter later in this program. Be aware of these limitations if you choose to work through this course using a sandbox account.
* **Easy setup and upgrades:** Getting started with any BigQuery account type is quick and easy. And if your needs evolve, you have the flexibility to modify your account status at any time. Additionally, projects can be retained even when transitioning between account types.

Choose the right BigQuery account type to match your specific needs and adapt as your requirements change!

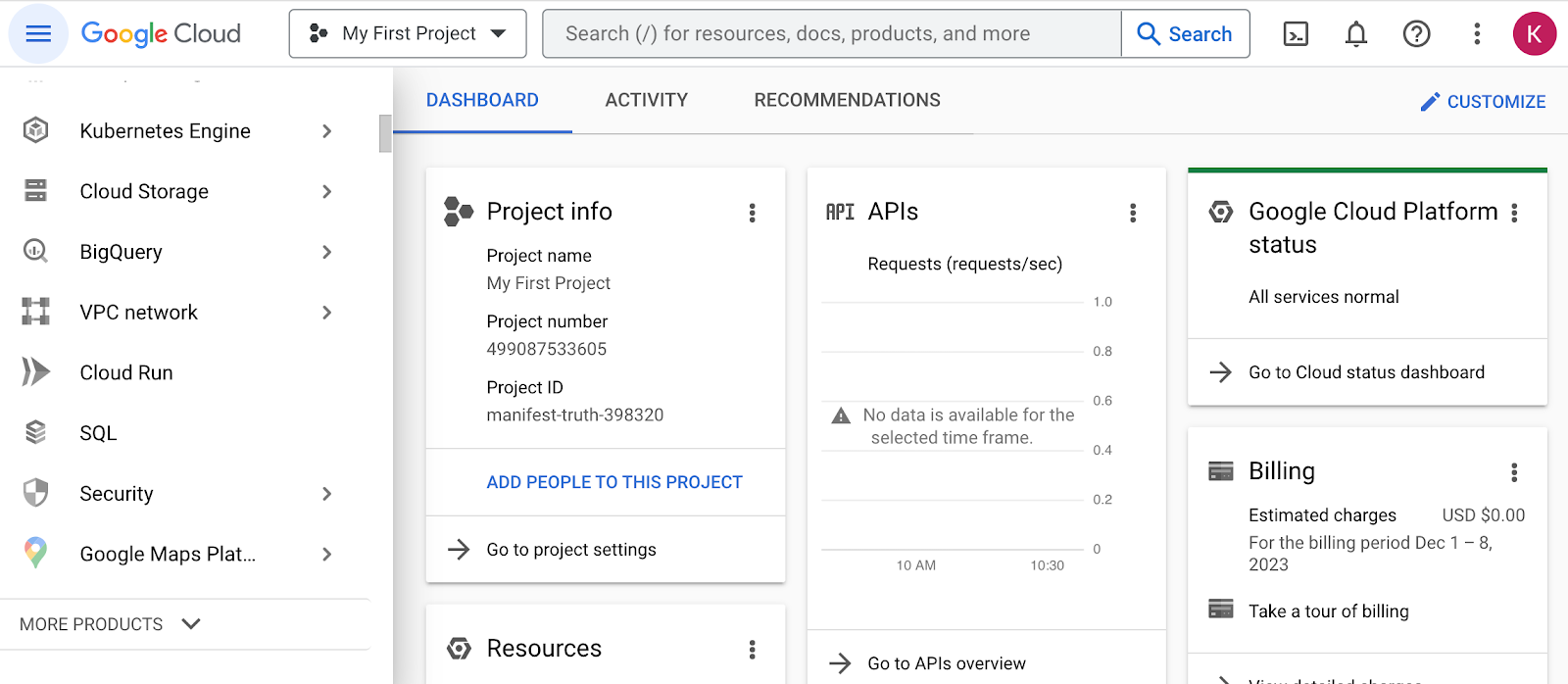
[**GET STARTED WITH BIGQUERY**](https://www.coursera.org/learn/data-preparation/supplement/7ctZ8/get-started-with-bigquery)

BigQuery is a data warehouse on the Google Cloud Platform used to query and filter large datasets, aggregate results, and perform complex operations. Throughout this program, you’re going to use BigQuery to practice your SQL skills and collect, prepare, and analyze data. At this point, you have set up your own account. Now, explore some of the important elements of the SQL workspace. This will prepare you for the upcoming activities in which you will use BigQuery. Note that BigQuery updates its interface frequently, so your console might be slightly different from what is described in this reading. That’s okay; use your troubleshooting skills to find what you need!

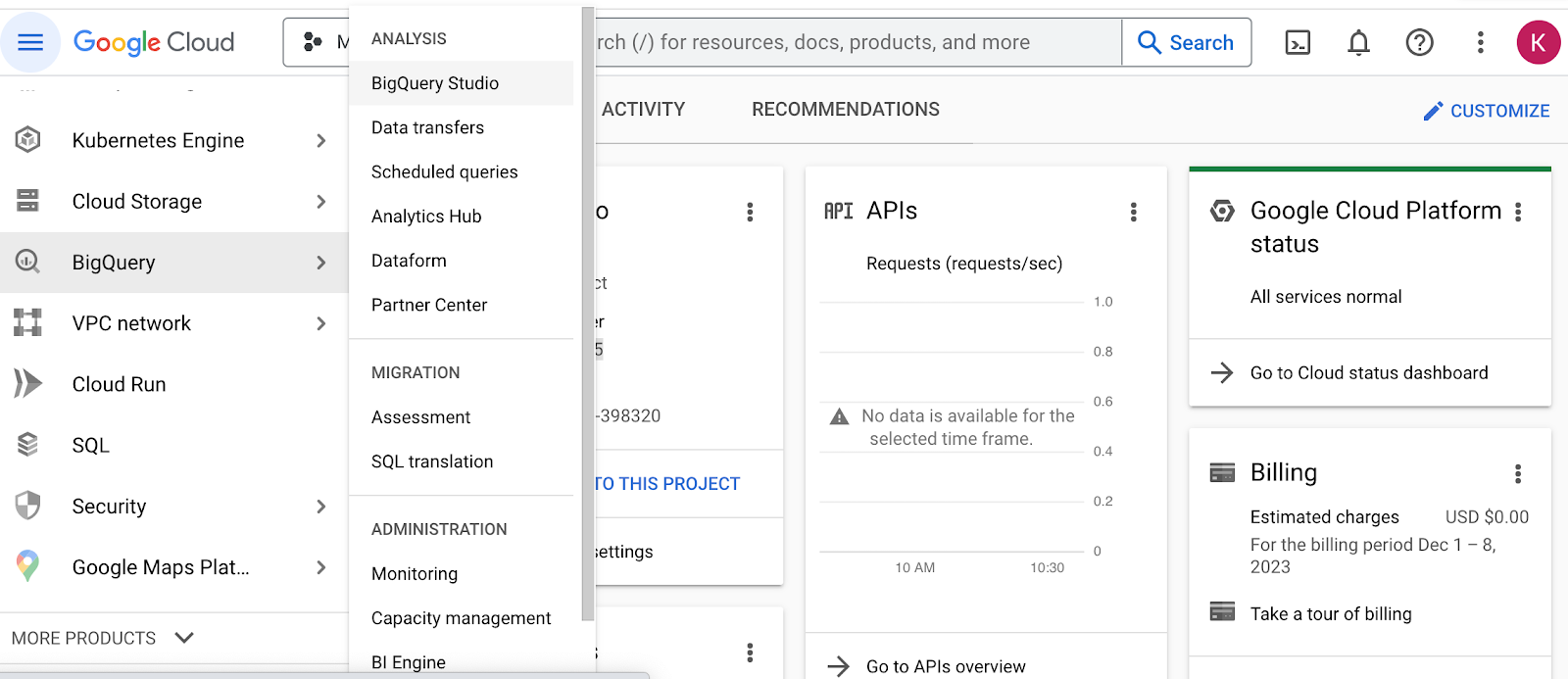
## 

## **Log in to BigQuery**

When you log in to BigQuery using the landing page, you will automatically open your project space. This is a high-level overview of your project, including the project information and the current resources being used. From here, you can check your recent activity.



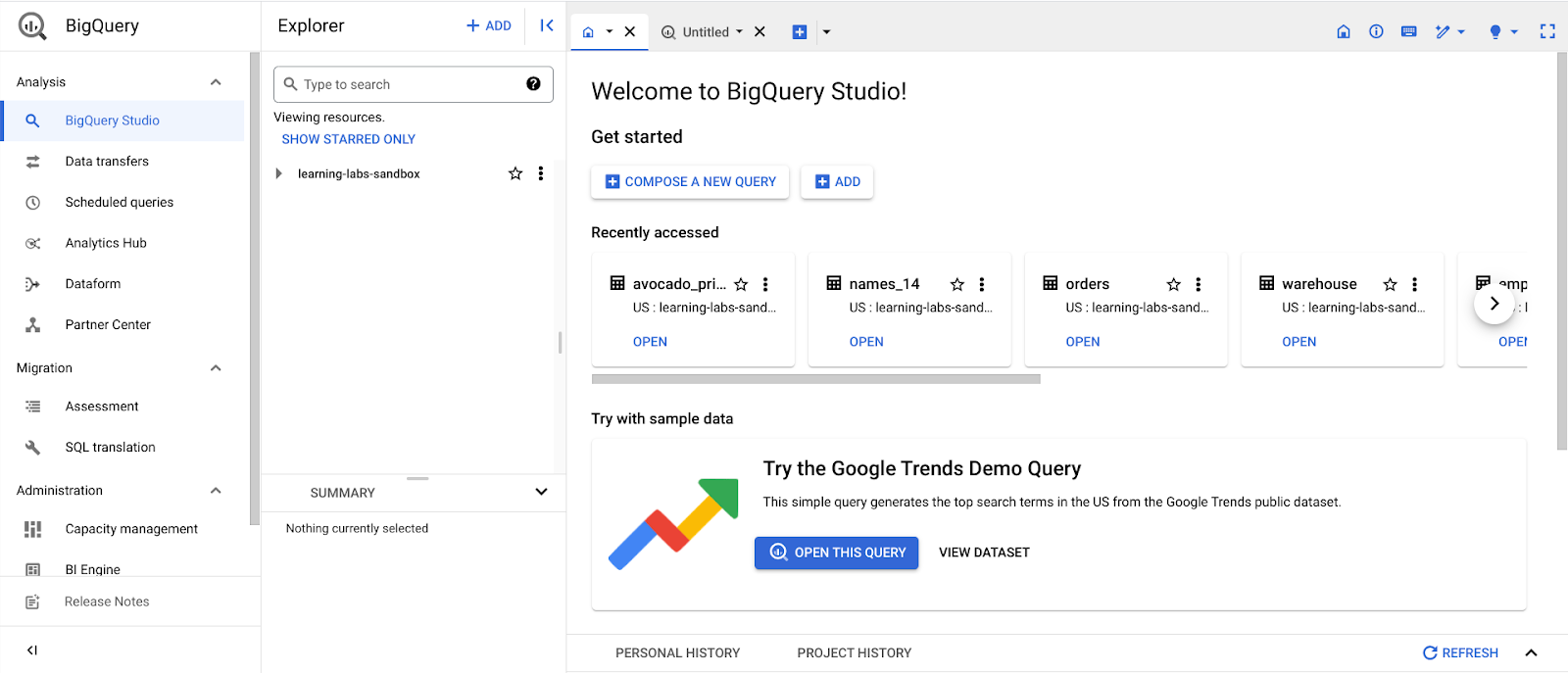
Navigate to your project’s BigQuery Studio by selecting BigQuery from the navigation menu and BigQuery Studio from the dropdown menu.



