Right now data is being generated all around

the world and we're talking tons of data.

Every minute of every day millions of

texts and hundreds of millions of emails are sent.

On top of that, millions of online searches are

made and videos viewed

and those numbers are only growing.

That's a lot of data.

Let's learn more about how it's made and used.

In this video, we'll talk about the ways that data can be

generated and how industries collect data themselves.

Every piece of information is data.

All that data is usually

generated as a result of our activity in the world.

These days, we spend a lot of time online.

With social media and mobile devices,

millions and millions of people are adding to

the huge amount of data out there, each and every day.

Think about it like this.

Every digital photo online is one piece of data.

Every photo itself holds even more data,

from the number of pixels to

the colors contained in each of those pixels.

But that's not the only way data is made.

We can also generate data by collecting information.

This data generation and

collection comes with a few more things to think about.

It needs to be done with consideration to

ethics so that we maintain people's rights and privacy.

We'll learn more about that later on.

For now, let's check out a real world example.

The United States Census Bureau uses

forms to collect data about the country's population.

This data is used for a number of reasons,

like funding for schools,

hospitals, and fire departments.

The Bureau also collects

information about things like U.S. businesses,

creating their own data in the process.

The great thing about this is that others can then use

the data for their own needs, including analysis.

The annual business survey

is used to figure out the needs of

businesses and how to provide

them with resources to help them succeed.

I actually generate data in

the analytics I do for the health care industry.

We run a lot of surveys to learn how patients

feel about certain things related to their health care.

For example, one survey asked how patients feel

about telemedicine versus in-person doctor visits.

The data we collected help the companies we work

with improve the care that their patients receive.

Survey data is just one example.

There's all kinds of data being generated all the

time, and there's lots of different ways to collect it.

Even something as simple as

an interview can help someone collect data.

Imagine you're in a job interview.

To impress the hiring manager,

you want to share information about yourself.

The hiring manager collects that data and

analyzes it to help them decide

whether to hire you or not.

But it goes both ways.

You could also collect your own data about

the company to help you decide if

the company is a good fit for you.

Or you can use the data you collect to come up

with thoughtful questions to ask the interviewer.

Scientists also generate data.

They use a lot of observations in their work.

For example, they might collect data by studying

animal behavior or looking

at bacteria under a microscope.

Earlier we talked about the forms that

the U.S. Census Bureau uses to collect data.

Forms, questionnaires and surveys are

commonly used ways to collect and generate data.

One thing to note: data that's generated

online doesn't always happen directly.

Have you ever wondered why some online ads seem to make

really accurate suggestions or

how some websites remember your preferences?

This is done using cookies,

which are small files stored on

computers that contain information about users.

Cookies can help inform advertisers about

your personal interests and

habits based on your online surfing,

without personally identifying you.

As a real world analyst,

you'll have all kinds of data right at

your fingertips and lots of it too.

Knowing how it's been generated can

help add context to the data,

and knowing how to collect it can make

the data analysis process more efficient.

Coming up, you'll learn how to decide

what data to collect for your analysis.

So stay tuned.

Welcome back. We've talked a lot

about all the data out there in the world.

But as a data analyst,

you'll need to decide what kind of data to

collect and use for every project.

With a nearly endless amount of data out there,

this can be quite a bit of

a data dilemma, but there's good news.

In this video, you'll learn

which factors to consider when collecting data.

Usually, you'll have a head start in

figuring out the right data for the job,

because the data you need will be given to you, or

your business task or problem

will narrow down your choices.

Let's start with a question like,

what's causing increased rush hour traffic in your city?

First, you need to know how the data will be collected.

You might use observations of traffic patterns to count

the number of cars on city streets

during particular times.

You notice that cars are getting

backed up on a specific street.

That brings us to data sources.

In our traffic example,

your observations would be first-party data.

This is data collected by

an individual or group using their own resources.

Collecting first-party data is typically

the preferred method because you

know exactly where it came from.

You might also have second-party data,

which is data collected by a group

directly from its audience and then sold.

In our example, if you aren't

able to collect your own data,

you might buy it from an organization that's

led traffic pattern studies in your city.

This data didn't start with you,

but it's still reliable because it came from

a source that has experience with traffic analysis.

The same can't always be said about third-party data or

data collected from outside

sources who did not collect it directly.

This data might have come from a number of

different sources before you investigated it.

It might not be as reliable,

but that doesn't mean it can't be useful.

You'll just want to make sure you check it for

accuracy, bias, and credibility.

Actually, no matter what kind of data you use,

it needs to be inspected

for accuracy and trustworthiness.

We'll learn more about that process later.

For now, just remember that the data you choose should

apply to your needs, and it must be approved for use.

As a data analyst,

it's your job to decide what data to use,

and that means choosing the data

that can help you find answers and

solve problems and not getting distracted by other data.

In our traffic example,

financial data probably wouldn't be that helpful,

but existing data about

high volume traffic times would be.

Okay. Now let's talk about how much data to collect.

In data analytics, a population refers to

all possible data values in a certain data set.

If you're analyzing data about car traffic in a city,

your population would be all the cars in that area.

But collecting data from

the entire population can be pretty challenging.

That's why a sample can be useful.

A sample is a part of a population

that is representative of the population.

You might collect a data sample about one spot

in the city and analyze the traffic there,

or you might pull a random sample from

all existing data in the population.

How you choose your sample will depend on your project.

As you collect data, you'll also want to

make sure you select the right data type.

For traffic data, an appropriate data type could

be the dates of traffic records stored in a date format.

The dates could help you figure

what days of the week there is

likely to be a high volume of traffic in the future.

We'll explore this topic in more detail soon.

Finally, you need to determine

the time frame for data collection.

In our example, if you needed an answer immediately,

you'd have to use historical data,

which is data that already exists.

But let's say you needed to track

traffic patterns over a long period of time.

That might affect the other decisions

you make during data collection.

Now you know more about

the different data collection considerations

you'll use as a data analyst.

Because of that, you'll be able to find

the right data when you start collecting it yourself.

There's still more to learn about

data collection, so stay tuned.

Selecting the right data

Following are some data-collection considerations to keep in mind for your analysis:

**How the data will be collected**

Decide if you will collect the data using your own resources or receive (and possibly purchase it) from another party. Data that you collect yourself is called first-party data.

**Data sources**

If you don’t collect the data using your own resources, you might get data from second-party or third-party data providers. **Second-party data** is collected directly by another group and then sold. **Third-party data** is sold by a provider that didn’t collect the data themselves. Third-party data might come from a number of different sources.

**Solving your business problem**

Datasets can show a lot of interesting information. But be sure to choose data that can actually help solve your problem question. For example, if you are analyzing trends over time, make sure you use time series data — in other words, data that includes dates.

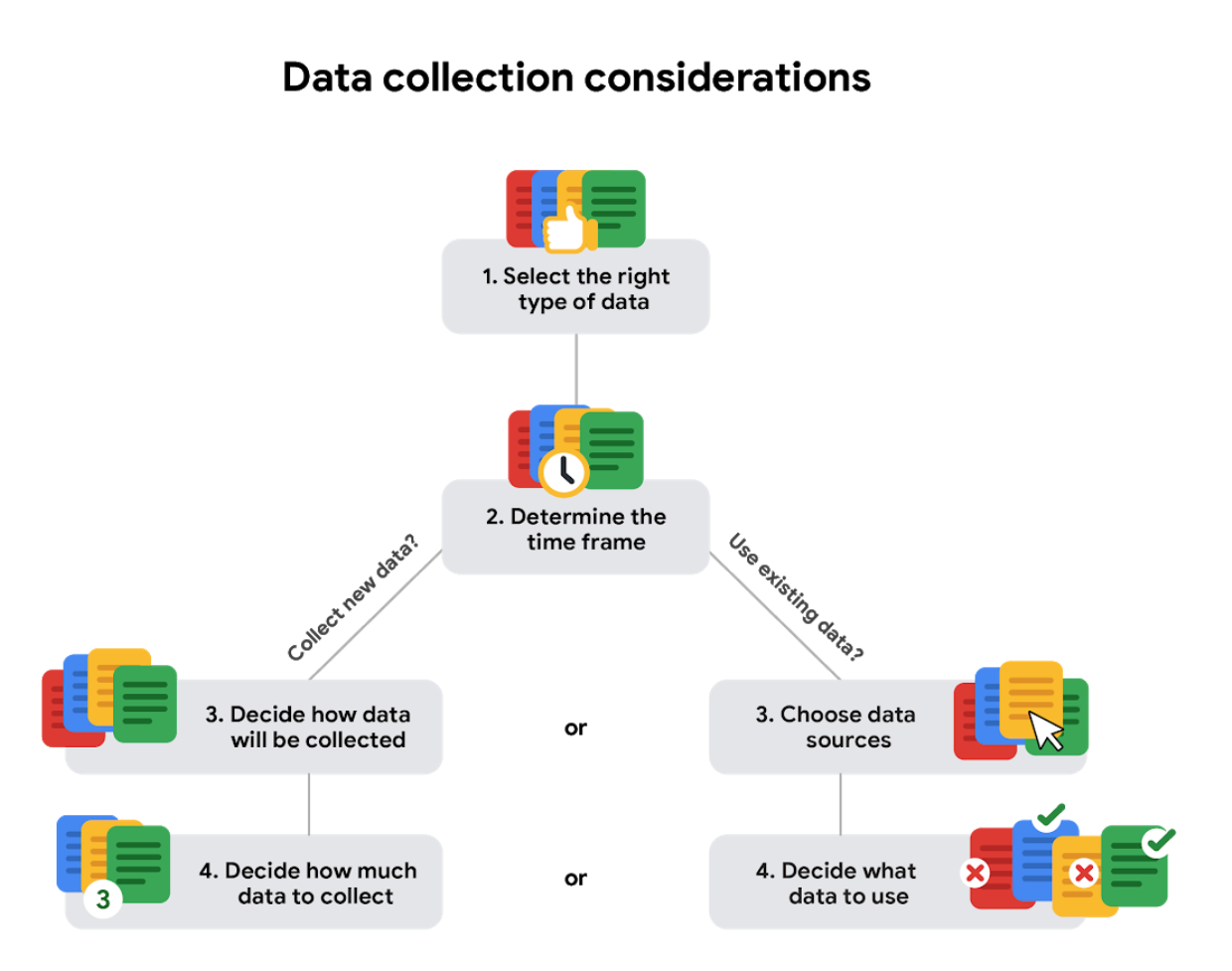
**How much data to collect**

If you are collecting your own data, make reasonable decisions about sample size. A random sample from existing data might be fine for some projects. Other projects might need more strategic data collection to focus on certain criteria. Each project has its own needs.

**Time frame**

If you are collecting your own data, decide how long you will need to collect it, especially if you are tracking trends over a long period of time. If you need an immediate answer, you might not have time to collect new data. In this case, you would need to use historical data that already exists.

Use the flowchart below if data collection relies heavily on how much time you have:



The data-collection process involves deciding what data to use, determining how much data to collect, and selecting the right data type. Which of the following are also steps in the data-collection process? Select all that apply.



Determining the time frame

**Correct**

Determining the time frame and choosing data sources are steps in the data collection process.



Creating data visualizations



Analyzing the data to answer business questions



Choosing data sources

**Correct**

Determining the time frame and choosing data sources are steps in the data collection process.

Which method of data-collection is most commonly used by scientists?

**0 / 1 point**



Questionnaires



Interviews



Surveys



Observations

**Incorrect**

Observation is the method of data-collection most often used by scientists.

Observations

2.

Question 2

Organizations such as the U.S. Centers for Disease Control (CDC) often use data collected from hospitals. What kind of data is the CDC using if it is collected by hospitals, then sold to the CDC for its own analysis?

**1 / 1 point**



Second-party data



Third-party data



Multiple-party data



First-party data

**Correct**

Data gathered by hospitals, then collected by the CDC, is an example of second-party data.

### 3.

Question 3

Fill in the blank: In data analytics, a \_\_\_\_\_ refers to all possible data values in a certain dataset.

**1 / 1 point**



representation



source



population



sample

**Correct**

In data analytics, a population refers to all possible data values in a certain dataset.

I don't know about you, but when

I'm choosing a movie to watch,

I sometimes get stuck between a couple of choices.

If I'm in the mood for excitement or suspense,

I might go for a thriller,

but if I need a good laugh,

I'll choose a comedy.

If I really can't decide between two movies,

I might even use some of my data analysis skills

to compare and contrast them.

Come to think of it, there really needs to be

more movies about data analysts.

I'd watch that, but since

we can't watch a movie about data,

at least not yet, we'll do

the next best thing: watch data about movies!

We're going to take a look at

this spreadsheet with movie data.

We know we can compare different movies and movie genres.

Turns out, you can do the same with

data and data formats.

Let's use our movie data spreadsheet

to understand how that works.

We'll start with quantitative and qualitative data.

If we check out column A, we'll find titles of the movies.

This is qualitative data because it can't be counted,

measured, or easily expressed using numbers.

Qualitative data is usually listed as

a name, category, or description.

In our spreadsheet, the movie titles and

cast members are qualitative data.

Next up is quantitative data,

which can be measured or

counted and then expressed as a number.

This is data with a certain quantity, amount, or range.

In our spreadsheet here,

the last two columns show the movies's

budget and box office revenue.

The data in these columns is listed in dollars,

which can be counted,

so we know that data is quantitative.

We can go even deeper into quantitative data

and break it down into discrete or continuous data.

Let's check out discrete data first.

This is data that's counted

and has a limited number of values.

Going back to our spreadsheet,

we'll find each movie's budget and

box office returns in columns M

and N. These are both examples of

discrete data that can be

counted and have a limited number of values.

For example, the amount of money

a movie makes can only be represented

with exactly two digits after

the decimal to represent cents.

There can't be anything between one and two cents.

Continuous data can be measured using a timer, and

its value can be shown as a decimal with several places.

Let's imagine a movie about data analysts

that I'm definitely going to star in someday.

You could express that movie's run time as

110.0356 minutes.

You could even add fractional data

after the decimal point if you needed to.

There's also nominal and ordinal data.

Nominal data is a type of

qualitative data that's categorized without a set order.

In other words, this data doesn't have a sequence.

Here's a quick example.

Let's say you're collecting data about movies.

You ask people if they've watched a given movie.

Their responses would be in the form of nominal data.

They could respond "Yes,"

"No," or "Not sure."

These choices don't have a particular order.

Ordinal data, on the other hand,

is a type of qualitative data with a set order or scale.

If you asked a group of people

to rank a movie from 1 to 5,

some might rank it as a 2,

others a 4, and so on.

These rankings are in order of

how much each person liked the movie.

Now let's talk about internal data,

which is data that lives within a company's own systems.

For example, if a movie studio had compiled all of

the data in the spreadsheet using

only their own collection methods,

then it would be their internal data.

The great thing about internal data is that it's

usually more reliable and easier to collect,

but in this spreadsheet,

it's more likely that the movie studio

had to use data owned or shared

by other studios and

sources because it includes movies they didn't make.

That means they'd be collecting external data.

External data is, you guessed it,

data that lives and

is generated outside of an organization.

External data becomes particularly valuable when

your analysis depends on as many sources as possible.

A great thing about this data is that it's structured.

Structured data is data

that's organized in a certain format,

such as rows and columns.

Spreadsheets and relational databases are

two examples of software

that can store data in a structured way.

You might remember our earlier exploration

of structured thinking,

which helps you add a framework to a problem so that

you can solve it in an organized and logical manner.

You can think of structured data in the same way.

Having a framework for the data makes the data

easily searchable and more analysis-ready.

As a data analyst,

you'll work with a lot of structured data,

which will usually be in the form of

a table, spreadsheet or relational database,

but sometimes you'll come across unstructured data.

This is data that is not organized in

any easily identifiable manner.

Audio and video files are examples of unstructured data

because there's no clear way to

identify or organize their content.

Unstructured data might have internal structure,

but the data doesn't fit neatly in rows and

columns like structured data. And there you have it!

Hopefully you're now more familiar with

data formats and how you might use them in your work.

In just a bit,

you'll continue to explore structured data and learn

even more about the data you'll

use most often as an analyst.

Coming soon to a screen near you.

An entertainment website displays a star rating for a movie based on user reviews. Users can select from one to five whole stars to rate the movie. The star rating is an example of what type of data? Select all that apply.



Discrete



Continuous



Ordinal



Nominal

The use of external data is particularly valuable in which circumstances?



When analysis involves data that hasn’t been cleaned



When analysis requires a lot of structured data



When analysis includes data from audio files



When analysis depends on as many data sources as possible

Data formats in practice

When you think about the word "format," a lot of things might come to mind. Think of an advertisement for your favorite store. You might find it in the form of a print ad, a billboard, or even a commercial. The information is presented in the format that works best for you to take it in. The format of a dataset is a lot like that, and choosing the right format will help you manage and use your data in the best way possible.

A screenshot of a computer

Description automatically generated with low confidence

**Data format examples**

As with most things, it is easier for definitions to click when we can pair them with real life examples. Review each definition first and then use the examples to lock in your understanding of each data format.

the following table highlights the differences between primary and secondary data and examples of each

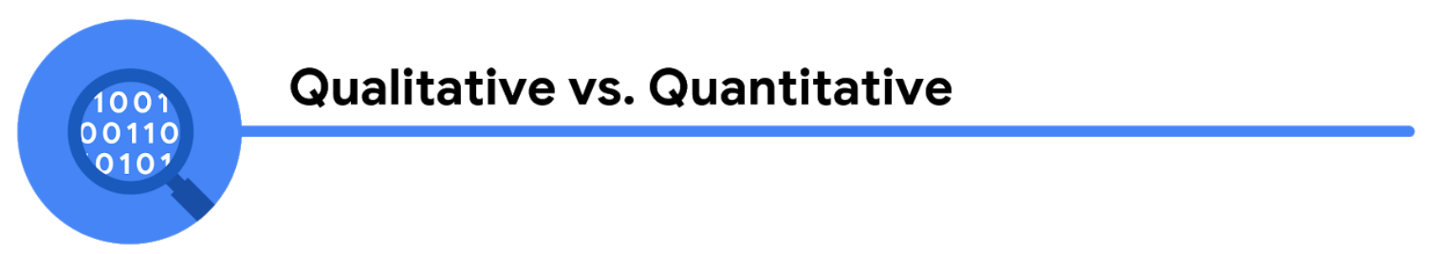
| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Primary data | Collected by a researcher from first-hand sources | - Data from an interview you conducted - Data from a survey returned from 20 participants - Data from questionnaires you got back from a group of workers |
| Secondary data | Gathered by other people or from other research | - Data you bought from a local data analytics firm’s customer profiles - Demographic data collected by a university - Census data gathered by the federal government |

the following table highlights the differences between internal and external data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Internal data | Data that lives inside a company’s own systems | - Wages of employees across different business units tracked by HR - Sales data by store location - Product inventory levels across distribution centers |
| External data | Data that lives outside of a company or organization | - National average wages for the various positions throughout your organization - Credit reports for customers of an auto dealership |

the following table highlights the differences between continuous and discrete data and examples of each

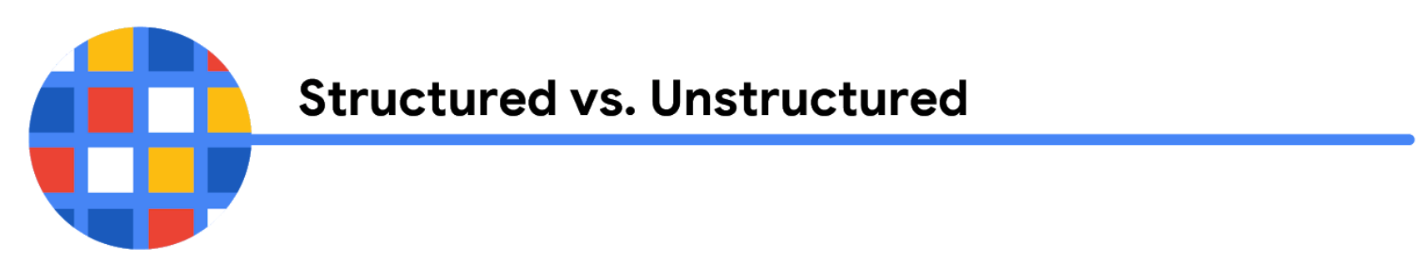
| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Continuous data | Data that is measured and can have almost any numeric value | - Height of kids in third grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature |
| Discrete data | Data that is counted and has a limited number of values | - Number of people who visit a hospital on a daily basis (10, 20, 200) - Room’s maximum capacity allowed - Tickets sold in the current month |

the following table highlights the differences between qualitative and quantitative data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Qualitative | Subjective and explanatory measures of qualities and characteristics | - Exercise activity most enjoyed - Favorite brands of most loyal customers - Fashion preferences of young adults |
| Quantitative | Specific and objective measures of numerical facts | - Percentage of board certified doctors who are women - Population of elephants in Africa - Distance from Earth to Mars |

the following table highlights the differences between nominal and ordinal data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Nominal | A type of qualitative data that isn’t categorized with a set order | - First time customer, returning customer, regular customer - New job applicant, existing applicant, internal applicant - New listing, reduced price listing, foreclosure |
| Ordinal | A type of qualitative data with a set order or scale | - Movie ratings (number of stars: 1 star, 2 stars, 3 stars) - Ranked-choice voting selections (1st, 2nd, 3rd) - Income level (low income, middle income, high income) |

the following table highlights the differences between structured and unstructured data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Structured data | Data organized in a certain format, like rows and columns | - Expense reports - Tax returns - Store inventory |
| Unstructured data | Data that isn’t organized in any easily identifiable manner | - Social media posts - Emails - Videos |

Hi, great to see you again!

Earlier, we compared some data formats, including structured and

unstructured data.

Most of the data being generated right now is actually unstructured.

Audio files, video files, emails, photos, and

social media are all examples of unstructured data.

These can be harder to analyze in their unstructured format.

But here's the good news,

you'll be working with structured data most of the time.

For example, if you need to analyze data about the unstructured data in emails,

photos, and social media sites,

it'll most likely be structured for analysis before you even get to it.

Because of that, I want to explore structured data a bit more.

As a quick refresher, structured data is data organized in a format like rows and

columns.

But there's definitely more to it than that.

Structured data works nicely within a data model, which is a model that is used for

organizing data elements and how they relate to one another.

What are data elements?

They're pieces of information, such as people's names,

account numbers, and addresses.

Data models help to keep data consistent and

provide a map of how data is organized.

This makes it easier for analysts and

other stakeholders to make sense of their data and use it for business purposes.

In addition to working well within data models, structured data is also useful for

databases.

This makes it easy for analysts to enter, query, and

analyze the data whenever they need to.

This also helps make data visualization pretty easy

because structured data can be applied directly to charts, graphs, heat maps,

dashboards and most other visual representations of data.

Alright, so now we know that spreadsheets and

databases that store data sets are widely used sources of structured data.

After you explore some other data structures,

you'll check out more data types using a spreadsheet.

The adventure continues!

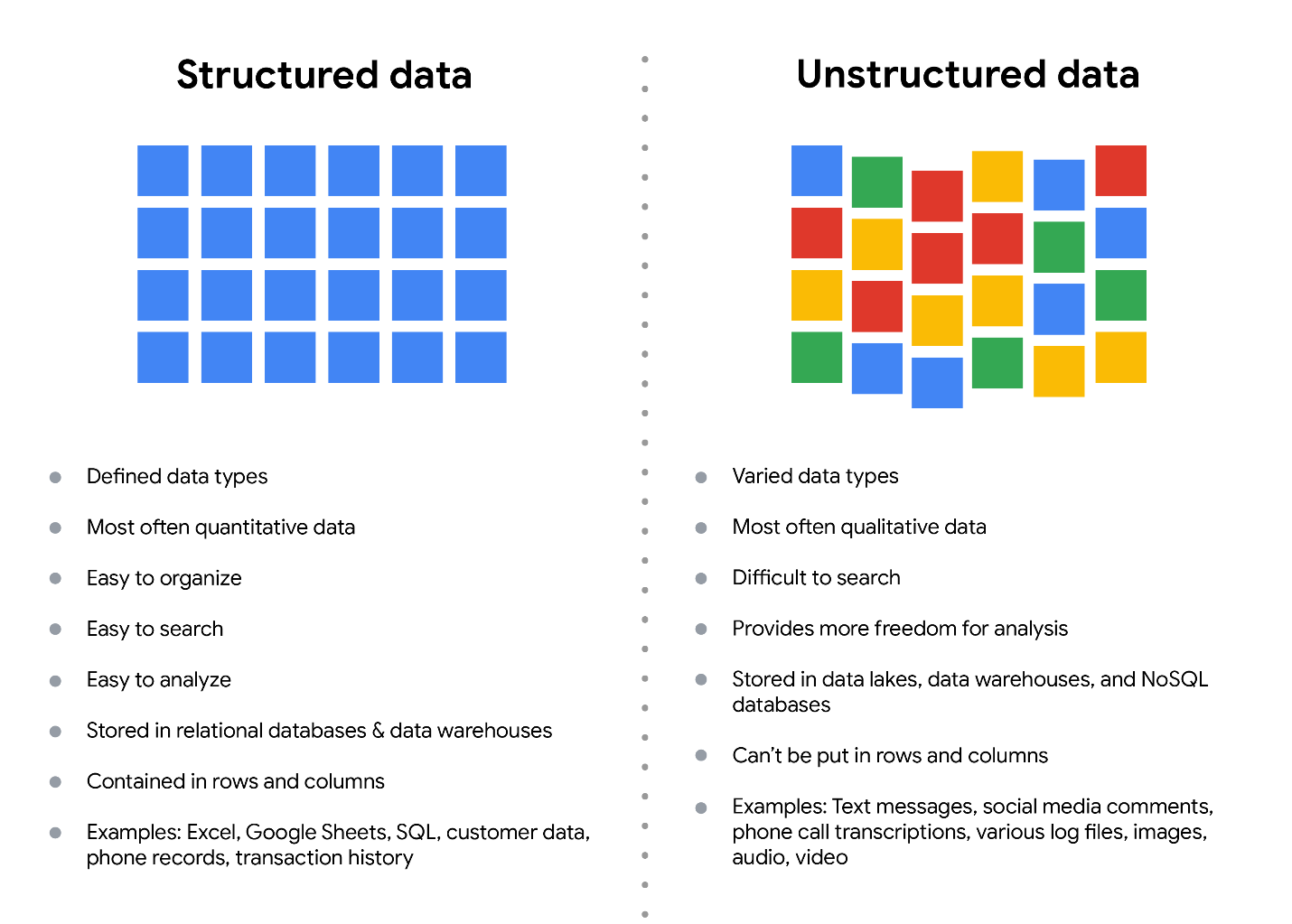
The structure of data

Data is everywhere and it can be stored in lots of ways. Two general categories of data are:

* **Structured data:** Organized in a certain format, such as rows and columns.
* **Unstructured data:** Not organized in any easy-to-identify way.

For example, when you rate your favorite restaurant online, you're creating structured data. But when you use Google Earth to check out a satellite image of a restaurant location, you're using unstructured data.

Here's a refresher on the characteristics of structured and unstructured data:

Structured data: - Defined data types - Most often quantitative data - Easy to organize - Easy to search - Easy to analyze - Stored in relational databases - Contained in rows and columns - Examples: Excel, Google Sheets, SQL, customer data, phone records, transaction history Unstructured data: - Varied data types - Most often qualitative data - Difficult to search - Provides more freedom for analysis - Stored in data lakes and NoSQL databases - Can't be put in rows and columns - Examples: Text messages, social media comments, phone call transcriptions, various log files, images, audio, video

**Structured data**

As we described earlier, **structured data** is organized in a certain format. This makes it easier to store and query for business needs. If the data is exported, the structure goes along with the data.

**Unstructured data**

**Unstructured data** can’t be organized in any easily identifiable manner. And there is much more unstructured than structured data in the world. Video and audio files, text files, social media content, satellite imagery, presentations, PDF files, open-ended survey responses, and websites all qualify as types of unstructured data.

**The fairness issue**

The lack of structure makes unstructured data difficult to search, manage, and analyze. But recent advancements in artificial intelligence and machine learning algorithms are beginning to change that. Now, the new challenge facing data scientists is making sure these tools are inclusive and unbiased. Otherwise, certain elements of a dataset will be more heavily weighted and/or represented than others. And as you're learning, an unfair dataset does not accurately represent the population, causing skewed outcomes, low accuracy levels, and unreliable analysis.

Data modeling levels and techniques

This reading introduces you to data modeling and different types of data models. Data models help keep data consistent and enable people to map out how data is organized. A basic understanding makes it easier for analysts and other stakeholders to make sense of their data and use it in the right ways.

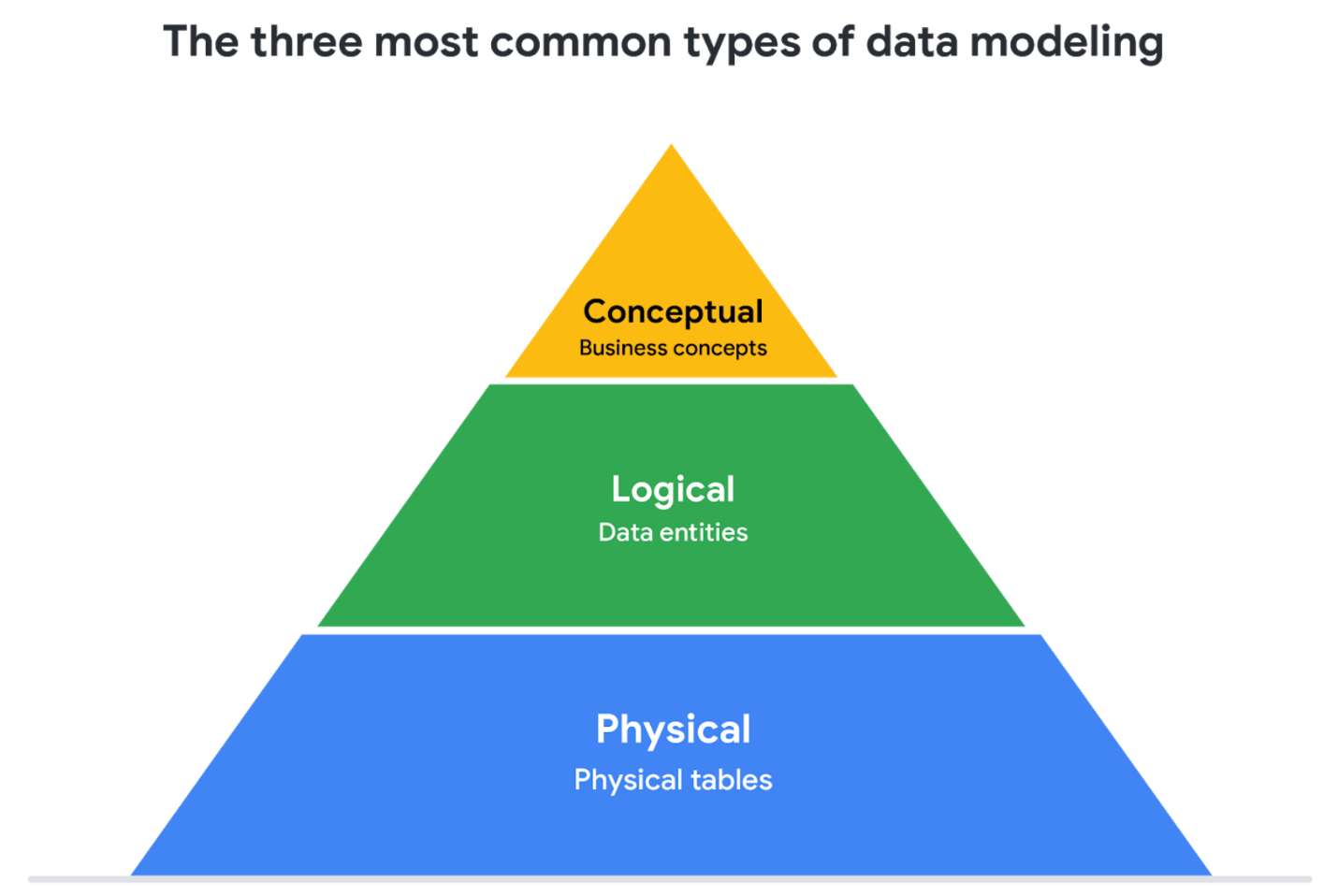
**Important note:** As a junior data analyst, you won't be asked to design a data model. But you might come across existing data models your organization already has in place.