## 

## **BigQuery Studio components**

Once you have navigated to BigQuery from the project space, most of the major components of the BigQuery console will be present: the **Navigation** pane, the **Explorer** pane, and the **SQL Workspace**.



### 

### 

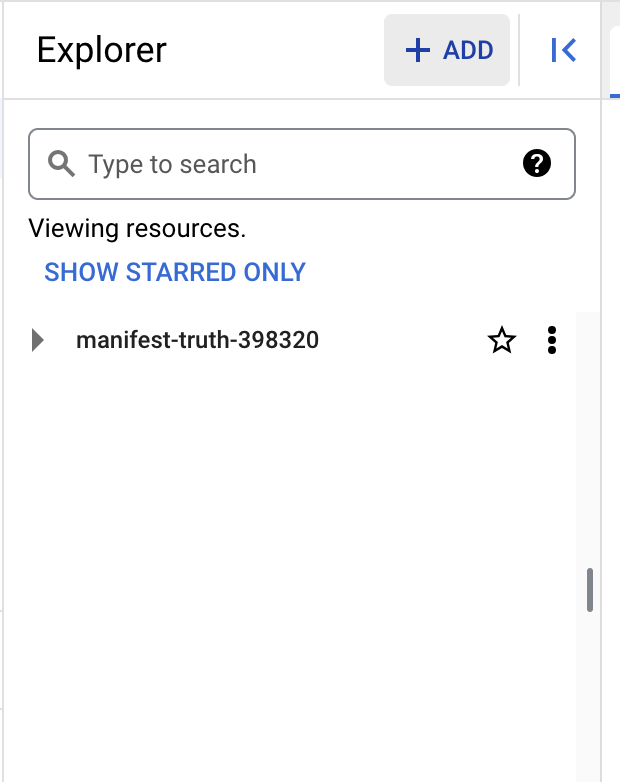
### 

### **The Navigation pane**

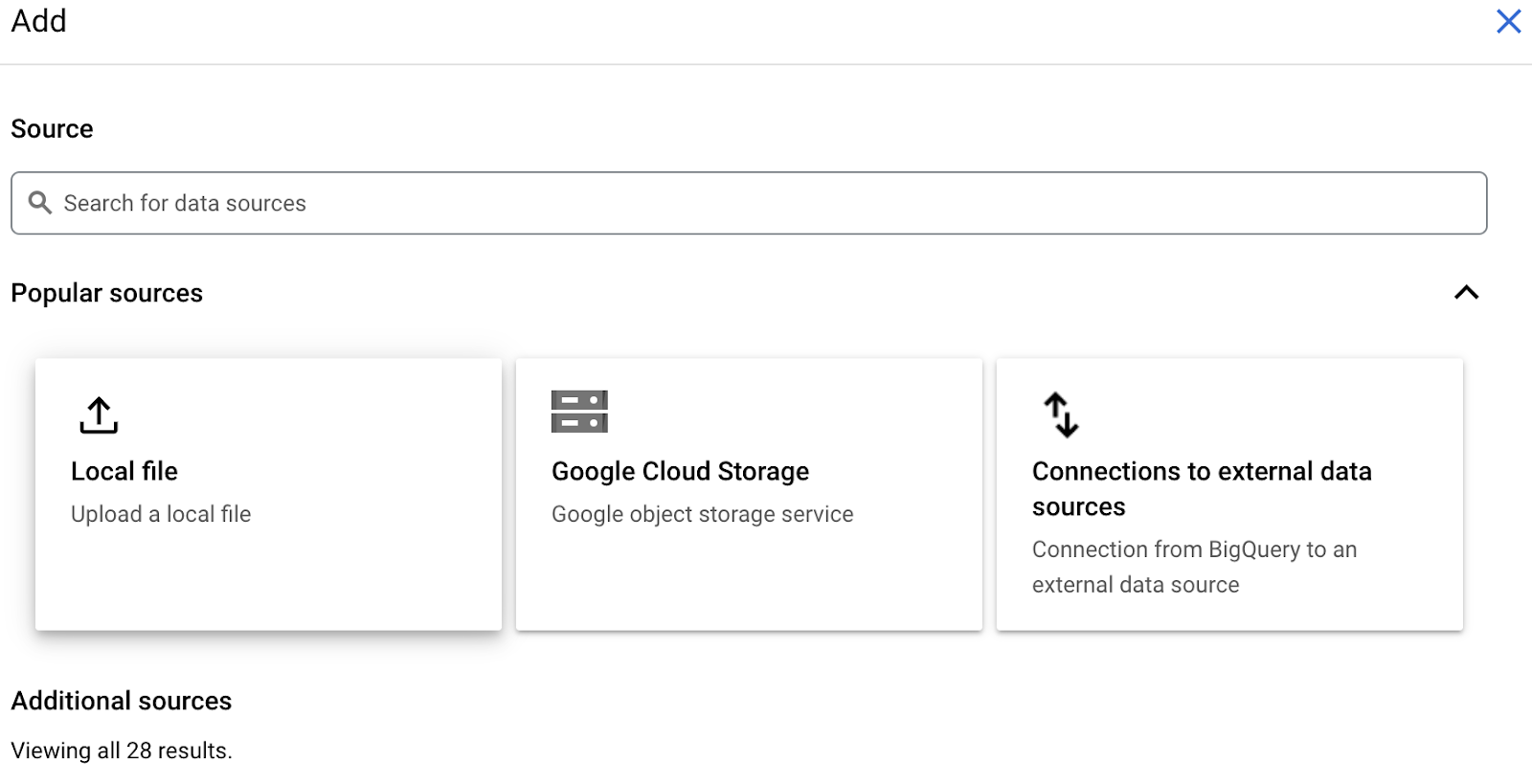
On the console page, find the **Navigation** pane. This is how you navigate from the project space to the BigQuery tool. This menu also contains a list of other Google Cloud Project (GCP) data tools. During this program, you will focus on BigQuery, but it’s useful to understand that the GCP has a collection of connected tools data professionals use every day.

### **The Explorer pane**

The **Explorer** pane lists your current projects and any starred projects you have added to your console. It’s also where you’ll find the **+ ADD** button, which you can use to add datasets.

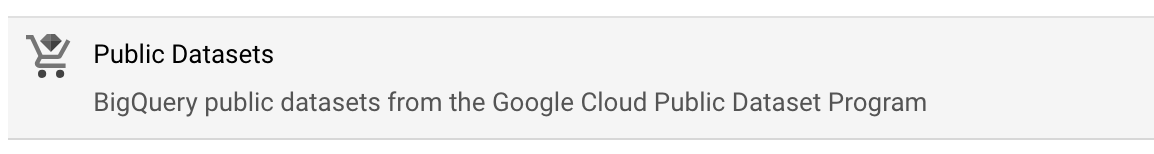


This button opens the **Add** dialog that allows you to open or import a variety of datasets.

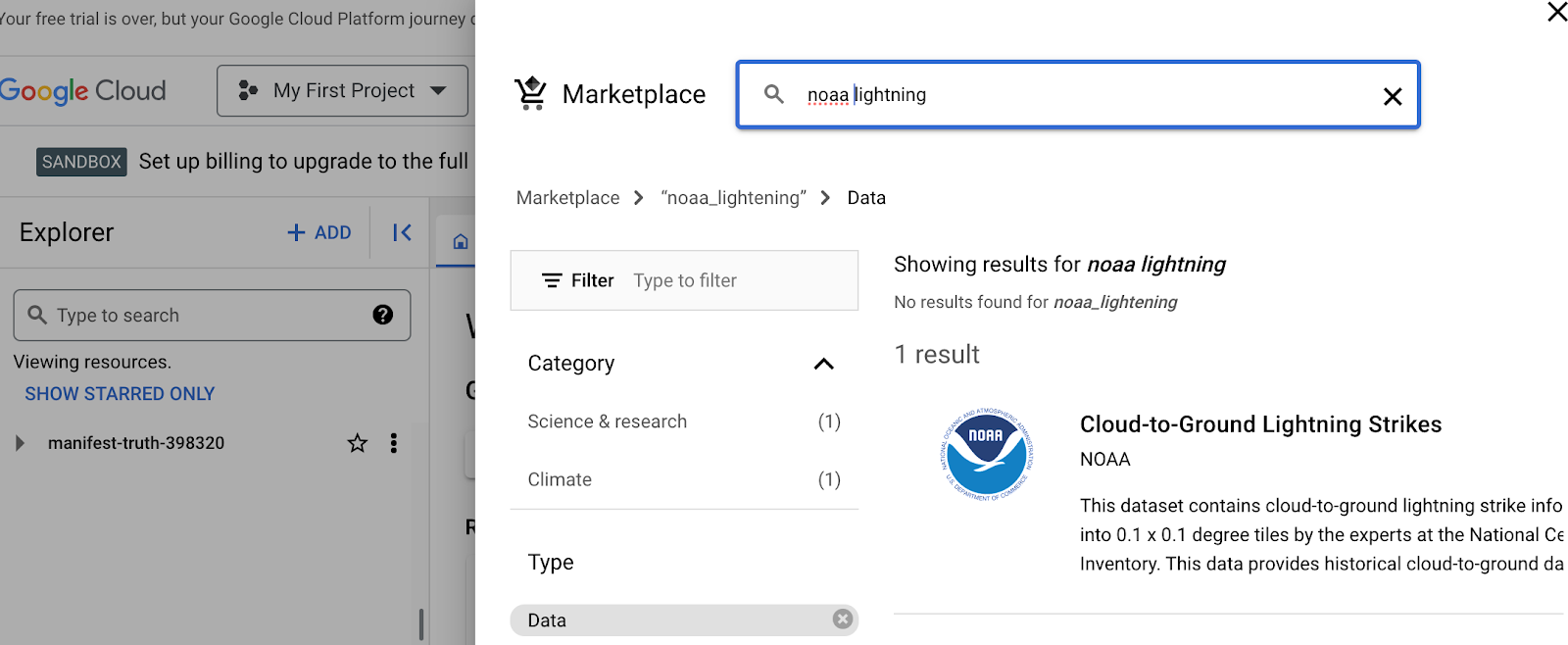


### **Add Public Datasets**

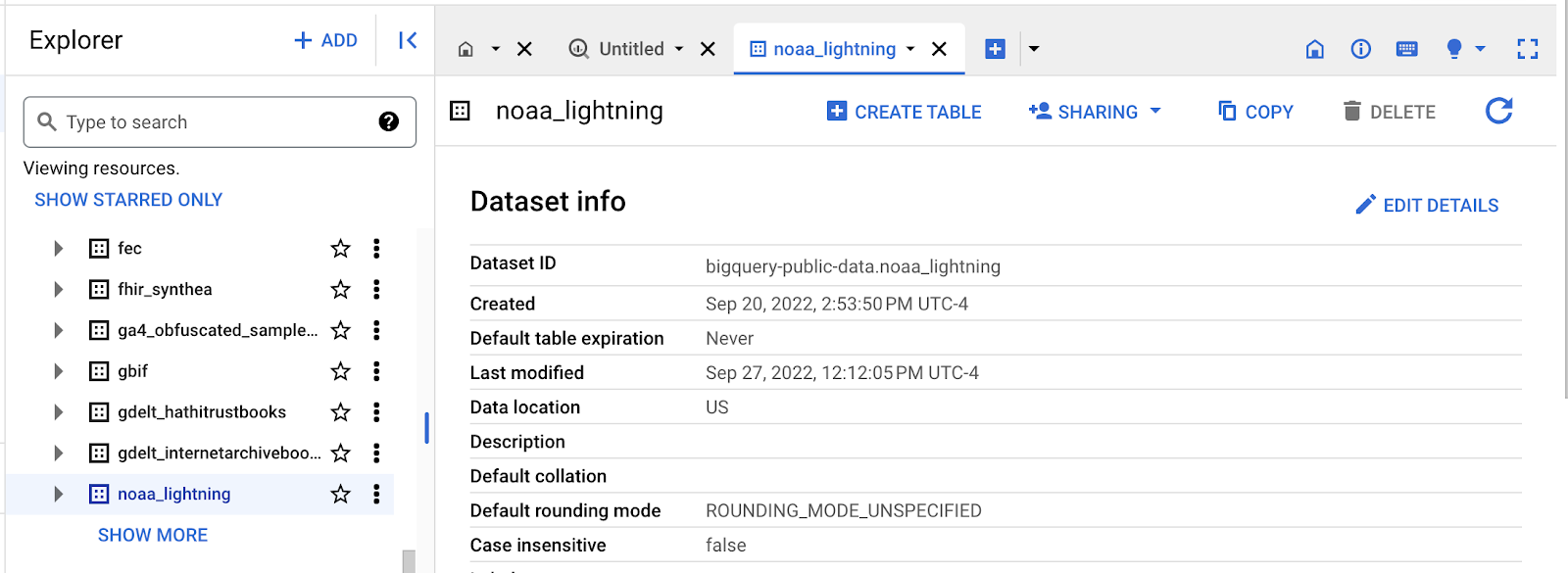
BigQuery offers a variety of public datasets from the Google Cloud Public Dataset Program. Scroll down the **Add** dialog to the **Public Datasets** option.



Select **Public Datasets**. This takes you to the **Public Datasets Marketplace**, where you can search for and select public datasets to add to your BigQuery console. For example, search for the "noaa lightning" dataset in the Marketplace search bar. When you search for this dataset, you will find NOAA’s Cloud-to-Ground Lightning Strikes data.



Select the dataset to read its description. Select **View dataset** to create a tab of the dataset’s information within the SQL workspace.

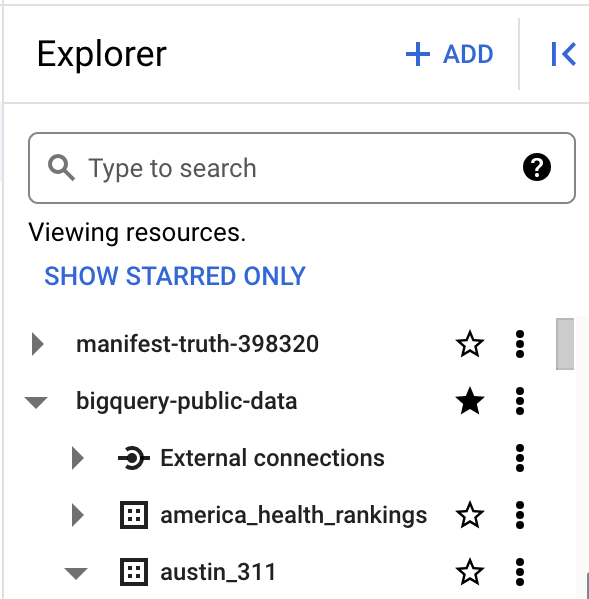


The Explorer Pane lists the noaa\_lightning and other public datasets.

### 

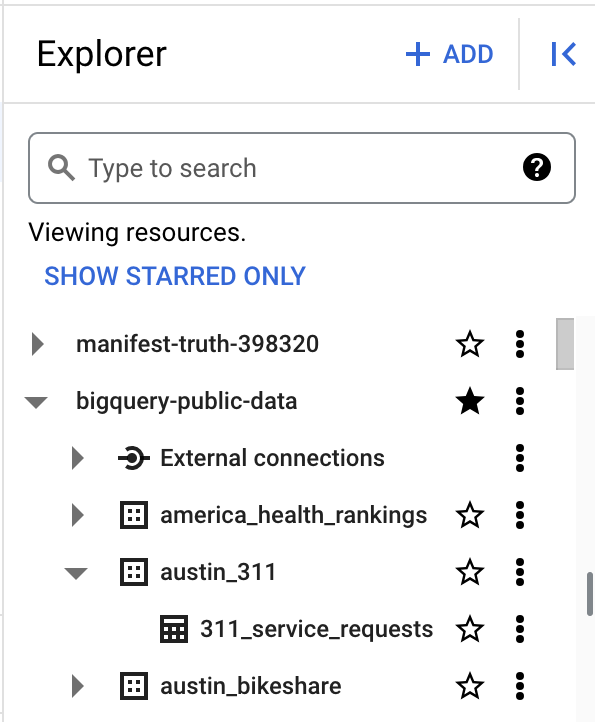
### **Star and examine Public Datasets**

You added the public noaa\_lightning dataset to your BigQuery Workspace, so the **Explorer** pane displays the noaa\_lightning dataset, along with the list of other public datasets. These datasets are nested under bigquery-public-data. Star bigquery-public-data by navigating to the top of the **Explorer** pane and selecting the star next to bigquery-public-data.



Starring bigquery-public-data will enable you to search for and add public datasets by scrolling in the **Explorer** pane or by searching for them in the **Explorer** search bar.

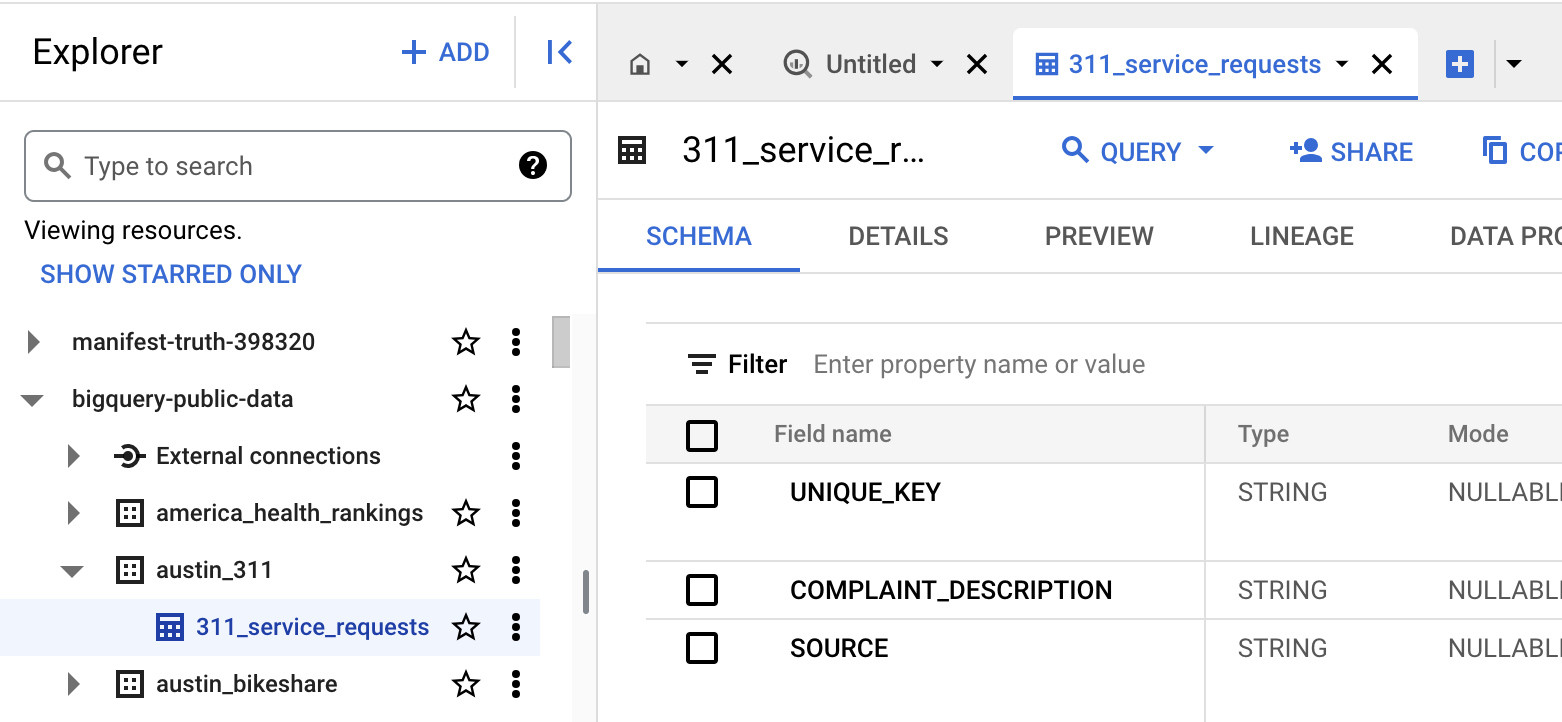
For example, you might want to select a different public dataset. If you select the second dataset, "austin\_311," it will expand to list the table stored in it, “311\_service\_requests.”