**What is data modeling?**

**Data modeling** is the process of creating diagrams that visually represent how data is organized and structured.  These visual representations are called **data models**. You can think of data modeling as a blueprint of a house. At any point, there might be electricians, carpenters, and plumbers using that blueprint. Each one of these builders has a different relationship to the blueprint, but they all need it to understand the overall structure of the house. Data models are similar; different users might have different data needs, but the data model gives them an understanding of the structure as a whole.

**Levels of data modeling**

Each level of data modeling has a different level of detail.



1. **Conceptual data modeling** gives a high-level view of the data structure, such as how data interacts across an organization. For example, a conceptual data model may be used to define the business requirements for a new database. A conceptual data model doesn't contain technical details.
2. **Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
3. **Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

More information can be found in this [comparison of data models.](https://www.1keydata.com/datawarehousing/data-modeling-levels.html)

**Data-modeling techniques**

There are a lot of approaches when it comes to developing data models, but two common methods are the **Entity Relationship Diagram (ERD)** and the **Unified Modeling Language (UML)** diagram. ERDs are a visual way to understand the relationship between entities in the data model. UML diagrams are very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations, and their relationships. As a junior data analyst, you will need to understand that there are different data modeling techniques, but in practice, you will probably be using your organization’s existing technique.

You can read more about ERD, UML, and data dictionaries in this [data modeling techniques article](https://dataedo.com/blog/basic-data-modeling-techniques).

**Data analysis and data modeling**

Data modeling can help you explore the high-level details of your data and how it is related across the organization’s information systems. Data modeling sometimes requires data analysis to understand how the data is put together; that way, you know how to map the data. And finally, data models make it easier for everyone in your organization to understand and collaborate with you on your data. This is important for you and everyone on your team!

### 1.

Question 1

Fill in the blank: The running time of a movie is an example of \_\_\_\_\_ data.

**1 point**



discrete



nominal



continuous



qualitative

### 2.

Question 2

What are the characteristics of unstructured data? Select all that apply.

**1 point**



Fits neatly into rows and columns



Is not organized



Has a clearly identifiable structure



May have an internal structure

### 3.

Question 3

Structured data enables data to be grouped together to form relations. This makes it easier for analysts to do what with the data? Select all that apply.

**1 point**



Search



Analyze



Rewrite



Store

### 4.

Question 4

Which of the following is an example of unstructured data?

**1 point**



Email message



Contact saved on a phone



GPS location



Rating of a local favorite restaurant

By now you've learned a lot about data.

From generated data, to collected data, to data formats,

it's good to know as much as you can

about the data you'll use for analysis.

In this video, we'll talk about another way

you can describe data: the data type.

A data type is a specific kind of data attribute

that tells what kind of value the data is.

In other words, a data type

tells you what kind of data you're working with.

Data types can be different

depending on the query language you're using.

For example, SQL allows for

different data types depending

on which database you're using.

For now though, let's focus on

the data types that you'll use in spreadsheets.

To help us out, we'll use

a spreadsheet that's already filled with data.

We'll call it "Worldwide

Interests in Sweets through Google Searches."

Now a data type in

a spreadsheet can be one of three things:

a number, a text

or string, or a Boolean.

You might find spreadsheet programs that classify them

a bit differently or include other types,

but these value types cover just about

any data you'll find in spreadsheets.

We'll look at all of these in just a bit.

Looking at columns B, D,

and F, we find number data types.

Each number represents the search interest

for the terms "cupcakes,"

"ice cream," and "candy" for a specific week.

The closer a number is to 100,

the more popular that search term was during that week.

One hundred represents peak popularity.

Keep in mind that in this case,

100 is a relative value,

not the actual number of searches.

It represents the maximum number

of searches during a certain time.

Think of it like a percentage on a test.

All other searches are then also valued out of 100.

You might notice this in other data sets as well.

Gold star for 100!

If you needed to, you could change the numbers into

percents or other formats, like currency.

These are all examples of number data types.

In column H, the data shows

the most popular treat for each week,

based on the search data.

So as we'll find in cell H4 for

the week beginning July 28th, 2019,

the most popular treat was ice cream.

This is an example of a text data type,

or a string data type,

which is a sequence of characters and

punctuation that contains textual information.

In this example, that information

would be the treats and people's names.

These can also include numbers, like

phone numbers or numbers in street addresses.

But these numbers wouldn't be used for calculations.

In this case they're treated like text, not numbers.

In columns C, E,

and G, it seems like we've got some text.

But the text here isn't a text or string data type.

Instead, it's a Boolean data type.

A Boolean data type is

a data type with only two possible values:

true or false.

Columns C, E, and G show

Boolean data for whether the

search interest for each week,

is at least 50 out of 100.

Here's how it works. To get this data,

we've created a formula that calculates

whether the search interest data in columns B,

D, and F is 50 or greater.

In cell B4, the search interest is 14.

In cell C4, we find the word false

because, for this week of data,

the search interest is less than 50.

For each cell in columns C, E,

and G, the only two possible values are true or false.

We could change the formula so

other words appear in these cells instead,

but it's still Boolean data.

You'll get a chance to read more

about the Boolean data type soon.

Let's talk about a common issue that

people encounter in spreadsheets:

mistaking data types with cell values.

For example, in cell B57,

we can create a formula to calculate data in other cells.

This will give us the average of the search interests

in cupcakes across all weeks in the dataset,

which is about 15.

The formula works because we

calculated using a number data type.

But if we tried it with a text or string data type,

like the data in column C, we'd get an error.

Error values usually happen if a mistake is

made in entering the values in the cells.

The more you know your data types and which ones to use,

the less errors you'll run into.

There you have it, a data type for everyone.

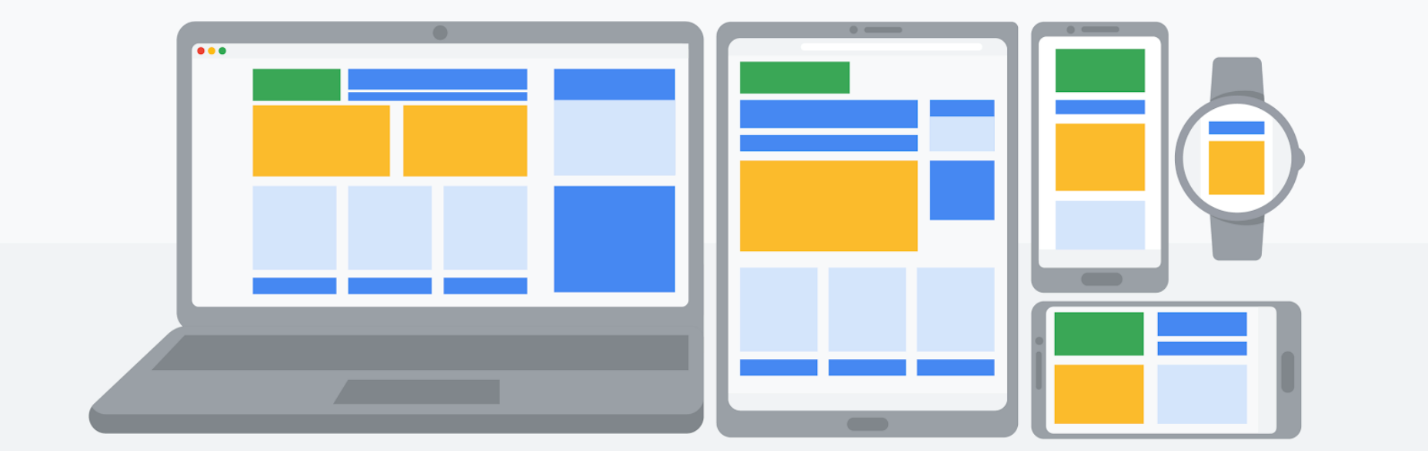
We're not done yet. Coming up,

we'll go deeper into the relationship between data types,

fields, and values. See you soon.

# Understanding Boolean logic

In this reading, you will explore the basics of Boolean logic and learn how to use multiple conditions in a Boolean statement. These conditions are created with Boolean operators, including AND, OR, and NOT. These operators are similar to mathematical operators and can be used to create logical statements that filter your results. Data analysts use Boolean statements to do a wide range of data analysis tasks, such as creating queries for searches and checking for conditions when writing programming code.

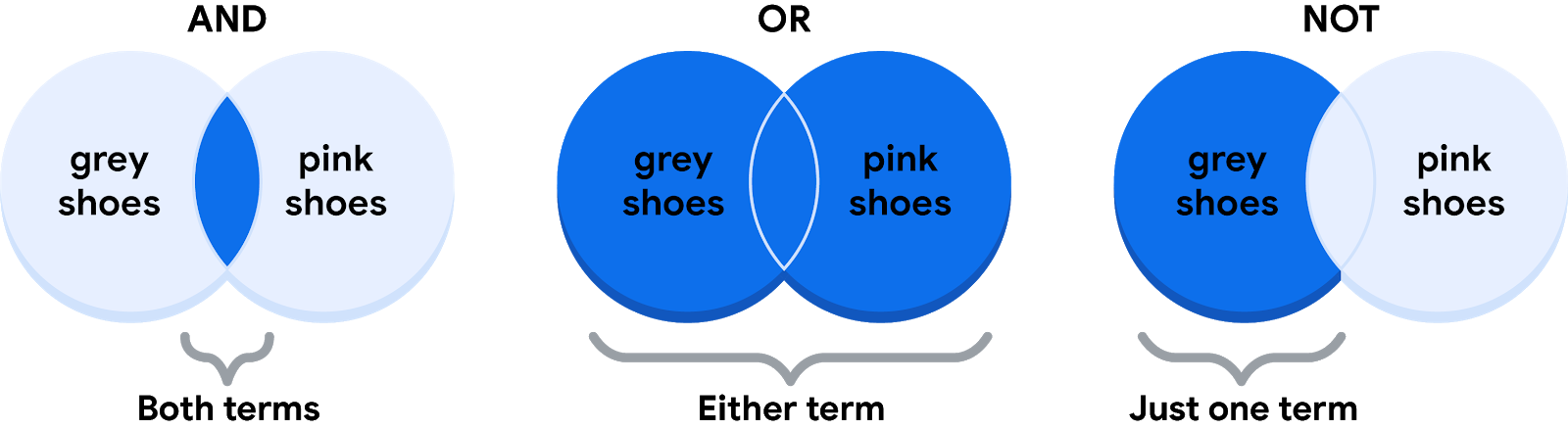


## Boolean logic example

Imagine you are shopping for shoes, and are considering certain preferences:

* You will buy the shoes only if they are pink and grey
* You will buy the shoes if they are entirely pink or entirely grey, or if they are pink and grey
* You will buy the shoes if they are grey, but not if they have any pink

Below are Venn diagrams that illustrate these preferences. AND is the center of the Venn diagram, where two conditions overlap. OR includes either condition. NOT includes only the part of the Venn diagram that doesn't contain the exception.



### **The AND operator**

Your condition is “If the color of the shoe has any combination of grey and pink, you will buy them.” The Boolean statement would break down the logic of that statement to filter your results by both colors. It would say “IF (Color=”Grey”) AND (Color=”Pink”) then buy them.” The AND operator lets you stack multiple conditions.

Below is a simple truth table that outlines the Boolean logic at work in this statement. In the **Color is Grey** column, there are two pairs of shoes that meet the color condition. And in the **Color is Pink** column, there are two pairs that meet that condition. But in the **If Grey AND Pink** column, there is only one pair of shoes that meets both conditions. So, according to the Boolean logic of the statement, there is only one pair marked true. In other words, there is one pair of shoes that you can buy.

| **Color is Grey** | **Color is Pink** | **If Grey AND Pink, then Buy** | **Boolean Logic** |
| --- | --- | --- | --- |
| Grey/True | Pink/True | True/Buy | True AND True = True |
| Grey/True | Black/False | False/Don't buy | True AND False = False |
| Red/False | Pink/True | False/Don't buy | False AND True = False |
| Red/False | Green/False | False/Don't buy | False AND False = False |

### **The OR operator**

The OR operator lets you move forward if either one of your two conditions is met. Your condition is “If the shoes are grey or pink, you will buy them.” The Boolean statement would be “IF (Color=”Grey”) OR (Color=”Pink”) then buy them.” Notice that any shoe that meets either the **Color is Grey** or the **Color is Pink** condition is marked as true by the Boolean logic. According to the truth table below, there are three pairs of shoes that you can buy.

| **Color is Grey** | **Color is Pink** | **If Grey OR Pink, then Buy** | **Boolean Logic** |
| --- | --- | --- | --- |
| Red/False | Black/False | False/Don't buy | False OR False = False |
| Black/False | Pink/True | True/Buy | False OR True = True |
| Grey/True | Green/False | True/Buy | True OR False = True |
| Grey/True | Pink/True | True/Buy | True OR True = True |

### **The NOT operator**

Finally, the NOT operator lets you filter by subtracting specific conditions from the results. Your condition is "You will buy any grey shoe except for those with any traces of pink in them." Your Boolean statement would be “IF (Color="Grey") AND (Color=NOT “Pink”) then buy them.” Now, all of the grey shoes that aren't pink are marked true by the Boolean logic for the **NOT Pink** condition. The pink shoes are marked false by the Boolean logic for the **NOT Pink** condition. Only one pair of shoes is excluded in the truth table below.

| **Color is Grey** | **Color is Pink** | **Boolean Logic for NOT Pink** | **If Grey AND (NOT Pink), then Buy** | **Boolean Logic** |
| --- | --- | --- | --- | --- |
| Grey/True | Red/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Black/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Green/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Pink/True | Not True = False | False/Don't buy | True AND False = False |

## The power of multiple conditions

For data analysts, the real power of Boolean logic comes from being able to combine multiple conditions in a single statement. For example, if you wanted to filter for shoes that were grey or pink, and waterproof, you could construct a Boolean statement such as: “IF ((Color = “Grey”) OR (Color = “Pink”)) AND (Waterproof=“True”).”  Notice that you can use parentheses to group your conditions together.

Whether you are doing a search for new shoes or applying this logic to your database queries, Boolean logic lets you create multiple conditions to filter your results. And now that you know a little more about how Boolean logic is used, you can start using it!

## Additional Reading/Resources

* Learn about who pioneered Boolean logic in this historical article: [Origins of Boolean Algebra in the Logic of Classes](https://www.maa.org/press/periodicals/convergence/origins-of-boolean-algebra-in-the-logic-of-classes-george-boole-john-venn-and-c-s-peirce).
* Find more information about using AND, OR, and NOT from these [tips for searching with Boolean operators](https://libguides.mit.edu/c.php?g=175963&p=1158594).

Here's a riddle for you.

What do a music playlist, a calendar agenda, and an email inbox have in common?

I'll give you a hint.

It's not a weekly jam session.

The answer is they're all arranged in tables.

Go ahead and check out your email inbox or a favorite playlist, or

look at your calendar agenda.

There's tables in every one!

A data table, or tabular data, has a very simple structure.

It's arranged in rows and columns.

You can call the rows "records" and the columns "fields."

They basically mean the same thing,

but records and fields can be used for any kind of data table, while rows and

columns are usually reserved for spreadsheets.

When talking about structured databases,

people in data analytics usually go with "records" and "fields."

Sometimes a field can also refer to a single piece of data,

like the value in a cell.

In any case,

you'll hear both versions of these terms used throughout this program and your job.

Let's go back to our playlist example.

We'll use the new terms we just introduced.

So each song is a record.

Each record has the same fields as the other records in the same order.

In other words, the playlist has the same information about each song.

Play video starting at :1:10 and follow transcript1:10

Each song characteristic, like the title and the artist, is a field.

Each separate field has the same data type, but

different fields can have different types.

Let me show you what I mean.

For the song list, the song titles are a text or string type, while

the song's length could be a number type if you're using it for calculations. Or

it could be a date and time type.

The column for favorites is Boolean

since it has two possible values: favorite or not favorite.

We can view spreadsheets in the same way.

The records in a spreadsheet might be about all sorts of things:

clients, products, invoices, or anything else.

Each record has several fields,

which reveal more about the clients, products, or invoices.

The value in every cell contains a specific piece of data,

like the address of a client or the dollar amount of an invoice.

As a data analyst, lots of data will come your way, and records, fields, and

values in data tables will help you navigate analysis.

Understanding the structures of the tables you're working with is a part of that.

And hopefully, while you're working hard on your analysis and

those tables, you can have a little fun with a different data table:

the one with your favorite playlist!

You probably use the words "wide" and "long" all the time.

You might use "wide" to describe the size of something from side to side, like

a wide river.

But a river can also travel great distances,

so you might call it "long" as well.

Wait! Before you stop the video,

I promise you didn't accidentally click in the wrong course.

I'm not here to teach you words you already know. But the words "wide" and

"long" can be used to describe data, too.

So I am here to help you understand wide data and long data.

So far you've dealt with data arranged mostly in a wide format.

With wide data, every data subject has a single row with multiple columns to hold

the values of various attributes of the subject.

Here's some wide data in a spreadsheet.

You might remember we discussed this data about the population of Latin and

Caribbean countries earlier.

For this data set, each row provides all of the population information about one

country. Each column shows the population for a different year.

Play video starting at ::57 and follow transcript0:57

Wide data lets you easily identify and quickly compare different columns.

In our example, the data is arranged alphabetically by country, so

you can compare the annual populations of Antigua and Barbuda, Aruba, and

the Bahamas by just checking out the values in each column.

The wide data format also makes it easy to find and

compare the countries' populations at different periods of time.

For example, by sorting the data,

we discover that Brazil had the highest population of all countries in 2010, and

the British Virgin Islands had the lowest population of all countries in 2013.

Okay, now let's explore this data in a long format.

Here the data is no longer organized into columns by year.

All the years are now in one column with each country, like Argentina, appearing in

multiple rows, one for each year of data.

This is how long data usually looks.

Long data is data in which each row is one time point per subject, so

each subject will have data in multiple rows.

Our spreadsheet is formatted to show each year of population data.

Play video starting at :2:1 and follow transcript2:01

Here we see Antigua and Barbuda first.

Long data is a great format for storing and organizing data when there's multiple

variables for each subject at each time point that we want to observe.

With this long data format, we can store and

analyze all of this data using fewer columns. Plus, if we added a new variable,

like the average age of a population, we'd only need one more column.

If we'd use a wide data format instead, we would have needed 10 more columns, one for

each year.

The long data format keeps everything nice and compact.

If you're wondering which format you should use,

the simple answer is, "it depends."

Play video starting at :2:37 and follow transcript2:37

Sometimes you'll have to transform wide data into a long data format, or

other times vice versa.

You'll probably work with both formats in your job. And

you'll definitely revisit both formats again later in this program.

That reminds me: earlier we define data as a collection of facts.

As you've discovered over the last few videos, that collection of facts can take

on lots of different formats, structures, types, and more.

Play video starting at :3:3 and follow transcript3:03

Learning about all of the ways that data can be presented will be a big help

to you throughout the data analysis process.

The more you work with data in all its forms,

the quicker you'll start to recognize which data to use, and when to use it.

Play video starting at :3:16 and follow transcript3:16

And in just a bit,

you'll use all that data stored in your brain to help you take an assessment.

After that, you'll learn how to identify and avoid bias in data and

how to embrace credibility, integrity and ethics.

The data adventure moves forward. I'm so glad you're moving with it!

Transforming data

**What is data transformation?**

A woman presenting data, a hand holding a medal, two people chatting, a ship's wheel being steered, two people high-fiving each other

In this reading, you will explore how data is transformed and the differences between wide and long data. **Data transformation** is the process of changing the data’s format, structure, or values. As a data analyst, there is a good chance you will need to transform data at some point to make it easier for you to analyze it.

Data transformation usually involves:

* Adding, copying, or replicating data
* Deleting fields or records
* Standardizing the names of variables
* Renaming, moving, or combining columns in a database
* Joining one set of data with another
* Saving a file in a different format. For example, saving a spreadsheet as a comma separated values (CSV) file.

**Why transform data?**

Goals for data transformation might be:

* Data **organization**: better organized data is easier to use
* Data **compatibility**: different applications or systems can then use the same data
* Data **migration**: data with matching formats can be moved from one system to another
* Data **merging**: data with the same organization can be merged together
* Data **enhancement**: data can be displayed with more detailed fields
* Data **comparison**: apples-to-apples comparisons of the data can then be made

**Data transformation example: data merging**

Mario is a plumber who owns a plumbing company. After years in the business, he buys another plumbing company. Mario wants to merge the customer information from his newly acquired company with his own, but the other company uses a different database. So, Mario needs to make the data compatible. To do this, he has to transform the format of the acquired company’s data. Then, he must remove duplicate rows for customers they had in common. When the data is compatible and together, Mario’s plumbing company will have a complete and merged customer database.

**Data transformation example: data organization (long to wide)**

To make it easier to create charts, you may also need to transform long data to wide data. Consider the following example of transforming stock prices (collected as long data) to wide data.

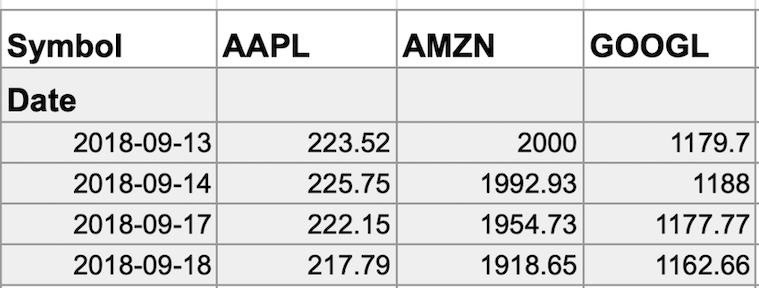
**Long data** is data where **each row contains a single data point** for a particular item. In the long data example below, individual stock prices (data points) have been collected for Apple (AAPL), Amazon (AMZN), and Google (GOOGL) (particular items) on the given dates.

**Long data example: Stock prices**



**Wide data** is data where **each row contains multiple data points** for the particular items identified in the columns.

**Wide data example: Stock prices**



With data transformed to wide data, you can create a chart comparing how each company's stock changed over the same period of time.

You might notice that all the data included in the long format is also in the wide format. But wide data is easier to read and understand. That is why data analysts typically transform long data to wide data more often than they transform wide data to long data. The following table summarizes when each format is preferred:

| **Wide data is preferred when** | **Long data is preferred when** |
| --- | --- |
| Creating tables and charts with a few variables about each subject | Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank |
| Comparing straightforward line graphs | Performing advanced statistical analysis or graphing |

### 1.

Question 1

Fill in the blank: Internet search engines are an everyday example of how Boolean operators are used. The Boolean operator \_\_\_\_\_ expands the number of results when used in a keyword search.

**1 / 1 point**



AND



NOT



OR



WITH

**Correct**

The Boolean operator OR expands the number of results when used in a keyword search.

### 2.

Question 2

Which of the following statements accurately describes a key difference between wide and long data?

**1 / 1 point**



Every wide data subject has a single column that holds the values of subject attributes. Every long data subject has multiple columns.



Wide data subjects can have data in multiple columns. Long data subjects can have multiple rows that hold the values of subject attributes.



Wide data subjects can have multiple rows that hold the values of subject attributes. Long data subjects can have data in multiple columns.



Every wide data subject has multiple columns. Every long data subject has data in a single column.

**Correct**

Wide data subjects can have data in multiple columns. Long data subjects can have multiple rows that hold the values of subject attributes.

### 3.

Question 3

What does data transformation enable data analysts to accomplish?

**0 / 1 point**



Retrieve the data faster



Restore the data after it has been lost



Change the structure of the data



Inspect the data for accuracy

A

**Action-oriented question:** A question whose answers lead to change

**Agenda:** A list of scheduled appointments

**Algorithm:** A process or set of rules followed for a specific task

**Analytical skills:** Qualities and characteristics associated with using facts to solve problems

**Analytical thinking:** The process of identifying and defining a problem, then solving it by using data in an organized, step-by-step manner

**Attribute**: A characteristic or quality of data used to label a column in a table

**Audio file:** Digitized audio storage usually in an MP3, AAC, or other compressed format

**AVERAGE**: A spreadsheet function that returns an average of the values from a selected range

B

**Big data:** Large, complex datasets typically involving long periods of time, which enable data analysts to address far-reaching business problems

**Boolean data:** A data type with only two possible values, usually true or false

**Borders**: Lines that can be added around two or more cells on a spreadsheet

**Business task:** The question or problem data analysis resolves for a business

C

**Cell reference:** A cell or a range of cells in a worksheet typically used in formulas and functions

**Cloud:** A place to keep data online, rather than a computer hard drive

**Context:** The condition in which something exists or happens

**Continuous data:** Data that is measured and can have almost any numeric value

**Cookie:** A small file stored on a computer that contains information about its users

**COUNT**: A spreadsheet function that counts the number of cells in a range that meet a specified criteria

D

**Dashboard:** A tool that monitors live, incoming data

**Data:** A collection of facts

**Data analysis:** The collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making

**Data analysis process:** The six phases of ask, prepare, process, analyze, share, and act whose purpose is to gain insights that drive informed decision-making

**Data analyst:** Someone who collects, transforms, and organizes data in order to draw conclusions, make predictions, and drive informed decision-making

**Data analytics:** The science of data

**Data design:** How information is organized

**Data-driven decision-making:** Using facts to guide business strategy

**Data ecosystem:** The various elements that interact with one another in order to produce, manage, store, organize, analyze, and share data

**Data element:** Apiece of information in a dataset

**Data-inspired decision-making:** Exploring different data sources to find out what they have in common

**Data life cycle:** The sequence of stages that data experiences, which include plan, capture, manage, analyze, archive, and destroy

**Data model:** A tool for organizing data elements and how they relate to one another

**Data science:** A field of study that uses raw data to createnew ways of modeling and understanding the unknown

**Data strategy:** The management of the people, processes, and tools used in data analysis

**Data visualization:** The graphical representation of data

**Data type:** An attribute that describes a piece of data based on its values, its programming language, or the operations it can perform

**Database:** A collection of data stored in a computer system

**Dataset:** A collection of data that can be manipulated or analyzed as one unit

**Digital photo:** An electronic or computer-based image usually in BMP or JPG format

**Discrete data:** Data that is counted and has a limited number of values

E

**Equation:** A calculation that involves addition, subtraction, multiplication, or division (also called a math expression)

**External data:** Data that lives, and is generated, outside of an organization

F

**Fairness:** A quality of data analysis that does not create or reinforce bias

**Field:** A single piece of information from a row or column of a spreadsheet;in a data table, typically a column in the table

**Fill handle:** A box in the lower-right-hand corner of a selected spreadsheet cell that can be dragged through neighboring cells in order to continue an instruction

**Filtering:** The process of showing only the data that meets a specified criteria while hiding the rest

**First-party data:** Data collected by an individual or group using their own resources

**Formula:** A set of instructions used to perform a calculation using the data in a spreadsheet

**Function:** A preset command that automatically performs a specified process or task using the data in a spreadsheet

G

**Gap analysis:** A method for examining and evaluating the current state of a process in order to identify opportunities for improvement in the future

H

**Header:** The first row in a spreadsheet that labels the type of data in each column

I

**Internal data:** Data that lives within a company’s own systems

J

K

L

**Leading question:** A question that steers people toward a certain response

**Long data:** A dataset in which each row is one time point per subject, so each subject has data in multiple rows

M

**Math expression:** A calculation that involves addition, subtraction, multiplication, or division (also called an equation)

**Math function**: A function that is used as part of a mathematical formula

**MAX:** A spreadsheet function that returns the largest numeric value from a range of cells

**Measurable question:** A question whose answers can be quantified and assessed

**Metric:** A single, quantifiable type of data that is used for measurement

**Metric goal:** A measurable goal set by a company and evaluated using metrics

**MIN**: A spreadsheet function that returns the smallest numeric value from a range of cells

N

**Nominal data:** A type of qualitative data that is categorized without a set order

O

**Observation:** The attributes that describe a piece of data contained in a row of a table

**Open data:** Data that is available to the public

**Operator:** A symbol that names the operation or calculation to be performed

**Order of operations:** Using parentheses to group together spreadsheet values in order to clarify the order in which operations should be performed

**Ordinal data:** Qualitative data with a set order or scale

**Ownership**: The aspect of data ethics that presumes individuals own the raw data they provide and have primary control over its usage, processing, and sharing

P

**Pivot chart:** A chart created from the fields in a pivot table

**Pivot table:** A data summarization tool used to sort, reorganize, group, count, total, or average data

**Pixel:** In digital imaging, a small area of illumination on a display screen that, when combined with other adjacent areas, forms a digital image

**Population:** In data analytics, all possible data values in a dataset

**Problem domain:** The area of analysis that encompasses every activity affecting or affected by a problem

**Problem types:** The various problems that data analysts encounter, including categorizing things, discovering connections, finding patterns, identifying themes, making predictions, and spotting something unusual

Q

**Qualitative data:** A subjective and explanatory measure of a quality or characteristic

**Quantitative data:** A specific and objective measure, such as a number, quantity, or range

**Query:** A request for data or information from a database

**Query language:** A computer programming language used to communicate with a database

R

**Range:** A collection of two or more cells in a spreadsheet

**Record:** A collection of related data in a data table, usually synonymous with row

**Reframing:** The process of restating a problem or challenge, then redirecting it toward a potential resolution

**Relational database:** A database that contains a series of tables that can be connected to form relationships

**Relevant question:** A question that has significance to the problem to be solved

**Report:** A static collection of data periodically given to stakeholders

**Return on investment (ROI):** A formula that uses the metrics of investment and profit to evaluate the success of an investment

**Revenue:** The total amount of income generated by the sale of goods or services

**Root cause:** The reason why a problem occurs

S

**Sample:** In data analytics, a segment of a population that is representative of the entire population

**Scope of work (SOW):** An agreed-upon outline of the tasks to be performed during a project

**Second-party data:** Data collected by a group directly from its audience and then sold

**Small data:** Small, specific data points typically involving a short period of time, which are useful for making day-to-day decisions

**SMART methodology:** A tool for determining a question’s effectiveness based on whether it is specific, measurable, action-oriented, relevant, and time-bound

**Social media:** Websites and applications through which users create and share content or participate in social networking

**Sorting:** The process of arranging data into a meaningful order to make it easier to understand, analyze, and visualize

**Specific question:** A question that is simple, significant, and focused on a single topic or a few closely related ideas

**Spreadsheet:** A digital worksheet

**SQL:** (Refer to Structured Query Language)

**Stakeholders:** People who invest time and resources into a project and are interested in its outcome

**String data type:** A sequence of characters and punctuation that contains textual information (Refer to Text data type)

**Structured data:** Data organized in a certain format such as rows and columns

**Structured Query Language:** A computer programming language used to communicate with a database

**Structured thinking:** The process of recognizing the current problem or situation, organizing available information, revealing gaps and opportunities, and identifying options

**SUM:** A spreadsheet function that adds the values of a selected range of cells

T

**Technical mindset:** The ability to break things down into smaller steps or pieces and work with them in an orderly and logical way

**Text data type:** A sequence of characters and punctuation that contains textual information (also called string data type)

**Third-party data:** Data provided from outside sources who didn’t collect it directly

**Time-bound question:** A question that specifies a timeframe to be studied

**Turnover rate:** The rate at which employees voluntarily leave a company

U

**Unfair question:** A question that makes assumptions or is difficult to answer honestly

**United States Census Bureau:** An agency in the U.S. Department of Commerce that serves as the nation’s leading provider of quality data about its people and economy

**Unstructured data:** Data that is not organized in any easily identifiable manner

V

**Video file:** A collection of images, audio files, and other data usually encoded in a compressed format such as MP4, MV4, MOV, AVI, or FLV

**Visualization:** (Refer to Data visualization)

W

**Wide data:** A dataset in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject

X

Y

Z

### 1.

Question 1

If you have a short time frame for data collection and need an answer immediately, you likely will have to use historical data.