The Explorer pane with the “bigquery-public data” and “austin\_311” datasets expanded, revealing the “311\_service\_requests” table

When you select a table, its information is displayed in the SQL Workspace. Select the 311\_service\_requests table to examine several tabs that describe it, including:

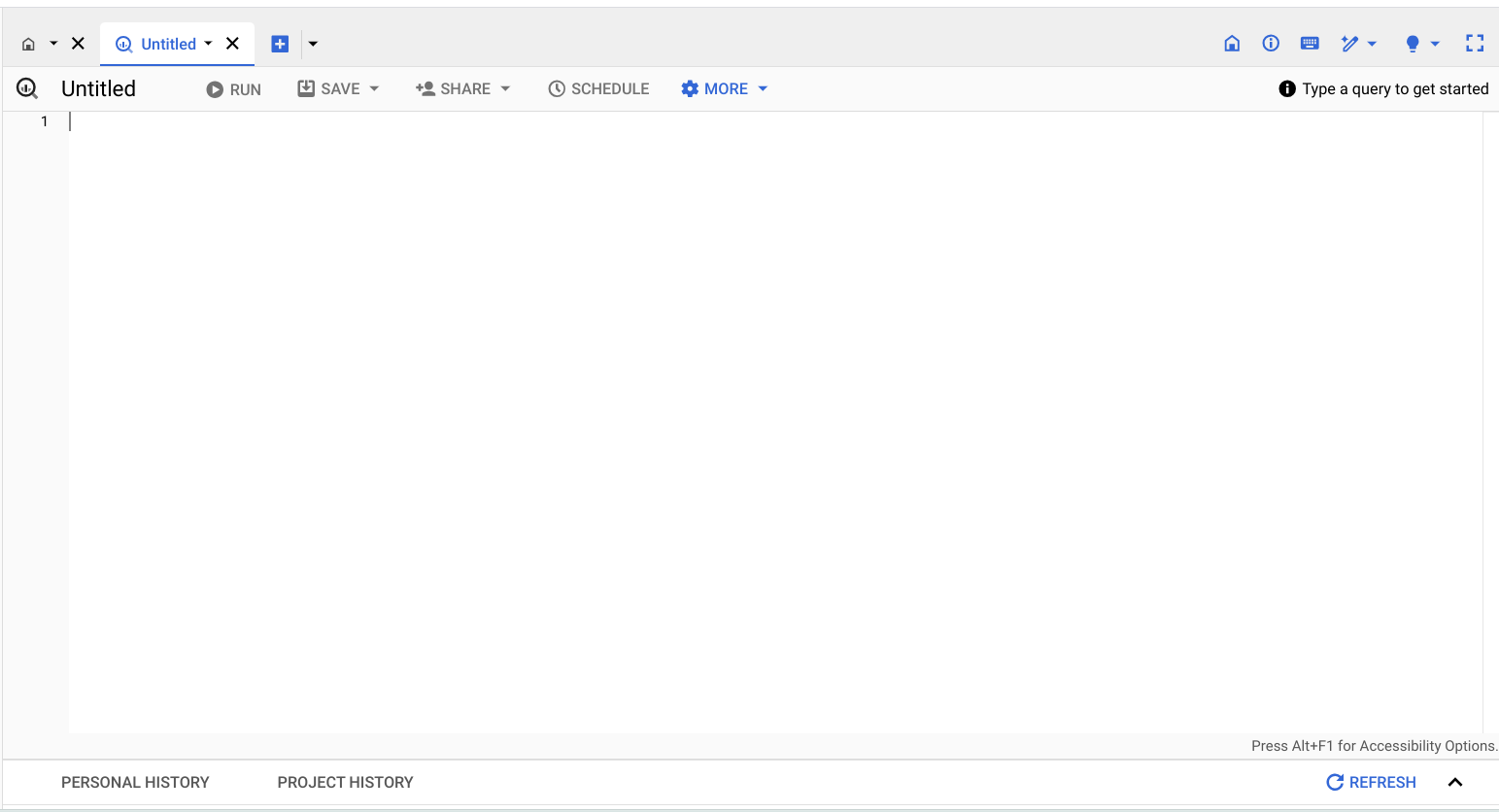
* **Schema**, which displays the column names in the dataset
* **Details**, which contains additional metadata, such as the creation date of the dataset
* **Preview**, which shows the first rows from the dataset



Additionally, you can select the **Query** button from the menu bar in the SQL Workspace to query this table.

### **The SQL Workspace**

The final menu pane in your console is the SQL Workspace. This is where you will actually write and execute queries in BigQuery.



The SQL Workspace also gives you access to your personal and project history, which stores a record of the queries you’ve run. This can be useful if you want to return to a query to run it again or use part of it in another query.

## **Upload your data**

In addition to offering access to public datasets, BigQuery also gives you the ability to upload your own data directly into your workspace. Access this feature by opening the **+ ADD** menu again or by clicking the three vertical dots next to your project’s name in the Explorer pane. This will give you the option to create your own dataset and upload your own tables. You will have the opportunity to upload your own data in an upcoming activity to practice using this feature!

## **Key takeaways**

BigQuery's SQL workspace allows you to search for public datasets, run SQL queries, and even upload your own data for analysis. Whether you're working with public datasets, running SQL queries, or uploading your own data, BigQuery’s SQL workspace offers a range of features to support all kinds of data analysis tasks. Throughout this program, you will be using BigQuery to practice your SQL skills, so being familiar with the major components of your BigQuery console will help you navigate it effectively in the future!

[**STEP-BY-STEP: BIGQUERY IN ACTION**](https://www.coursera.org/learn/data-preparation/supplement/PZ6Sv/step-by-step-bigquery-in-action)

This reading provides you with the steps the instructor performs in the following video, [BigQuery in action](https://www.coursera.org/learn/data-preparation/lecture/H877e/bigquery-in-action). The video focuses on how to create a query to view a small section of data from a large dataset.

Keep this guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

## **What you'll need**

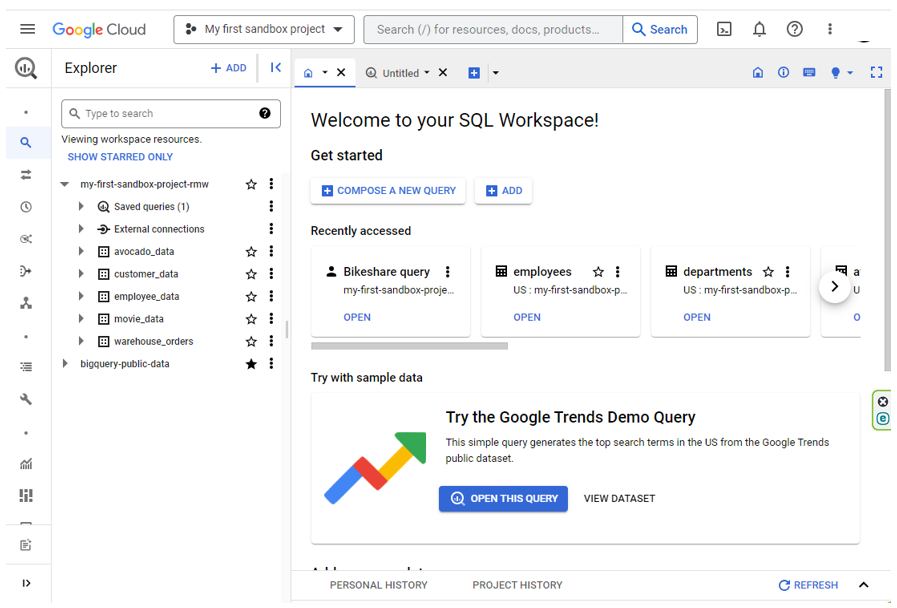
To follow along with the examples in this video, log in to your BigQuery account and follow the instructions to star **bigquery-public-data** in **The Explorer pane** section of the previous reading, [Get Started with BigQuery](https://www.coursera.org/learn/data-preparation/supplement/7ctZ8/get-started-with-bigquery).

Empty alt text.