**1 point**



True



False

### 2.

Question 2

Continuous data is measured and has a limited number of values.

**1 point**



False



True

### 3.

Question 3

Fill in the blank: The question “Where did you vacation last year?” is an example of collecting \_\_\_\_\_ data.

**1 point**



nominal quantitative



nominal qualitative



real quantitative



real qualitative

### 4.

Question 4

Fill in the blank: A data analyst is assembling a quarterly report for executives in their company. They evaluate their needs and determine they will only use \_\_\_\_\_ due to its reliability as well as the ability to easily collect it.

**1 point**



third party data



external data



structured data



internal data

### 5.

Question 5

Fill in the blank: A relational database is an example of \_\_\_\_\_ data.

**1 point**



open



closed



unstructured



structured

### 6.

Question 6

A Boolean data type must have a numeric value.

**1 point**



True



False

### 7.

Question 7

What do the columns contain in long data?

**1 point**



Different formats



The data types



The values and the context for the values



Specific constraints

### 8.

Question 8

Which of the following are examples of data transformation? Select all that apply.

**1 point**



Transforming data from wide format to long format



Saving data in a spreadsheet



Deleting data for security reasons



Changing a file type from .CSV to .XLS

Module 3

Hello again. So far, you've seen how

data can be gathered and analyzed to solve all

kinds of problems. 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. So, ready to tap into

the incredible power of databases? Let's go!

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.

As you can see, a table can only have one primary key but

it can have multiple foreign keys.

Understanding primary and foreign keys can be tricky, so

you'll have more opportunities to practice coming up.

But 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.

And 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.

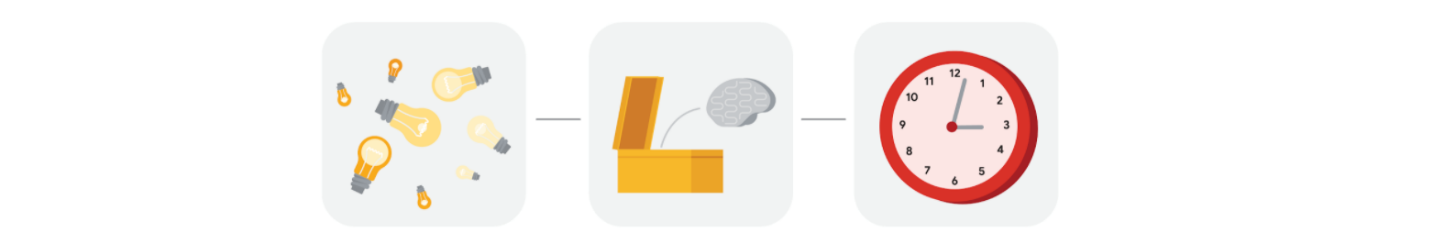
That will be a great opportunity to improve your understanding of primary and

foreign keys, database organization and

how you might use databases in your future analytics career.

**Maximize databases in data analytics**

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.

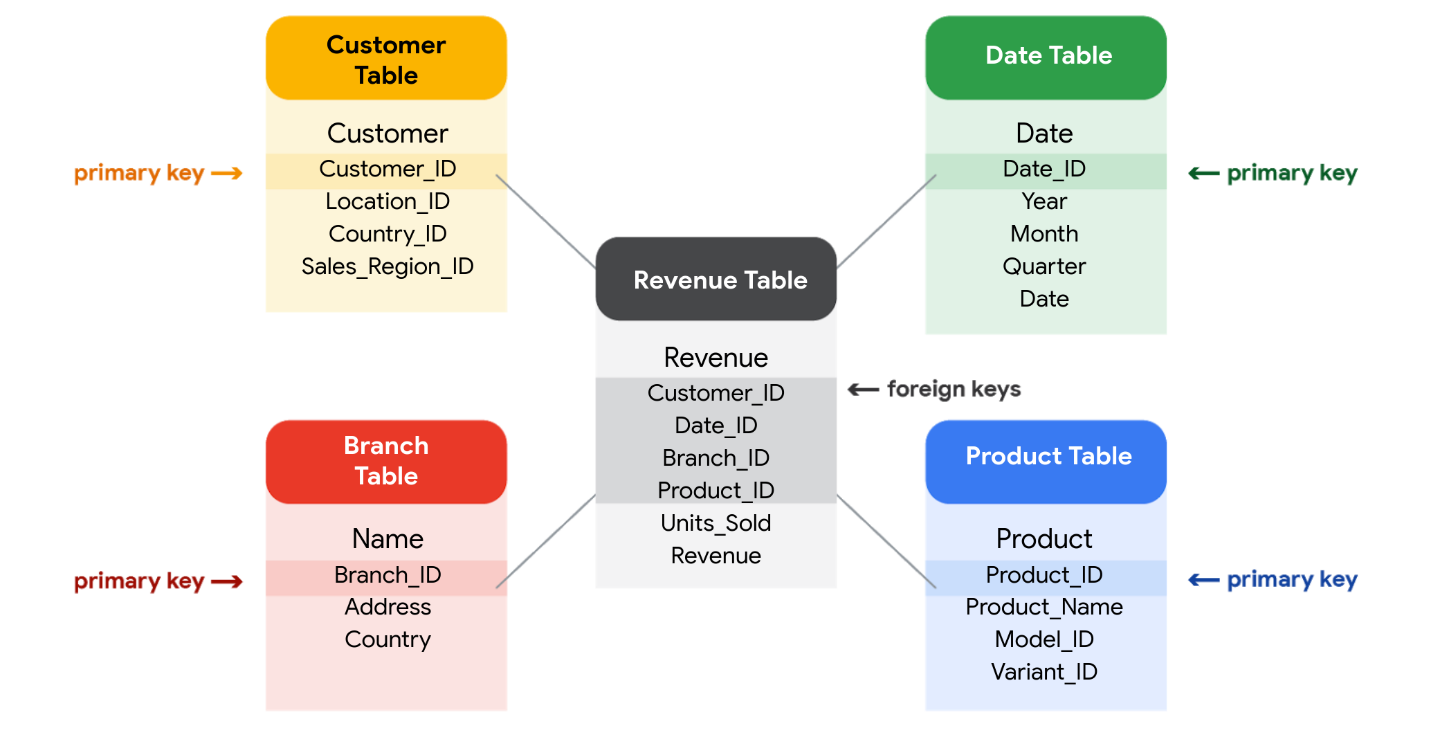
**Normalization** is a process of organizing data in a relational database. For example, creating tables and establishing relationships between those tables. It is applied to eliminate data redundancy, increase data integrity, and reduce complexity in a database.

**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

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=1706745600&Signature=TrO4UcoJhO6cT4aqKZnCNZL0El3ygjzHz9hVePyXyEWPHywVY1LKYCWSm1IuAMthUXd5Rqf~hwmMfhz82OoRQVhkbAKaoY-pG-SWsnNXX3RgrL4hx2Y-6~1XXKM4nJJcHKsbjKWpI3RnpJ6uaVqxCgLqr-Axy4uVrt8HbbTKqDw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/jmigEulNR7yooBLpTYe8Cw_9ecaf818f1a74b7987fe6a7d9af3c1f1_SalesByTemp.xlsx?Expires=1706745600&Signature=TrO4UcoJhO6cT4aqKZnCNZL0El3ygjzHz9hVePyXyEWPHywVY1LKYCWSm1IuAMthUXd5Rqf~hwmMfhz82OoRQVhkbAKaoY-pG-SWsnNXX3RgrL4hx2Y-6~1XXKM4nJJcHKsbjKWpI3RnpJ6uaVqxCgLqr-Axy4uVrt8HbbTKqDw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[SalesByDay](https://d3c33hcgiwev3.cloudfront.net/B3ofmLtERPq6H5i7RFT6Pg_1ca5eec9c08941518e2c16034a2e65f1_SalesByDay.xlsx?Expires=1706745600&Signature=Vl5v7qVl8YlNUoPIcQykRFJBHJEVKhp0OUcZfHY~QQhv7MgBXQ97fdOp~ocuyo-qY~iUu4rVabJ2fMvDWGf960bVp-vYlLmX8dWpRICCIMnIbSmsJkPPW0pbVQO1KqEyNMlIsy0rG2oFJiaYDJRKaCvz8~g8aQPB~8Pyji73ST0_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/B3ofmLtERPq6H5i7RFT6Pg_1ca5eec9c08941518e2c16034a2e65f1_SalesByDay.xlsx?Expires=1706745600&Signature=Vl5v7qVl8YlNUoPIcQykRFJBHJEVKhp0OUcZfHY~QQhv7MgBXQ97fdOp~ocuyo-qY~iUu4rVabJ2fMvDWGf960bVp-vYlLmX8dWpRICCIMnIbSmsJkPPW0pbVQO1KqEyNMlIsy0rG2oFJiaYDJRKaCvz8~g8aQPB~8Pyji73ST0_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[SalesByFlavor](https://d3c33hcgiwev3.cloudfront.net/DHN9hYWCSDCzfYWFgvgwgg_b0e0d35f6a4f4bde9c84ecd0dd69c0f1_SalesByFlavor.xlsx?Expires=1706745600&Signature=DzeZSjcHqGIFs3u7i4Ju79ynUFGpLyUjMpZY4O1Bh7vdSUYJHTHHB0Qd11xs~iEtggn3vtB2XTRKqnYb0UdEUAKbrqjVhs7WVQp8gq-I2mvccZ7J8JtFei9TLcUvTAfjkPgDXtO8xTJ~EUlYncaNCUg~8c8xH441gVzslZ7-4dU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/DHN9hYWCSDCzfYWFgvgwgg_b0e0d35f6a4f4bde9c84ecd0dd69c0f1_SalesByFlavor.xlsx?Expires=1706745600&Signature=DzeZSjcHqGIFs3u7i4Ju79ynUFGpLyUjMpZY4O1Bh7vdSUYJHTHHB0Qd11xs~iEtggn3vtB2XTRKqnYb0UdEUAKbrqjVhs7WVQp8gq-I2mvccZ7J8JtFei9TLcUvTAfjkPgDXtO8xTJ~EUlYncaNCUg~8c8xH441gVzslZ7-4dU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

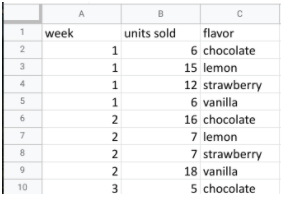


## Inspect the data

### 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.

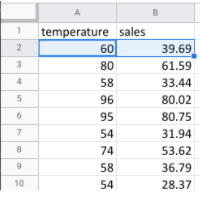
Click the **flavors** tab on your spreadsheet to view the relevant data. The **flavors** sheet has three columns and 209 rows of data. The column headers are **week**, **units sold**, and **flavor**. This dataset did not come with a data description, so you have to figure out the significance of the 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 **temperatures** 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).

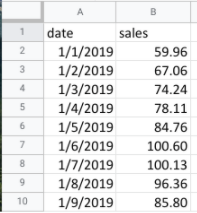


So, which is it? It’s probably a daily snapshot because there are 365 entries for temperature, and multiple rows with the same temperature and different sales values. This implies that each entry is for a single day and not a summary of multiple days. However, without more information, you can’t be certain. 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.

### 1.

Question 1

Fill in the blank: A relational database contains a series of \_\_\_\_\_ that can be connected to form relationships.

1 point

tables

cells

fields

spreadsheets

### 2.

Question 2

What is the term for an identifier that references a database column in which each value is unique?

1 point

Foreign key

Primary key

Relation

Field

### 3.

Question 3

What process do data professionals use to eliminate data redundancy, increase data integrity, and reduce complexity in a database?

1 point

Iteration

Composition

Manipulation

Normalization

### 4.

Question 4

Fill in the blank: When using a relational database, data analysts write \_\_\_\_\_ to request data from the related tables.

1 point

programs

keys

queries

relationships

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."

Play video starting at ::41 and follow transcript0:41

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.

Pretty amazing, right? Here's 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."

Play video starting at :1:13 and follow transcript1:13

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:

descriptive, structural, and administrative.

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.

Play video starting at :2:27 and follow transcript2:27

It would also include the book's author and title.

Next is structural metadata, which

is metadata that 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.

Finally, we have administrative metadata.

Administrative metadata 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.

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

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.

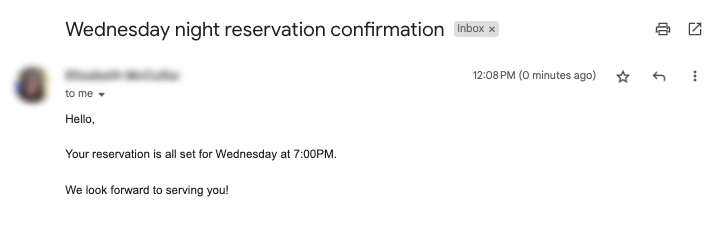
### **Photos**

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.

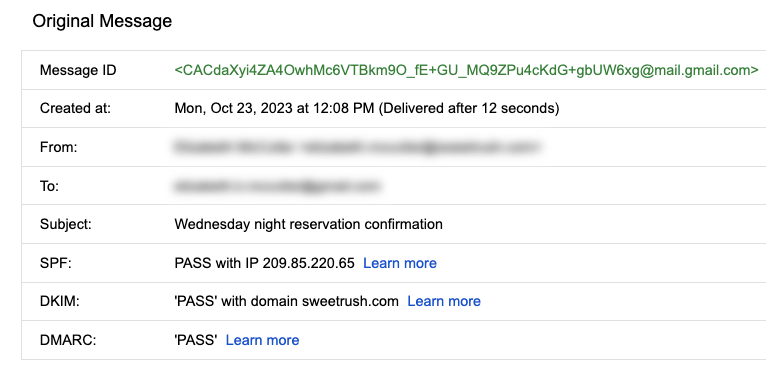
Image with accompanying Information pop-up that displays its description, the date and time the image was taken, its size, the device on which it was taken, and an option to add the geolocation of the image.

### **Emails**

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

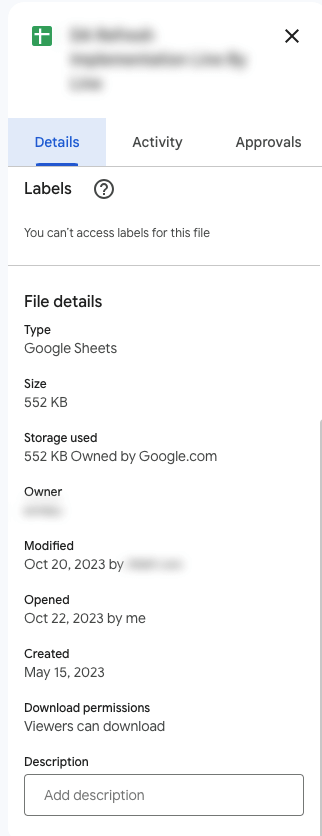


Emails also contain hidden metadata that includes server names, IP addresses, HTML format, and software details. This image includes hidden email metadata such as the message ID and when the email was created.