## **Example 1: Preview a section from a table viewer**

A database is a collection of data stored in a computer system. Query languages such as SQL enable communication between databases and data analysts. You discovered earlier that a relational database is made up of several tables that may be joined together to create relationships. Primary and foreign keys serve as representations of these relationships. To extract data from these tables, data analysts use queries. To learn more about that, explore BigQuery in action:

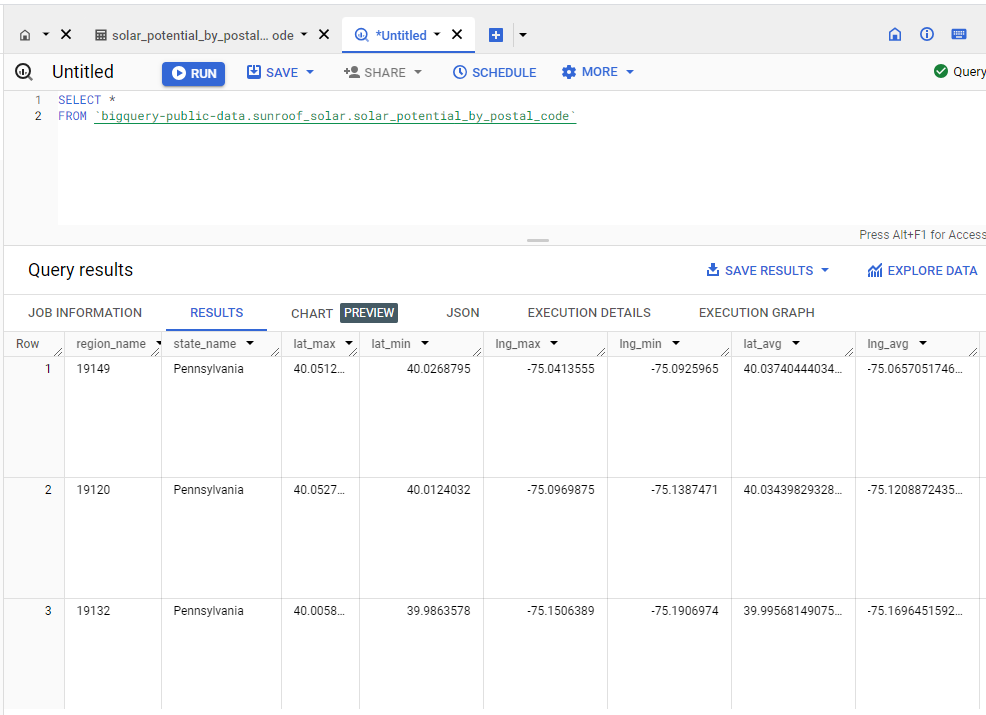
1. Log in to [BigQuery](https://console.cloud.google.com/bigquery) and go to your console. You should find the **Welcome to your SQL Workspace!** landing page open. Select **COMPOSE A NEW QUERY** In the Bigquery console. Make sure that no tabs are open so that the entire workspace is displayed, including the **Explorer** pane.
2. Enter **sunroof** in the search bar. In the search results, expand **sunroof\_solar** and then select the **solar\_potential\_by\_postal\_code** dataset.
3. Observe the **Schema tab** of the **Explorer** pane to explore the table fields.
4. Select the **Preview** tab to view the regions, states, yearly sunlight, and more.



## **Example 2: Writing a query**

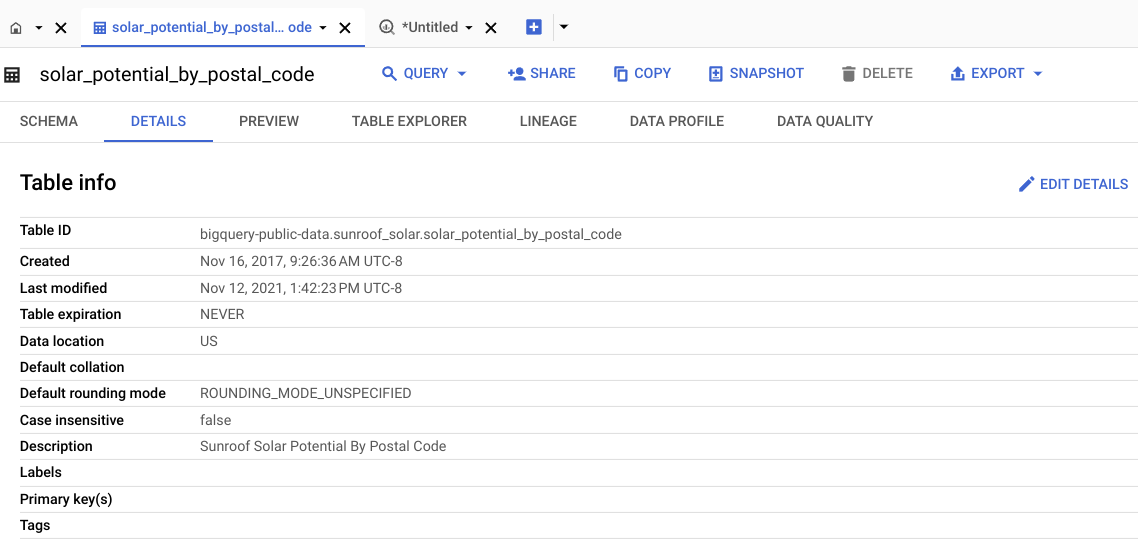
In order to view the entire dataset, you will need to write a query.

1. The first step is finding out the complete, correct path to the table you want to work with. Select the **ellipses** (three vertical dots) by the dataset **solar\_potential\_by\_postal\_code**, then select **Query**. A new tab will populate on your screen. Select the tab. The path to the table should be written inside two backticks.
2. Select the full path by highlighting the text including the backticks and copy it. (**Note:** You can also get the full path to the project, database, and table directly by clicking the ellipses next to the table's name in the **Explorer** panel on the left and selecting **Copy ID**.)
3. Now, click on the **plus sign** to create a new query. Notice that BigQuery doesn’t automatically generate a **SELECT** statement in this window. Enter **SELECT** and add a space after it.
4. Put an asterisk **\*** after **SELECT** to indicate you want to return the entire dataset. The asterisk lets the database know to include all columns. Without this shortcut, you would have to manually enter every column name!
5. Next, press the **Enter/Return** key and enter **FROM** on the second line. **FROM** indicates where the data is coming from. After **FROM**, add another space.
6. Paste in the path to the table that you copied earlier. It will read **`bigquery-public-data.sunroof\_solar.solar\_potential\_by\_postal\_code`**
7. Execute the query by selecting the **RUN** button.



**Important!**

Many of the public databases on BigQuery are living records and, as such, are periodically updated with new data. Throughout this course (and others in this certificate program), if your results differ from those you encounter in videos or screenshots, there's a good chance it is due to a data refresh. You can verify when a table has been refreshed by selecting it from the **Explorer** panel and clicking **Details**. You'll find the date the table was created, when it was last modified, as well as other useful information.



## **Example 3: Use SQL to view a piece of data**

If the project doesn’t require every field to be completed, you can use SQL to see a particular piece, or pieces, of data. To do this, specify a certain column name in the query.

1. For example, you might only need data from Pennsylvania. You’d begin your query the same way you just did in the previous examples: Click on the **plus sign**, enter **SELECT**, add a space, an asterisk (**\***), and then press **Enter/Return**.
2. Enter **FROM** and then paste **`bigquery-public-data.sunroof\_solar.solar\_potential\_by\_postal\_code`**. Press **Enter/Return**.
3. This time, add **WHERE**. It will be on the same line as the **FROM** statement. Add a space and enter **state\_name** with a space before state and a space after name. **state\_name** is a column name in the table.
4. Because you only want data from Pennsylvania, add **=** and **'Pennsylvania'** on the same line as **state\_name**. In SQL, single quotes represent the beginning and ending of a string.
5. Execute the query with the **RUN** button.
6. Review the data on solar potential for Pennsylvania. Scroll through the query results.

**Example 3 Code:**

1 SELECT \*

2 FROM `bigquery-public-data.sunroof\_solar.solar\_potential\_by\_postal\_code` WHERE

3 state\_name = 'New Jersey'

Keep in mind that SQL queries can be written in a lot of different ways and still return the same results. You might discover other ways to write these queries!

[**BIGQUERY IN ACTION**](https://www.coursera.org/learn/data-preparation/lecture/H877e/bigquery-in-action)

It’s just the video doing the same actions as the previous topic.

[**HANDS-ON ACTIVITY: INTRODUCTION TO BIGQUERY**](https://www.coursera.org/learn/data-preparation/quiz/OsfsT/hands-on-activity-introduction-to-bigquery)

## **Activity overview**

You have recently been introduced to BigQuery, a data warehouse on Google Cloud that data analysts use to query, filter large datasets, aggregate results, and perform complex operations. In this activity, you will explore the BigQuery interface; upload public data to your console; and write some simple SQL queries using **SELECT**, **FROM**, and **WHERE**.

By the time you complete this activity, you will be more familiar with writing queries in the BigQuery interface. This will enable you to practice SQL, which is important for working with databases in your career as a data analyst.



### 

### **Step-By-Step Instructions**

Follow the instructions to complete each step of the activity. Then answer the questions at the end of the activity before going to the next course item.

### 

### **Step 1: Get a BigQuery account**

For this activity, you will need a BigQuery account. If you haven’t made one already, follow the instructions from the [Using BigQuery](https://www.coursera.org/learn/data-preparation/supplement/DYOQK/using-bigquery) reading.

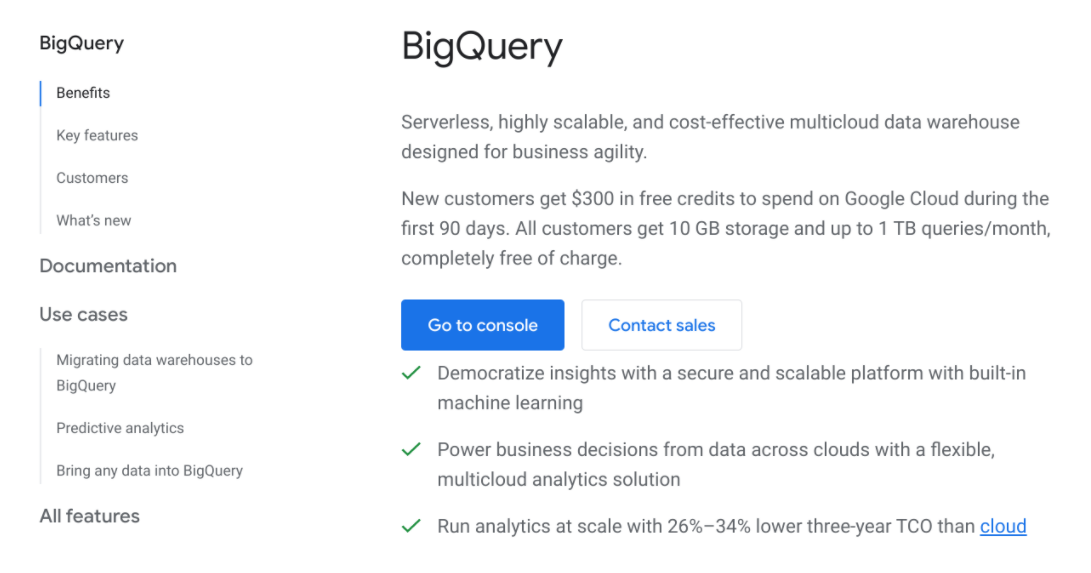
Once you have your account, start exploring!

### 

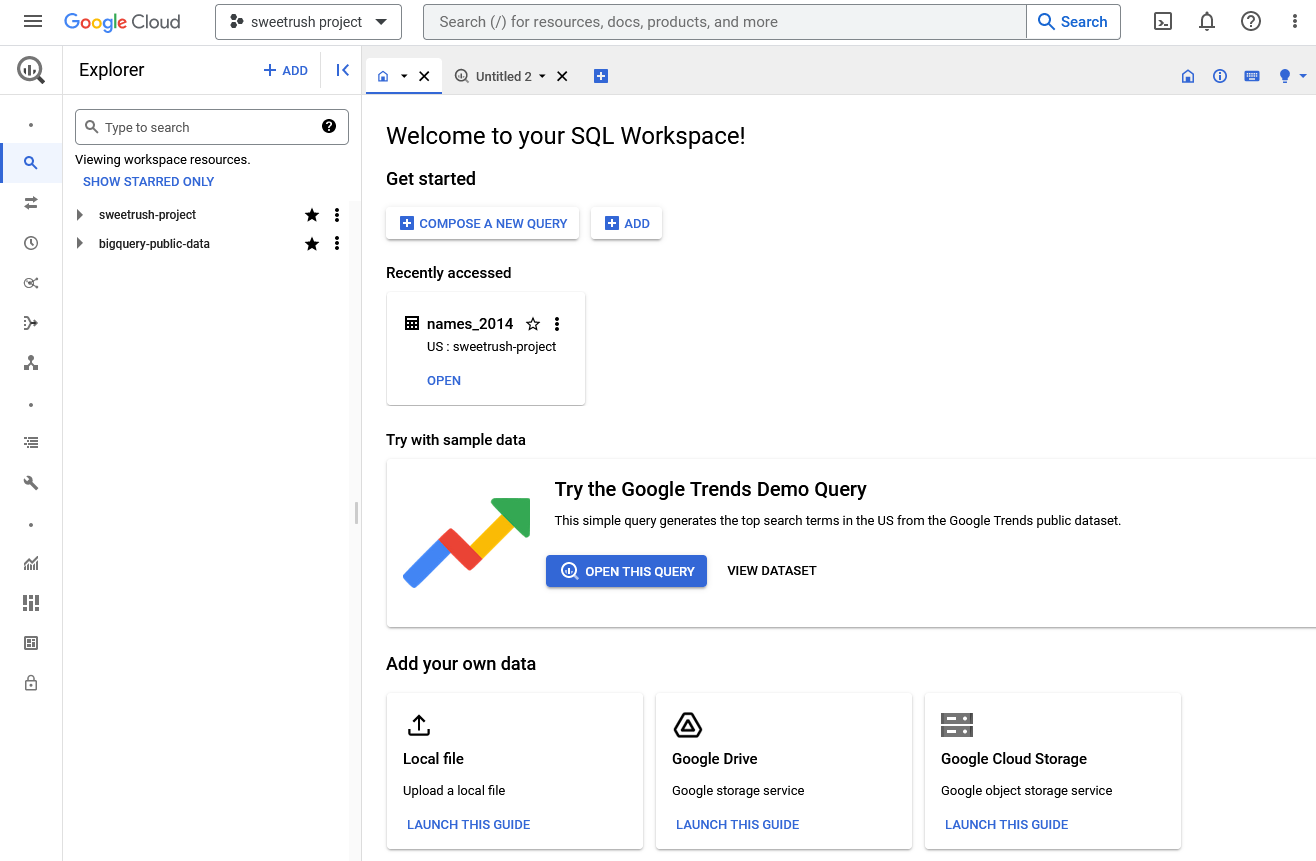
### **Step 2: Open your BigQuery console**

1. Log in to [BigQuery](https://cloud.google.com/bigquery).

2. Select the Go to console button on the BigQuery homepage. This will open a new tab with your console.



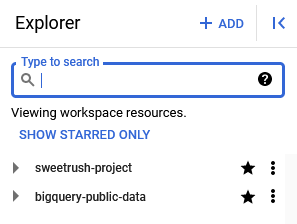
3. Take a moment to explore your console. The Explorer menu includes a search bar you can use to find resources, pinned projects, and the + ADD button for adding data. The Editor welcome page is where you will navigate to a query editor, try sample data, add local data files, add Google cloud storage, or add other external connections. You can also find your job history, query history, and saved queries here.



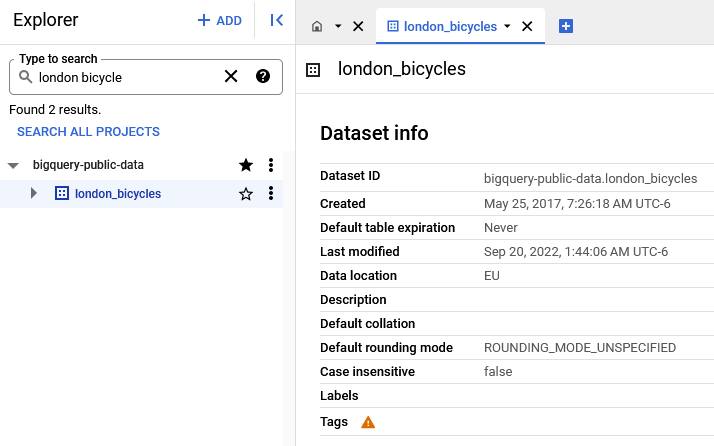
### **Step 3: Access public data in BigQuery**

In order to start writing queries, you will need some data to work with. Once you’re familiar with the BigQuery interface, you can access a public dataset directly from your console.

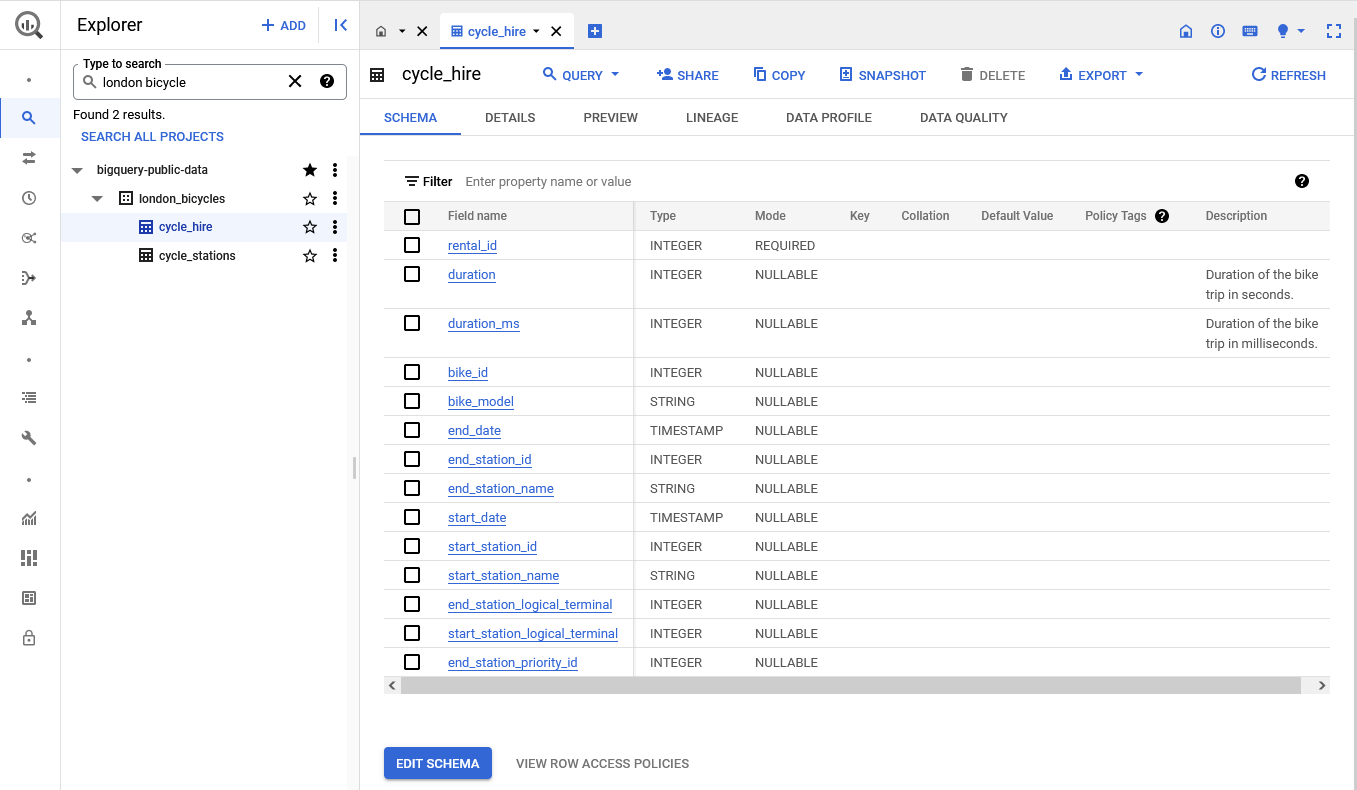
1. Select the search bar in the Explorer pane.



2. Enter “london bicycle” in the search box and press enter; this will return the **london\_bicycles** database from the Greater London Authority. Select the database for more details. If you cannot find it, make sure you're searching in all projects. The **london\_bicycles** database is in the **bigquery-public-data** project.



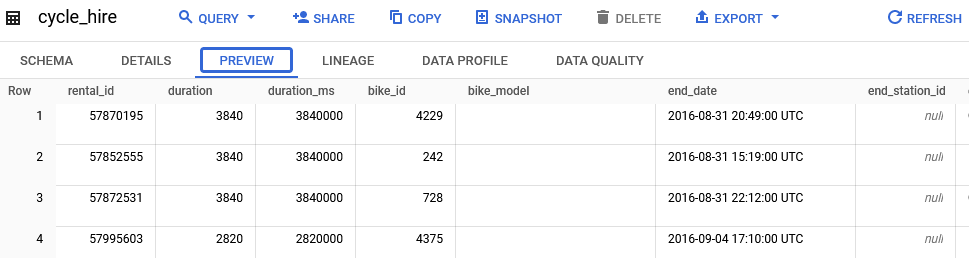
3. Select the arrow to the left of the **london\_bicycles** database name. This expands the dataset to reveal two table names: **cycle\_hire** and **cycle\_stations**. Select the **cycle\_hire** table name within the Explorer pane.



Screenshot of BigQuery. The titles cycle\_hire and cycle\_stations are shown beneath london\_bicycles in the explorer pane. Cycle\_hire is selected.