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

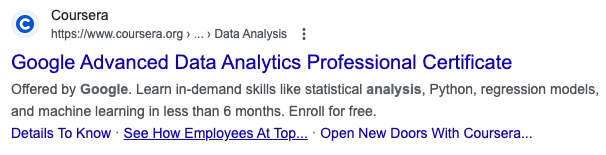
### **Spreadsheets and electronically created documents**

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.

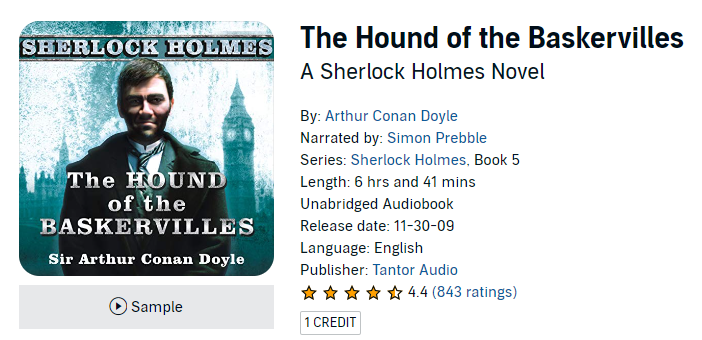
### **Websites**

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

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.

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.

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.

Play video starting at :1:15 and follow transcript1:15

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 of 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.

[SOUND]

### 1.

Question 1

A large company has several databases across its many departments. What kind of metadata describes how many locations contain a certain piece of data?

1 point

Descriptive

Representative

Structural

Administrative

### 2.

Question 2

A large metropolitan high school gives each of its students an ID number to differentiate them in its database. What kind of metadata are the ID numbers?

1 point

Structural

Descriptive

Administrative

Representative

### 3.

Question 3

An international nonprofit organization wants to merge third-party data with its own data. Which of the following actions will help make this process successful? Select all that apply.

1 point

Replace the incoming data’s metadata with its own company metadata.

Use metadata to evaluate the third-party data’s quality and credibility.

Use metadata to standardize the datasets.

Alter the internal metadata to more closely reflect the incoming metadata.

### 4.

Question 4

Fill in the blank: Data \_\_\_\_\_ is a process data professionals use to ensure the formal management of their organization’s data assets.

1 point

governance

organization

storage

sourcing

In this video, 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 learn that there are two basic types of data

used by data analysts: internal and external.

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.

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.

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.

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.

Coming up, we'll learn how to import all the data you

collect from different sources into a spreadsheet.

# 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 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.

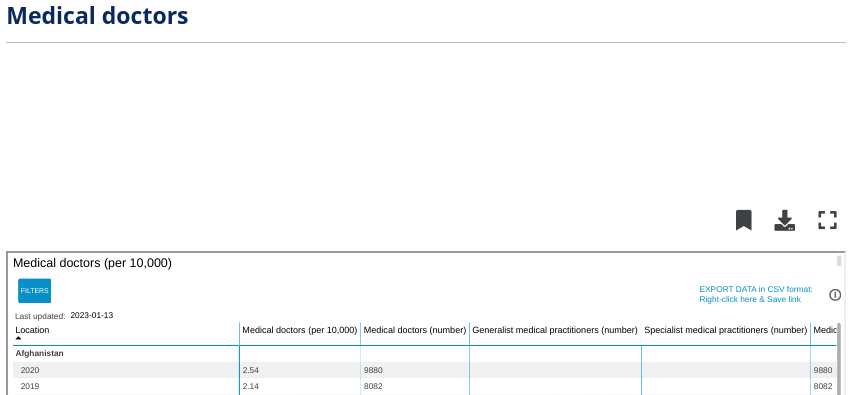
### Example 2: Download data from the Global Health Observatory

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.
9. At this point, you've learned all about internal and external data, and
10. how to prepare it for use.
11. Now, we'll go through the process of actually importing data from
12. different sources.
13. Sometimes you want to upload a spreadsheet from your files, such as a CSV file.
14. CSV stands for comma-separated values.
15. A CSV file saves data in a table format.
16. Now, let's bring that file into a fresh spreadsheet.
17. Play video starting at ::29 and follow transcript0:29
18. We'll start by selecting a file, then import.
19. Play video starting at ::37 and follow transcript0:37
20. Then we'll choose to upload a file.
21. Play video starting at ::41 and follow transcript0:41
22. Navigate to it, open it and insert it as a new sheet.
23. Play video starting at ::52 and follow transcript0:52
24. CSV files use plain text and they're delineated by characters.
25. So each column or field is clearly distinct from another when importing.
26. As you learned, CSVs are comma-separated, and
27. usually the spreadsheet app will auto-detect those separations.
28. But sometimes, you might need to indicate that the separator is another character or
29. a space by selecting the different options in this window.
30. Play video starting at :1:15 and follow transcript1:15
31. Also, if you are planning to work with the data set,
32. you would usually convert to text, numbers or other options here.
33. But plain text is okay for reporting purposes.
34. So we can leave those fields alone.
35. Finally, select Import data.
36. Play video starting at :1:33 and follow transcript1:33
37. Now our CSV file's ready to work with in our spreadsheet.
38. I spend most of my time at work analyzing spreadsheets full of healthcare
39. information.
40. I typically start by looking at a larger data set.
41. Then I pull a subset of it into a spreadsheet so I can work with it.
42. Maybe I want to analyze year over year growth in user demand on Google Search for
43. certain healthcare services, like telemedicine.
44. Or maybe I want to look at data sets from external healthcare organizations or
45. agencies for more insight into this trend.
46. For example, with telemedicine,
47. maybe I'll look at a spreadsheet that lists telemedicine providers.
48. There are so many ways spreadsheets can help you find the insights you need.
49. One source I use a lot is the World Health Organization's data repository.
50. Play video starting at :2:16 and follow transcript2:16
51. This is a place where anybody can access open-source data.
52. As you can see, there's tons of data available.
53. You can search by theme, category, indicator and country.
54. You can also access World Health Organization metadata if you want to learn
55. more about the data in the repository.
56. Play video starting at :2:35 and follow transcript2:35
57. For our example, we'll look at medical doctors by country and year.
58. Play video starting at :2:42 and follow transcript2:42
59. This information would be useful for
60. a data analysis project looking into how many doctors are available to treat
61. patients within a certain population compared to other populations.
62. To get this data, we'll start on this webpage, which contains the data set we
63. want. Then we'll download the data as a CSV file.
64. Play video starting at :3:5 and follow transcript3:05
65. Then open a new spreadsheet and import the file by selecting File, Import.
66. Play video starting at :3:18 and follow transcript3:18
67. Next, upload your file and select Import Data.
68. Play video starting at :3:33 and follow transcript3:33
69. After reviewing the data to make sure it looks clean, we can title it and
70. begin our work.
71. Play video starting at :3:39 and follow transcript3:39
72. I know this is a lot of information to take in, but
73. you'll get much more comfortable with this the more you practice.
74. Coming up, we'll learn how to sort and
75. filter your data to focus on the information relevant to you.

# 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.

A screen shot of a computer

Description automatically generated

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.

# xplore 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.

### 1.

Question 1

What are some key benefits of open-data initiatives? Select all that apply.

1 point

Help educate citizens about important issues

Support innovation and economic growth

Limit opportunities for collaboration

Make government activities more transparent

### 2.

Question 2

What type of file saves data in a table format?

1 point

Compatible scientific variables (.csv)

Cell-structured variables (.csv)

Calculated spreadsheet values (.csv)

Comma-separated values (.csv)

### 3.

Question 3

Bringing data from a .csv file into a spreadsheet is an example of what process?

1 point

Filing data

Normalizing data

Importing data

Editing data

### 4.

Question 4

In Google Sheets, what function enables a data analyst to specify a range of cells in one spreadsheet to be duplicated in another?

1 point

**CELLRANGE**

**SPECIFY**

**IMPORTRANGE**

**DUPLICATE**

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.

Play video starting at :1:38 and follow transcript1:38

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.

Play video starting at :1:55 and follow transcript1:55

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.

Play video starting at :2:7 and follow transcript2:07

Looks good to me. Now let's sort the entire spreadsheet.

We'll sort by city first.

To do this, select the city column,

Play video starting at :2:19 and follow transcript2:19

then use the drop-down arrow to sort the sheet.

Select A to Z.

Play video starting at :2:27 and follow transcript2:27

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,

Play video starting at :3:4 and follow transcript3:04

then choose data and sort range.

Play video starting at :3:11 and follow transcript3:11

In the dialog box,

make sure that "Data has header row" is highlighted.

Play video starting at :3:18 and follow transcript3:18

That way row A, city,

states, sales rep, and

auto parts won't be part of the sort.

Play video starting at :3:28 and follow transcript3:28

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 the sort order A to Z.

Play video starting at :3:48 and follow transcript3:48

Finally, select Sort.

Play video starting at :3:54 and follow transcript3:54

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.

Play video starting at :5:2 and follow transcript5:02

Then select okay.

Play video starting at :5:7 and follow transcript5:07

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.

Play video starting at :5:22 and follow transcript5:22

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.

### 1.

Question 1

What is the process for arranging data into a meaningful order to make it easier to understand, analyze, and visualize?

1 point

Filtering

Reframing

Sorting

Prioritizing

### 2.

Question 2

A data analyst is reviewing a national database of real estate sales. They are only interested in sales of condominiums. How can the analyst narrow their scope?

1 point

Filter out condominium sales

Sort by non-condominium sales

Filter out non-condominium sales

Sort by condominium sales

### 3.

Question 3

A data analyst works for a rental car company. They have a spreadsheet that lists car ID numbers and the dates cars were returned. How should they sort the spreadsheet to find the most recently returned cars?

1 point

By return date, in descending order

By car numerical ID, in ascending order

By return date, in ascending order

By car numerical ID, in descending order

### 4.

Question 4

Fill in the blank: To keep a header row at the top of a spreadsheet, highlight the row and select \_\_\_\_\_ from the View menu.

1 point

lock

set

pin

freeze

Hi and welcome back.

Throughout this course, you've seen how Big Query can be

used to view and analyze data from tons of sources.

Now we're going to explore

the different account 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

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.

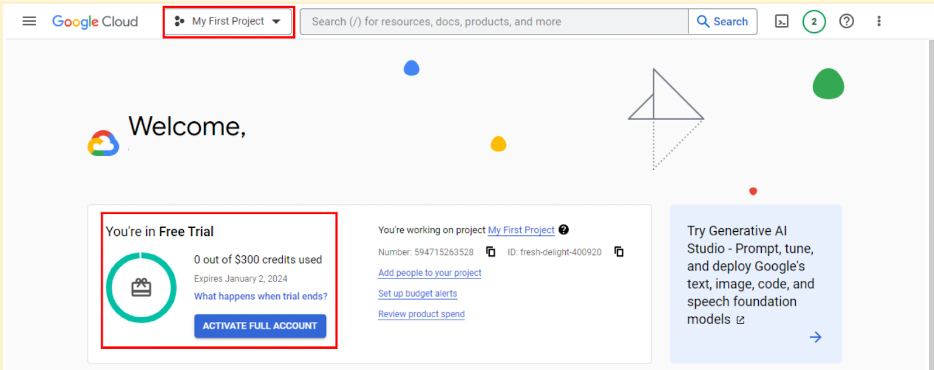
### **Free-of-charge trial**

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.

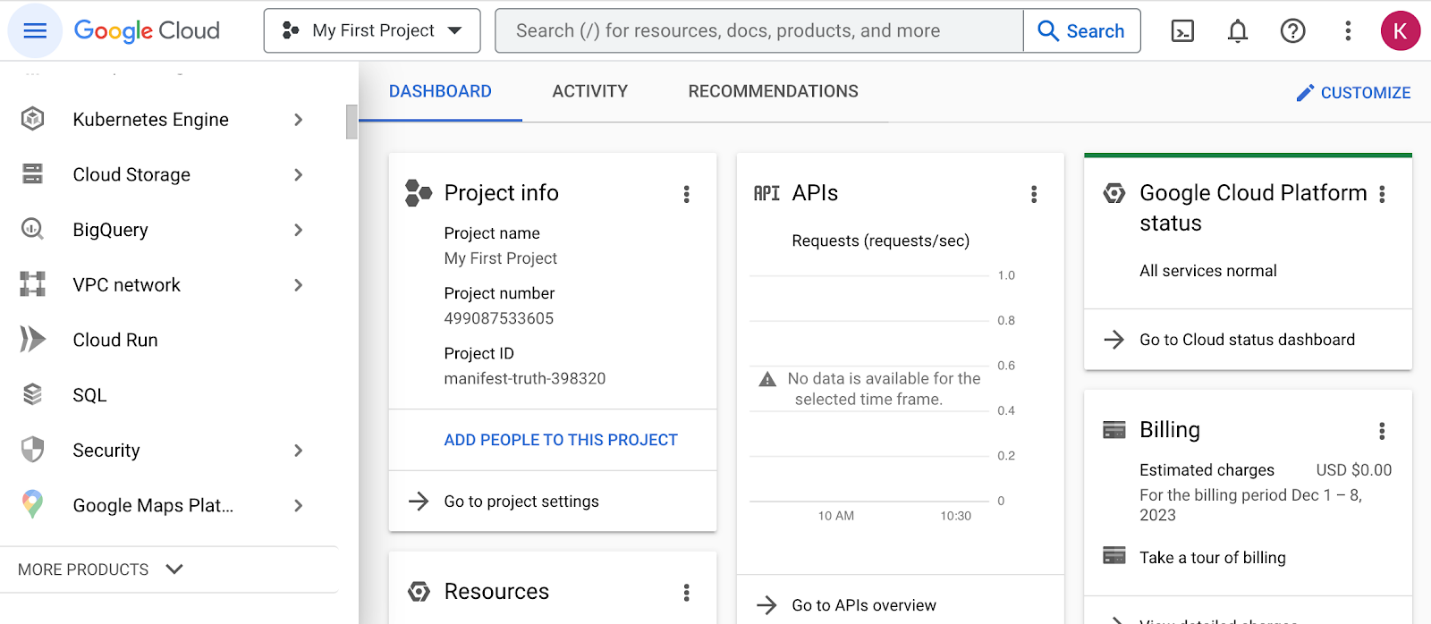
Choose the right BigQuery account type to match your specific needs and adapt as your requirements change!

# Get started with BigQuery

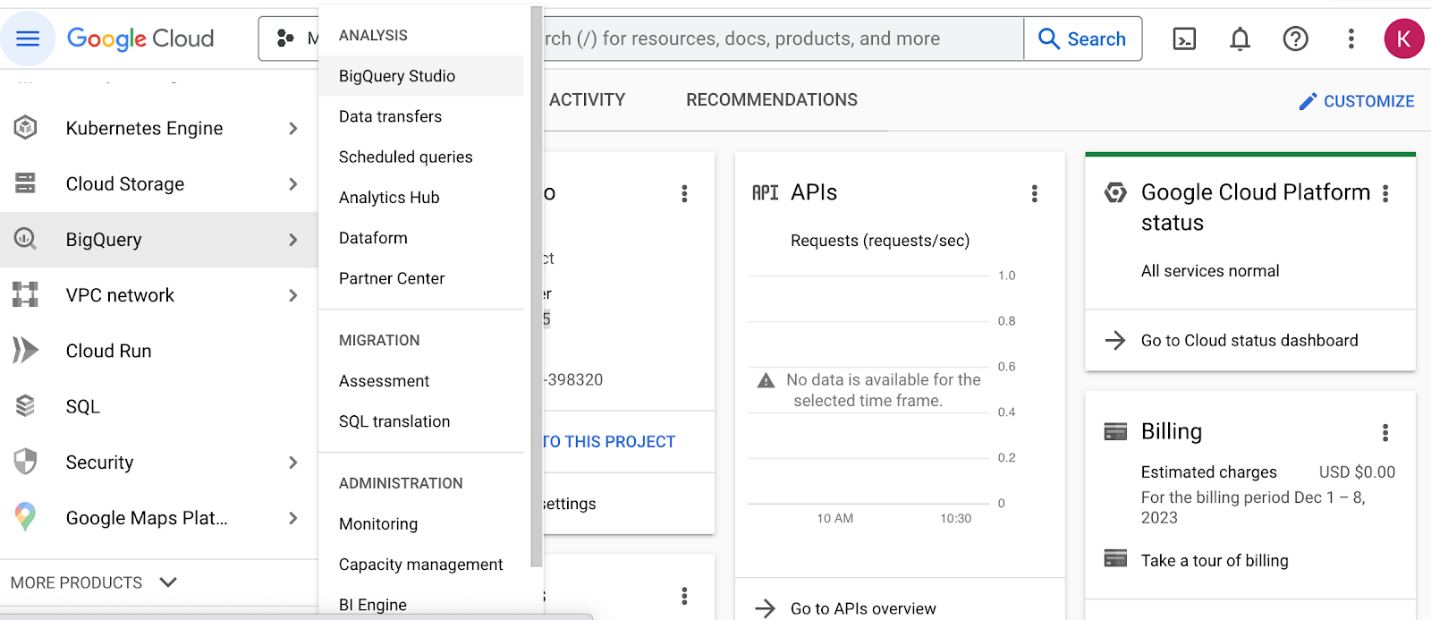
BigQuery is a data warehouse on the Google Cloud Platform used to query and filter large datasets, aggregate results, and perform complex operations. Throughout this program, you’re going to use BigQuery to practice your SQL skills and collect, prepare, and analyze data. At this point, you have set up 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.

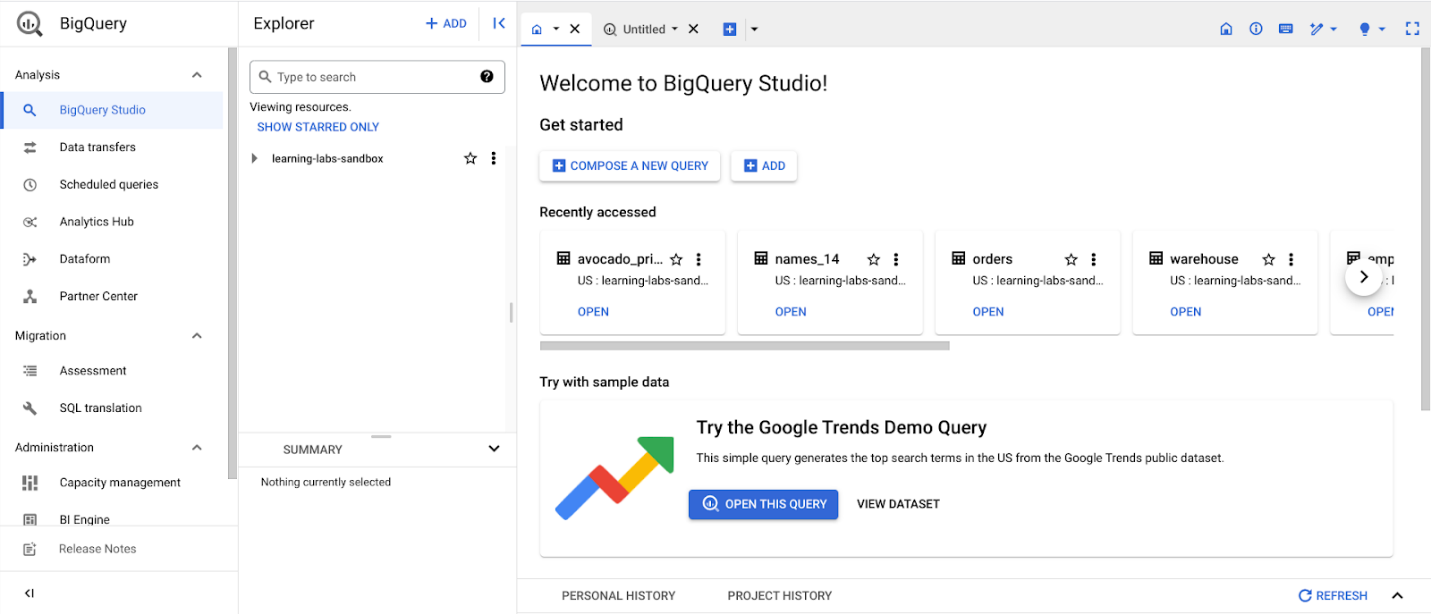


Navigate to your project’s BigQuery Studio by selecting BigQuery from the navigation menu and BigQuery Studio from the dropdown menu.



## BiqQuery 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**.

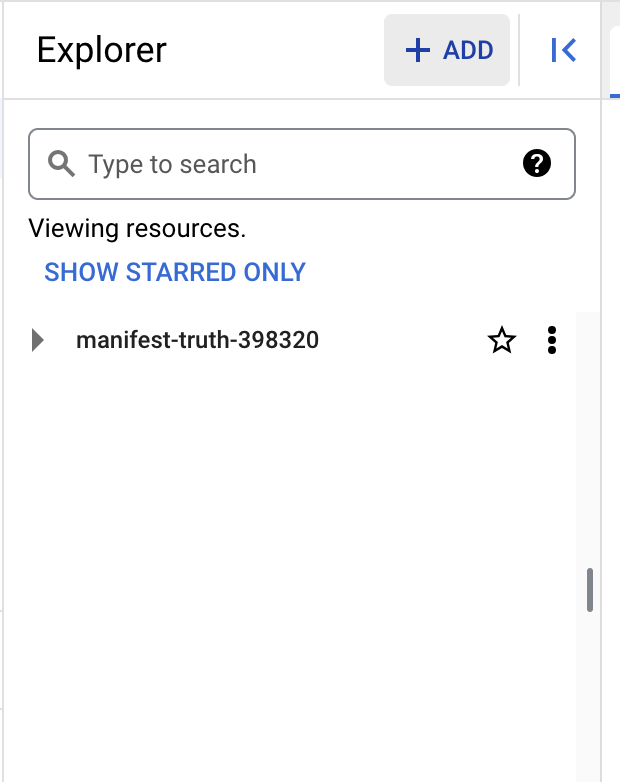


### The Navigation pane

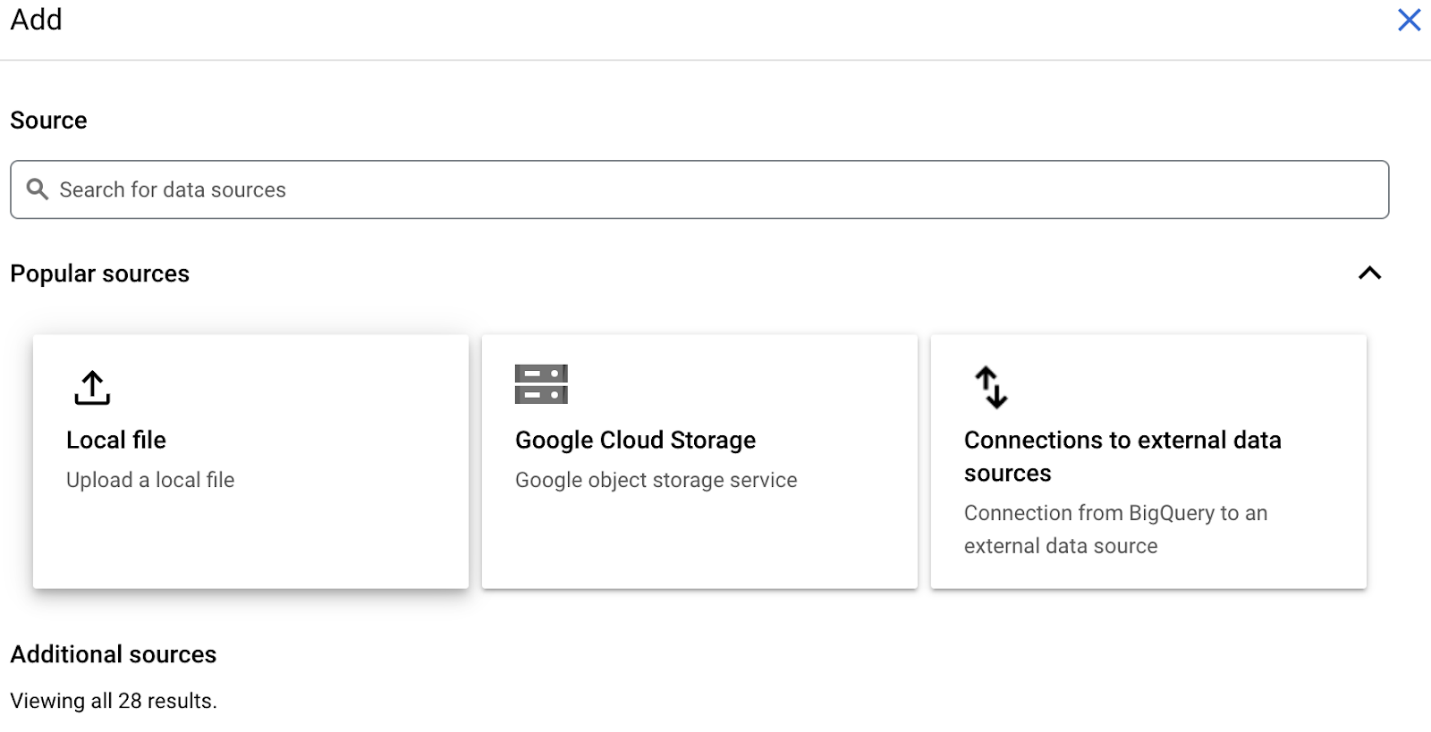
On the console page, find the **Navigation** pane. This is how you navigate from the project space to the BigQuery tool. This menu also contains a list of other Google Cloud Project (GCP) data tools. During this program, you will focus on BigQuery, but it’s useful to understand that the GCP has a collection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

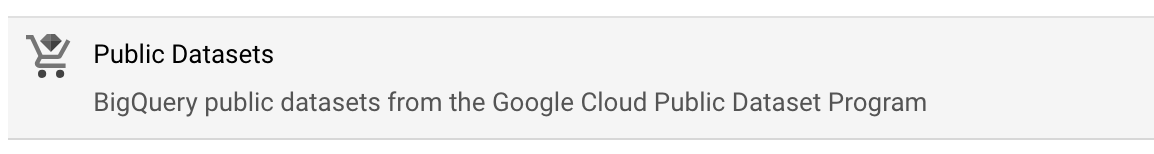


This button opens the **Add** dialog that allows you to open or import a variety of datasets.

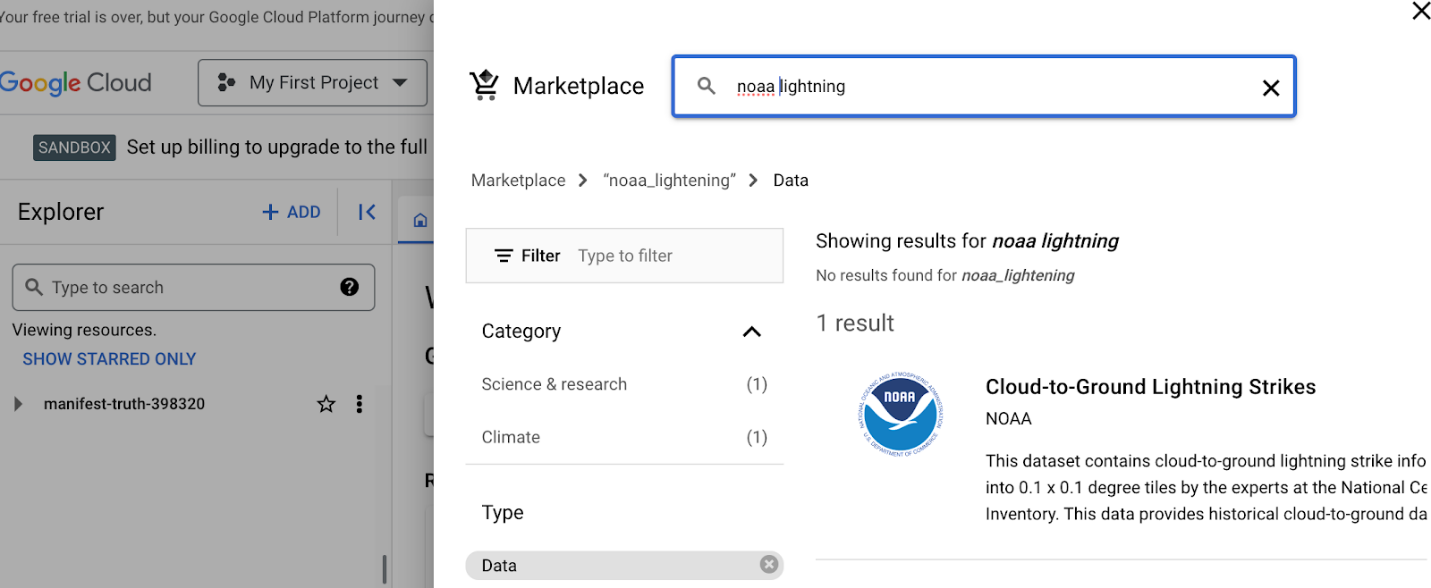


### Add Public Datasets

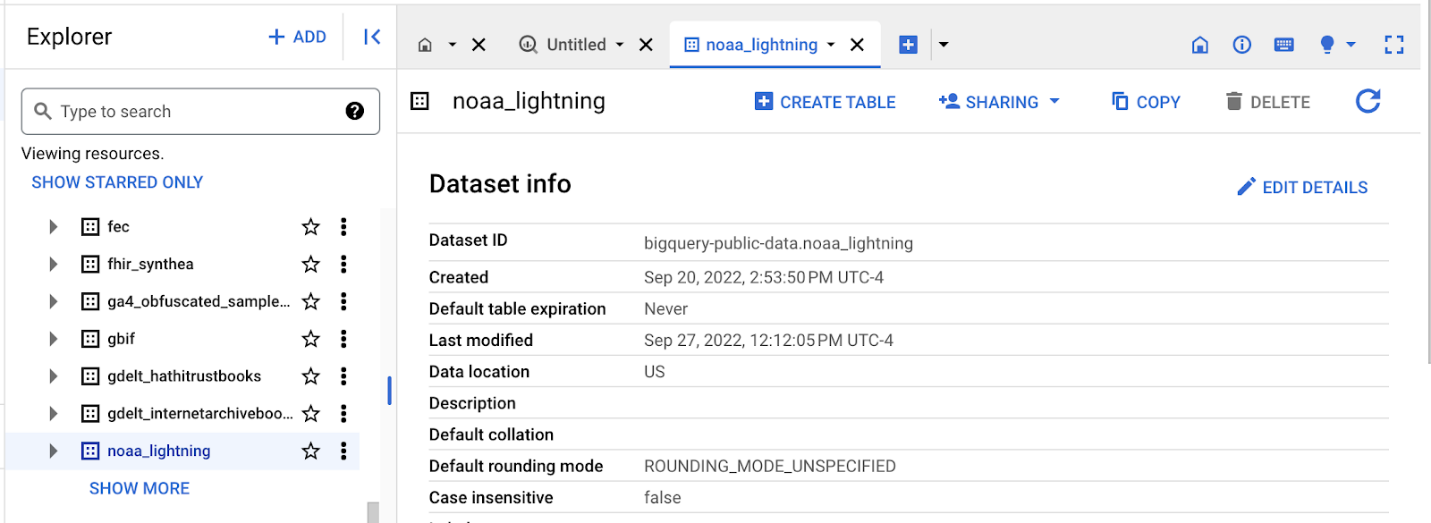
BigQuery offers a variety of public datasets from the Google Cloud Public Dataset Program. Scroll down the **Add** dialog to the **Public Datasets** option.