This will pull the **cycle\_hire** schema into the console. Take a moment to explore the field names and the associated information.

4. Now, select the PREVIEW tab to find a sample of the data that you’ll be working with.



Once you have finished previewing the data, write a query!

### 

### **Step 4: Review basic parts of a query**



So far, you’ve learned three basic parts of a query: **SELECT**, **FROM**, and **WHERE**. As a refresher:

* **SELECT** is the section of a query that indicates what data you want SQL to return to you.
* **FROM** is the section of a query that indicates which table the desired data comes from. You must provide a full path to the table. The path includes the project name, database name, and table name, each separated by a period.
* **WHERE** is the section of a query that indicates any filters you’d like to apply to your table.

### 

### **Step 5: Write a basic query**

Now, construct a simple command using the basic parts of a query you have already learned! For example, you can select a specific column from the **cycle\_hire** table, such as the **end\_station\_name** column.

1. Select the Blue + button or QUERY - In the new tab to start a new query.

2. Start your query with a **SELECT** clause, and indicate which column you want to select from the table; in this case, you’ll input **end\_station\_name**.

3. After you have indicated which column you are selecting, write your **FROM** clause. Specify the table you are querying from by inputting the following location: **`bigquery-public-data.london\_bicycles.cycle\_hire`;**

The completed query should appear like this:

1 SELECT

2 end\_station\_name

3 FROM

4 `bigquery-public-data.london\_bicycles.cycle\_hire`;

4. Run your completed query by selecting the blue RUN button.

This query may take a few seconds to execute. Once it has finished, you will find the list of station names you requested under the Query Results console pane.

### 

### **Step 6: Write a query to answer a question**

After running the first basic query, try answering a specific question about the data. For example, how many bike trips lasted for 20 minutes or longer?

1. Select the Blue + button or QUERY - In the new tab to start a new query. Start with your **SELECT** statement again. This time, include the two columns **duration**and **start\_station\_name** in the query. The data in these columns will tell where the trip started and the duration of the trip. Be sure to separate each column name with a comma.

2. Next, add your **FROM** statement. You will be using the same table as the previous query: **FROM `bigquery-public-data.london\_bicycles.cycle\_hire`;**

3. Finally, add a **WHERE** statement to specify that you want to filter for only bike rides 20 minutes or longer. If you check the preview of this data, you might notice that the **duration** is recorded in seconds, so you’ll specify 1200 seconds in your query. Write that as **WHERE duration >= 1200;**

Your completed query should be written like this:

1 SELECT

2 duration,

3 start\_station\_name

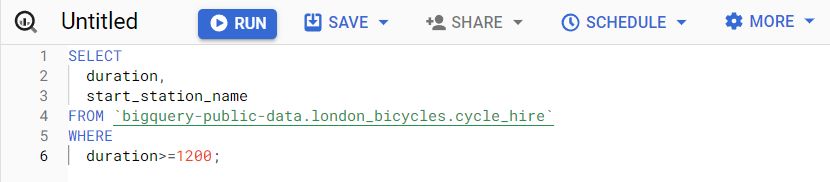
4 FROM

5 `bigquery-public-data.london\_bicycles.cycle\_hire`

6 WHERE

7 duration >= 1200;

4. Run your completed query by clicking the RUN button.



This query may take a few seconds to execute. Once it has finished, you will find a list of rides from this table that fit your criteria. There are millions of rows with bike trips that are 20 minutes or longer!

### 

### Optional Step 7: Up for a challenge?

If you’re comfortable using queries to answer questions, try creating and running queries to complete the tasks below:

* What is the name of the station whose **start\_station\_id** is 111?
* Return all the **rental\_id**s, station IDs, and station names that **bike\_id** 1710 started from.
* What is the **bike\_model** of **bike\_id** 58782?

### 

### Step 8: Check your work

Use the solutions doc to check your work: [Intro to BigQuery Solutions](https://docs.google.com/document/d/1Rw8gXT0E4Smo4huoOcahX5ZQqV_pV8zgES8Oltatr-Y/template/preview)

Or download the file directly here:

[Intro to BigQuery solutions](https://d3c33hcgiwev3.cloudfront.net/9KElwgWaT2yuZqs_AcEbpQ_a6e50e18806a49979215819d1a80f8f1_Intro-to-BigQuery-solutions.docx?Expires=1711584000&Signature=fZFXoqlSAmiTNTAShaZyNiZ-21h3opmlo~M8V1RDui5THXBZgSfElCjFdKKndC4FHvLAOzOgEjVZJKae5zV4hAhxZafS4QqETg-WpCmBlIYo4R88L0N682ZVUft5Ddn-P-DGaUOs8auLl6Yo7QsujImdK8xK24VF5qtRgyOtGnQ_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[DOCX File](https://d3c33hcgiwev3.cloudfront.net/9KElwgWaT2yuZqs_AcEbpQ_a6e50e18806a49979215819d1a80f8f1_Intro-to-BigQuery-solutions.docx?Expires=1711584000&Signature=fZFXoqlSAmiTNTAShaZyNiZ-21h3opmlo~M8V1RDui5THXBZgSfElCjFdKKndC4FHvLAOzOgEjVZJKae5zV4hAhxZafS4QqETg-WpCmBlIYo4R88L0N682ZVUft5Ddn-P-DGaUOs8auLl6Yo7QsujImdK8xK24VF5qtRgyOtGnQ_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

Save this reading for future reference. Feel free to download a .pdf version of this reading below:

[DAC3-In-depth-guide\_-SQL-best-practices.pdf](https://d3c33hcgiwev3.cloudfront.net/UwaGyGQoRLu9Dw_8BLUTiQ_4d31c4c09dc54520835973c4ac8240f1_DAC3-In-depth-guide_-SQL-best-practices.pdf?Expires=1711584000&Signature=hwAtZfNWqgLUVXU48jZQh29nZY86t~YwzMx5-0DEzpayla8L0r58YnrlFudn-4m0woZBLsdMloV~mk5qAJt6aU09~0S9PqFD8Os3rGpX2k0Xa0Ru3O3OadqYstQjoqu7WEbhgTGNl7nYWzR-5aRkcd6ATJpLWdEGUoiQp4Buaw4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[PDF File](https://d3c33hcgiwev3.cloudfront.net/UwaGyGQoRLu9Dw_8BLUTiQ_4d31c4c09dc54520835973c4ac8240f1_DAC3-In-depth-guide_-SQL-best-practices.pdf?Expires=1711584000&Signature=hwAtZfNWqgLUVXU48jZQh29nZY86t~YwzMx5-0DEzpayla8L0r58YnrlFudn-4m0woZBLsdMloV~mk5qAJt6aU09~0S9PqFD8Os3rGpX2k0Xa0Ru3O3OadqYstQjoqu7WEbhgTGNl7nYWzR-5aRkcd6ATJpLWdEGUoiQp4Buaw4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[**IN DEPTH GUIDE: SQL BEST PRACTICES**](https://www.coursera.org/learn/data-preparation/supplement/G5WAN/in-depth-guide-sql-best-practices)

These best practices include guidelines for entering SQL queries, developing documentation, and examples that demonstrate these practices. This is a great resource to have handy when you are using SQL yourself; you can just go straight to the relevant section to review these practices. Think of it like a SQL field guide!

## **Capitalization and case sensitivity**

With SQL, capitalization usually doesn’t matter. You could enter **SELECT** or select or **SeLeCT**. They all work!But if you **use capitalization as part of a consistent style your queries will look more professional.**

To enter SQL queries like a pro, it is always a good idea to use all caps for clause starters (e.g. **SELECT**, **FROM**, **WHERE**, etc.). Functions should also be in all caps (e.g. **SUM()**).

**Column names** should be all lowercase (refer to the section on **snake\_case** later in this guide).

Table names should be in CamelCase (refer to the section on CamelCase later in this guide).

This **helps keep your queries consistent and easier to read** while not impacting the data that will be pulled when you run them. The only time that capitalization does matter is when it is inside quotes (more on quotes below).

Vendors of SQL databases may use slightly different variations of SQL. These variations are called **SQL dialects**. Some SQL dialects are **case sensitive**. **BigQuery is** one of them.

Vertica is another. But most, like MySQL, PostgreSQL, and SQL Server, aren’t case sensitive. This means if you searched for **country\_code = ‘us’**, it will return all entries that have **'us'**, **'uS'**, **'Us'**, and **'US'**.

This isn’t the case with BigQuery. **BigQuery is case sensitive**, so that same search would only return entries where the **country\_code** is exactly **'us'**. If the **country\_code** is **'US'**, BigQuery wouldn’t return those entries as part of your result.

## 

## **Single or double quotes: ' ' or " "**

For the most part, it also doesn’t matter if you use single quotes **' '** or double quotes **" "** when referring to strings. For example, **SELECT** is a clause starter. If you put **SELECT** in quotes like **'SELECT'** or **"SELECT"**, then SQL will treat it as a text string. Your query will return an error because your query needs a **SELECT** clause.

But there are two situations where it does matter what kind of quotes you use:

1. When you want strings to be identifiable in *any* SQL dialect
2. When your string contains an apostrophe or quotation marks

Within each SQL dialect there are rules for what is accepted and what isn’t.

But a general rule across almost all SQL dialects is to **use single quotes for strings**. This helps get rid of a lot of confusion.

So if we want to reference the country US in a **WHERE** clause (e.g. **country\_code = 'US'**), then use single quotes around the string **'US'**.

The second situation is when your string has quotes inside it. Suppose you have a column **favorite\_food** in a table called **FavoriteFoods** and the other column corresponds to each **friend**.

| **friend** | **favorite\_food** |
| --- | --- |
| Rachel DeSantos | Shepherd’s pie |
| Sujin Lee | Tacos |
| Najil Okoro | Spanish paella |

You might notice how Rachel’s favorite food contains an apostrophe. If you were to use single quotes in a **WHERE** clause to find the friend who has this favorite food, it would look like this:

1 SELECT

2 friend

3 FROM

4 FavoriteFoods

5 WHERE

6 favorite\_food = 'Shepherd's pie'

**This won’t work.** If you run this query, you will get an error in return. This is because SQL recognizes a text string as something that starts with a quote **'** and ends with another quote **'**. So in the bad query above, SQL thinks that the **favorite\_food** you are looking for is '**Shepherd'**, because the apostrophe in Shepherd**'**s ends the string.

Generally speaking, this should be the only time you would use double quotes instead of single quotes. So your query would look like this instead:

1 SELECT

friend

2 FROM

FavoriteFoods

3 WHERE

favorite\_food = "Shepherd's pie"

SQL understands text strings as either starting with a single quote **'** or double quote **"**. Since this string starts with double quotes, SQL will expect another double quote to signal the end of the string. This keeps the apostrophe safe, so it will return "Shepherd's pie" and not 'Shepherd'.

## **Comments as reminders**

As you get more comfortable with SQL, you will be able to read and understand queries at a glance. But it never hurts to have comments in the query to remind yourself of what you are trying to do. And if you share your query, it also helps others understand it.

For example:

4

5

6

7

Info.date, -date is in spring format YYYY-MM-DD HH:MM:SS

Info.code -e.g. 'pub-###'

FROM

Publishers

You can use # in place of the two dashes, **--**, in the above query but keep in mind that # isn’t recognized in all SQL dialects (MySQL doesn’t recognize **#**). So it is best to use **--** and be consistent with it. When you add a comment to a query using **--**, the database query engine will ignore everything in the same line after **--**. It will continue to process the query starting on the next line.

## **snake\_case names for columns**

It is important to always make sure that the output of your query has easy-to-understand names. If you create a new column (say from a calculation or from concatenating new fields), the new column will receive a generic default name (e.g. **f0**). For example:

2

3

4

5

6

7

SUM(tickets),

COUNT(tickets),

SUM(tickets) AS total\_tickets,

COUNT(tickets) AS number\_of\_purchases

FROM

Purchases

Results are:

| **f0** | **f1** | **total\_tickets** | **number\_of\_purchases** |
| --- | --- | --- | --- |
| 8 | 4 | 8 | 4 |

The first two columns are named **f0** and **f1** because they weren’t named in the above query. SQL defaults to **f0**, **f1**, **f2**, **f3**, and so on. We named the last two columns **total\_tickets** and **number\_of\_purchases** so these column names show up in the query results. This is why it is always good to give your columns useful names, especially when using functions. After running your query, you want to be able to quickly understand your results, like the last two columns we described in the example.

On top of that, you might notice how the column names have an underscore between the words. Names should never have spaces in them. If **total\_tickets** had a space and looked like **total tickets** then SQL would throw a syntax error because it wouldn't know what to do with the second word (**tickets**). So, spaces are bad in SQL names. Never use spaces.

The best practice is to use snake\_case. This means that 'total tickets', which has a space between the two words, should be entered as **total\_tickets** with an underscore instead of a space.

## **CamelCase names for tables**

You can also use CamelCase capitalization when naming your table. CamelCase capitalization means that you capitalize the start of each word, like a two-humped (Bactrian) camel. So the table **TicketsByOccasion** uses CamelCase capitalization. Please note that the capitalization of the first word in CamelCase is *optional;* camelCase is also used. Some people differentiate between the two styles by calling CamelCase,PascalCase, and reserving camelCase for when the first word isn't capitalized, like a one-humped (Dromedary) camel; for example, **ticketsByOccasion**.

At the end of the day, CamelCase is a style choice. There are other ways you can name your tables, including:

* All lower or upper case, like **ticketsbyoccasion** or **TICKETSBYOCCASION**

With snake\_case, like t**ickets\_by\_occasion**

Keep in mind, the option with all lowercase or uppercase letters can make it difficult to read your table name, so it isn’t recommended for professional use.

The second option, snake\_case, is technically okay. With words separated by underscores, your table name is easy to read, but it can get very long because you are adding the underscores. It also takes more time to enter. If you use this table a lot, it can become a chore.

In summary, it is up to you to use snake\_case or CamelCase when creating table names. Just make sure your table name is easy to read and consistent. Also be sure to find out if your company has a preferred way of naming their tables. If they do, always go with their naming convention for consistency.

## **Indentation**

As a general rule, you want to keep the length of each line in a query <= 100 characters. This makes your queries easy to read. For example, check out this query with a line with >100 characters:

4

5

6

7

8

watch\_category, COUNT(movie\_title) AS number\_of\_movies

FROM

MovieTheater

GROUP BY

1

This query is hard to read and just as hard to troubleshoot or edit. Now, here is a query where we stick to the <= 100 character rule:

10

1

Now it is much easier to understand what you are trying to do in the **SELECT** clause. Sure, both queries will run without a problem because indentation doesn’t matter in SQL. But proper indentation is still important to keep lines short. And it will be valued by anyone reading your query, including yourself!

## **Multi-line comments**

If you make comments that take up multiple lines, you can use **--** for each line. Or, if you have more than two lines of comments, it might be cleaner and easier is to use **/\*** to start the comment and **\*/** to close the comment. For example, you can use the -- method like below:

7

table

Or, you can use the /\* \*/ method like below:

8

table

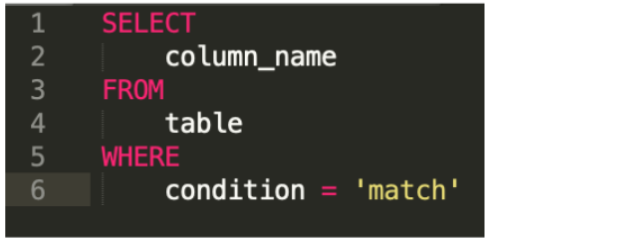
In SQL, it doesn’t matter which method you use. SQL ignores comments regardless of what you use: **#**, **--**, or **/\*** and **\*/**. So it is up to you and your personal preference. The **/\*** and **\*/** method for multi-line comments usually looks cleaner and helps separate the comments from the query. But there isn’t one right or wrong method.

## **SQL text editors**

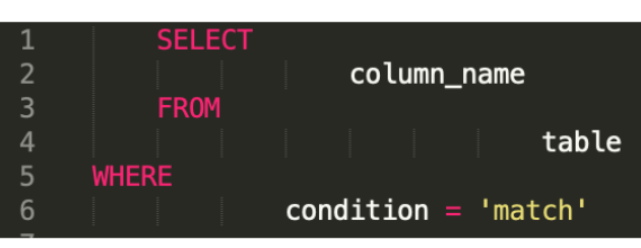
When you join a company, you can expect each company to use their own SQL platform and SQL dialect. The SQL platform they use (e.g. BigQuery, MySQL, or SQL Server) is where you will enter and run your SQL queries. But keep in mind that not all SQL platforms provide native script editors to enter SQL code. SQL text editors give you an interface where you can enter your SQL queries in an easier and color-coded way. In fact, all of the code we have been working with so far was entered with an SQL text editor!

## **Examples with Sublime Text**

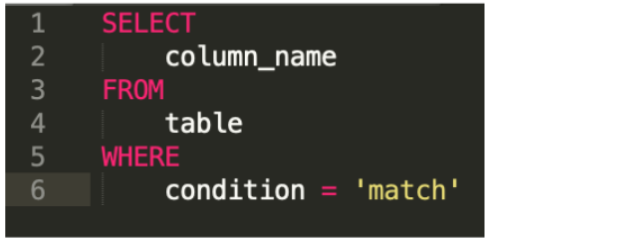
If your SQL platform doesn’t have color coding, you might want to think about using a text editor like [Sublime Text](https://www.sublimetext.com/) or [Atom](https://atom.io/). This section shows how SQL is displayed in Sublime Text. Here is a query in Sublime Text:



With Sublime Text, you can also do advanced editing like deleting indents across multiple lines at the same time. For example, suppose your query somehow had indents in the wrong places and looked like this:



This is really hard to read, so you will want to eliminate those indents and start over. In a regular SQL platform, you would have to go into each line and press BACKSPACE to delete each indent per line. But in Sublime, you can get rid of all the indents at the same time by selecting all lines and pressing Command (or CTRL in Windows) + [. This eliminates indents from every line. Then you can select the lines that you want to indent (i.e., lines 2, 4, and 6) by pressing the Command key (or the CTRL key in Windows) and selecting those lines. Then while still holding down the Command key (or the CTRL key in Windows), press ] to indent lines 2, 4, and 6 at the same time. This will clean up your query and make it look like this instead:



Sublime Text also supports regular expressions. **Regular expressions** (or **regex**) can be used to search for and replace string patterns in queries. We won’t cover regular expressions here, but you might want to learn more about them on your own because they are a very powerful tool.

You can begin with these resources:

* [Search and replace in Sublime Text](https://sublime-text-unofficial-documentation.readthedocs.io/en/latest/search_and_replace/search_and_replace_overview.html)
* [Regex tutorial](https://www.regular-expressions.info/tutorialcnt.html) (if you don’t know what regular expressions are)
* [Regex cheat sheet](https://jdhao.github.io/2019/02/28/sublime_text_regex_cheat_sheet/)

[**HANDS-ON ACTIVITY: CHOOSE THE RIGHT TOOL FOR THE JOB**](https://www.coursera.org/learn/data-preparation/quiz/KThkt/hands-on-activity-choose-the-right-tool-for-the-job)

[**HANDS-ON ACTIVITY: MORE PRACTICE WITH SQL**](https://www.coursera.org/learn/data-preparation/quiz/vGN91/hands-on-activity-more-practice-with-sql)

[**TEST YOUR KNOWLEDGE ON USING SQL WITH LARGE DATASETS**](https://www.coursera.org/learn/data-preparation/quiz/3UyRx/test-your-knowledge-on-using-sql-with-large-datasets)

a

**MODULE 3 CHALLENGE**

[**GLOSSARY TERMS FROM MODULE 3**](https://www.coursera.org/learn/data-preparation/supplement/MOjTP/glossary-terms-from-module-3)

a

**MODULE 4- ORGANIZE AND PROTECT DATA**

**BRING DATA TO ORDER**

[**CHALLENGE**](https://www.coursera.org/learn/data-preparation/exam/jZOOG/module-2-challenge)

**SECURE DATA**

**MODULE 4 CHALLENGE**

**MODULE 5- ENGAGE IN DATA COMMUNITY**

**CREATE OR ENHANCE YOUR ONLINE PRESENCE**

[**CHALLENGE**](https://www.coursera.org/learn/data-preparation/exam/jZOOG/module-2-challenge)

**BUILD A DATA ANALYTICS NETWORK**

**COURSE WRAP-UP**

**Stakeholders** are people that have invested time, interest, and resources into the projects that you'll be working on as a data analyst. In other words, **they hold stakes in what you're doing.** There's a good chance they'll need the work you do to perform their own needs.