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.

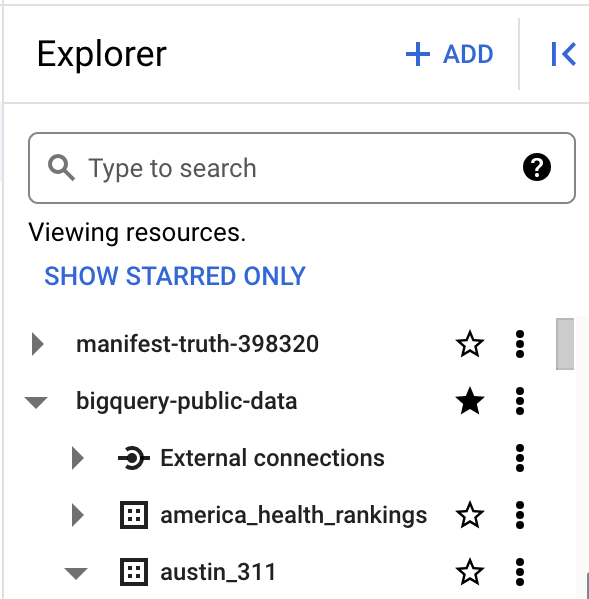


Select the dataset to read its description. Select **View dataset** to create a tab of the dataset’s information within the SQL workspace.

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.

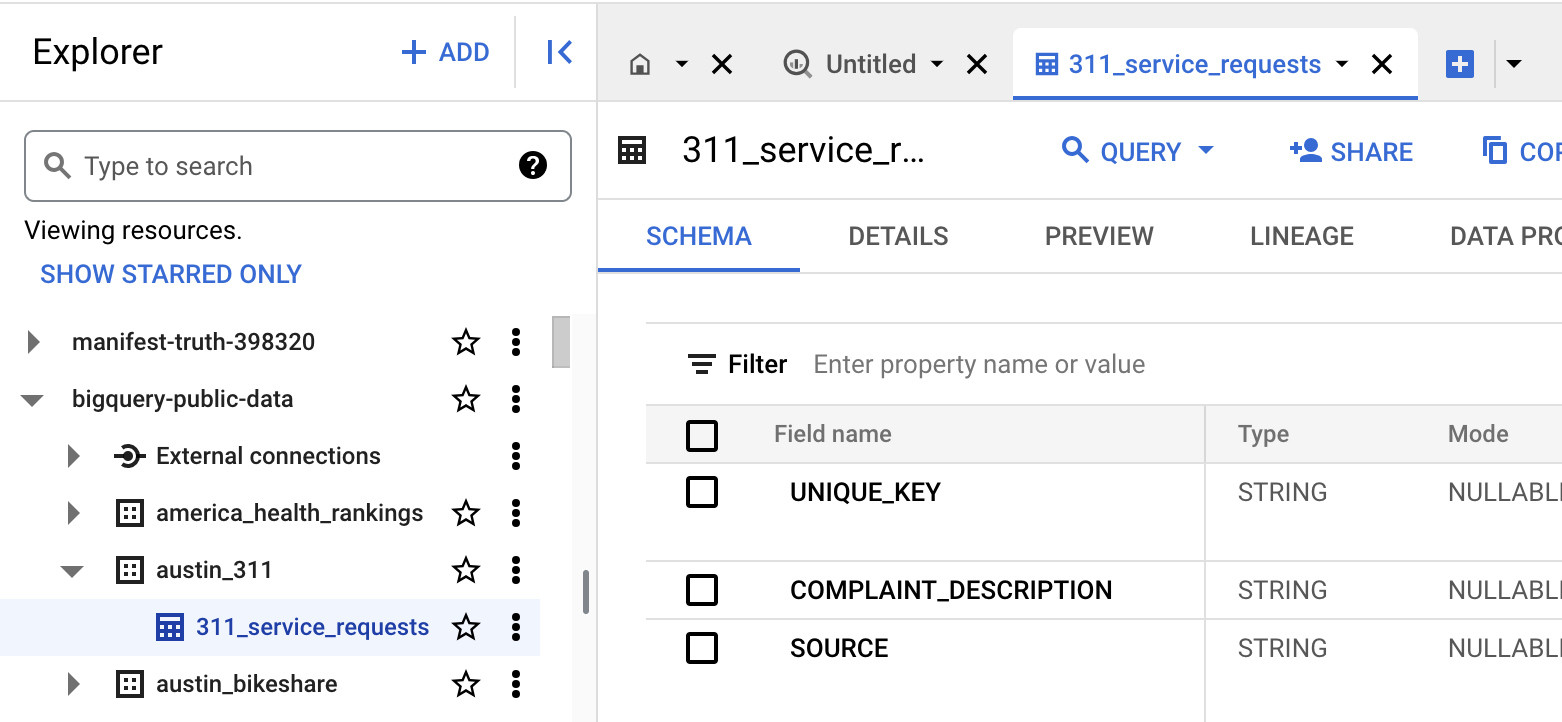
For example, you might want to select a different public dataset. If you select the second dataset, "austin\_311," it will expand to list the table stored in it, “311\_service\_requests.”

A screenshot of a search engine

Description automatically generatedThe 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:

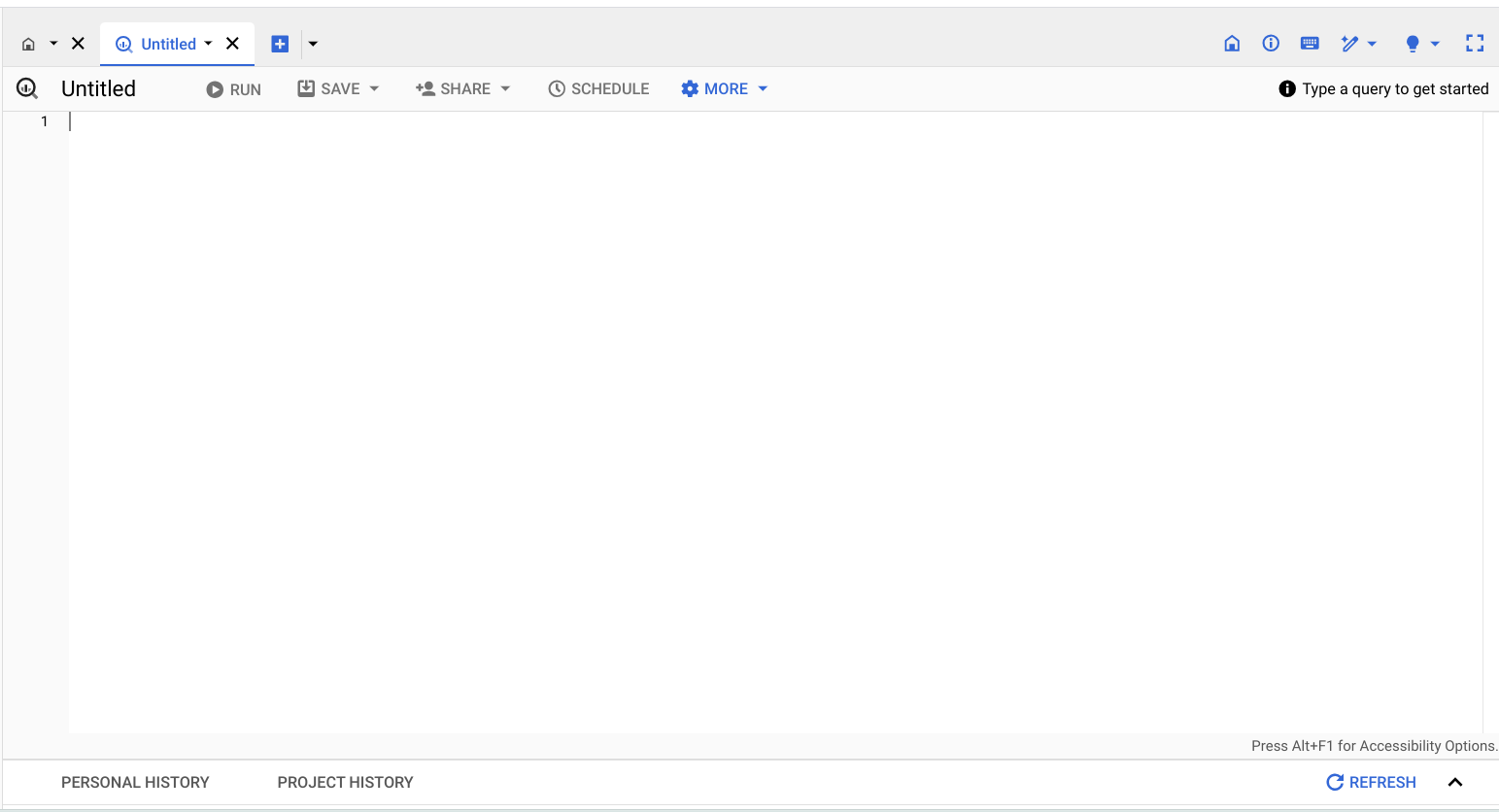
* **Schema**, which displays the column names in the dataset
* **Details**, which contains additional metadata, such as the creation date of the dataset
* **Preview**, which shows the first rows from the dataset



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

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:

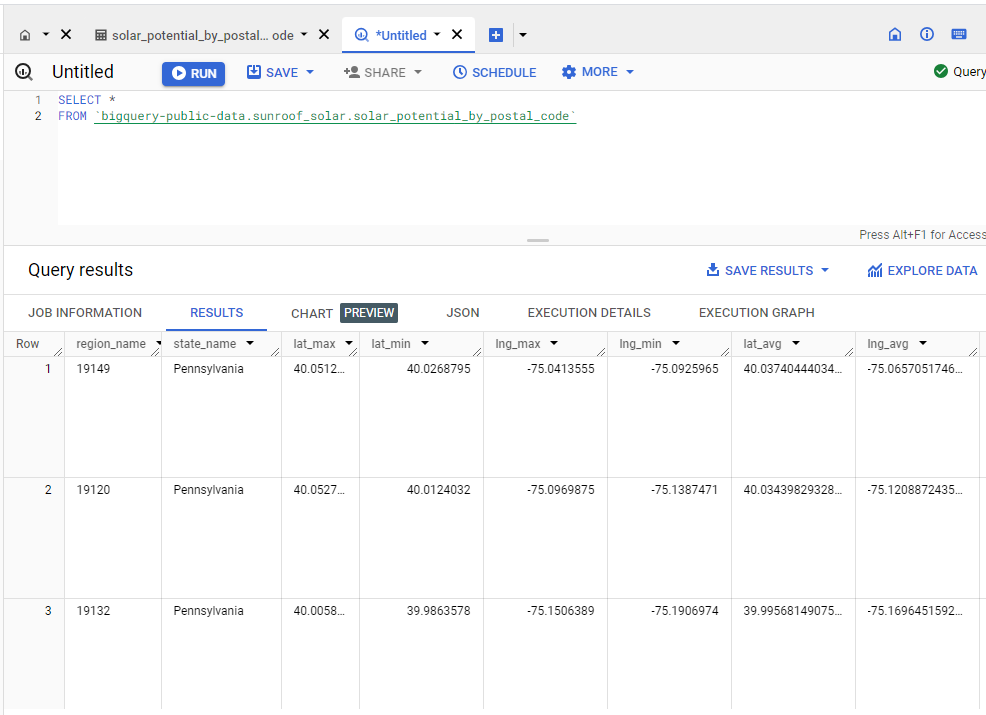
1. Log in to [BigQuery](https://cloud.google.com/bigquery?utm_source=google&utm_medium=cpc&utm_campaign=na-none-all-en-dr-sitelink-all-all-trial-b-gcp-1605212&utm_content=text-ad-none-any-DEV_c-CRE_665665924750-ADGP_Hybrid+%7C+BKWS+-+MIX+%7C+Txt_BigQuery-KWID_43700077225652791-kwd-33969409261-userloc_9030075&utm_term=KW_bigquery-ST_bigquery-NET_g-&gclid=Cj0KCQjw1aOpBhCOARIsACXYv-fA3H67Qgz-5mnVoHtT0fQF2s9qDD3fO89K00t_yzrJ1DRSafdvHDsaAhGCEALw_wcB&gclsrc=aw.ds) 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 name of the dataset. Select the **ellipses** by the dataset **solar\_potential\_by\_postal\_code**, then select **Query**. A new tab will populate on your screen. Select the tab. The name of the dataset should be written inside the two backticks.
2. Select the dataset name by highlighting the text including the backticks and copy it.
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 name of the dataset 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.



## 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.
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.

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!

You've learned how sorting and filtering data in

spreadsheets helps

data analysts customize the information.

Customizing data makes it more meaningful and

easier to understand, analyze, and visualize.

You also discovered that some spreadsheets

can be extremely long and complex.

Knowing how to zero in on the exact data you need while

setting aside the rest helps you focus on your analysis.

This is also true for databases.

Sometimes a data set is too large

to download or it won't fit in a spreadsheet,

so a data analyst will use SQL to create a query to

view the specific data that

they want from within the larger set.

We've learned that a database

is a collection of data stored in

a computer system and that

SQL stands for a Structured Query Language.

Data analysts use query languages

to communicate with the database.

In an earlier video, you also

learned that a relational database

contains a series of tables that can

be connected to foreign relationships.

These relationships are represented

by primary and foreign keys.

Data analysts write queries

in order to get data from these tables.

Let's see how this works.

We'll start with our Table Viewer.

Here we can see what public data-sets are available.

We'll scroll through the data before we start using it to

get a feel for what it's all

about and to make sure it's clean.

Play video starting at :1:22 and follow transcript1:22

Some Table Viewers let you preview

a few rows before even writing a query.

This is helpful if you want to take a quick look

to be sure the data set will be right for your project.

To show you how this works, let's

check out a sample data set.

This one shows how much sunlight hits rooftops in a year.

This would be very

useful for a data analyst working on

a solar energy project, for example.

We'll start by previewing the data set.

Then we'll select a subset of this data,

where we find regions,

states, yearly sunlight, and more.

Now to see the entire data set, let's write a query.

The first step is finding out

the complete correct name of the data set.

To do this, select the data set,

solar potential by postal code and select query table.

Play video starting at :2:15 and follow transcript2:15

The name of the data set is

shown inside the two backticks.

This is to help us read the query more easily.

We can also remove the backticks in

this case and our query would still run.

The words you see before the dot represent

the database name and

the words after the dot represent the table name.

Let's select and copy

the data set name now because we'll need it in a second.

Play video starting at :2:43 and follow transcript2:43

Now we'll click on the plus sign to compose a new query.

Play video starting at :2:49 and follow transcript2:49

Most queries begin with the word

select then we add a space.

Because we want to see the entire data set,

we'll put an asterisk next.

The asterisk says, we want to include all columns.

This is a great shortcut because without it,

we'd have to type in every single field name.

Next, we'll press "Return" and type FROM.

Play video starting at :3:17 and follow transcript3:17

FROM does just what it sounds like,

it indicates where the data is coming from.

After that, we'll add another space.

Now we paste in the name of the data set that we copied

earlier and finally run the query.

Play video starting at :3:39 and follow transcript3:39

Now we can carefully inspect

the data set before we begin working with it.

One important thing to keep in mind,

SQL queries can be written in a lot of different ways,

but still provide the same results.

For example, we could have written

this query as one long line of instructions like

this and we'd still get the same results.

Play video starting at :4:6 and follow transcript4:06

The additional lines and spaces

don't impact the query's outcome

but they keep your query organized and

easier to read for yourself and others.

Now, if the project doesn't require all of these fields,

we can use SQL to view

a particular piece or pieces of data.

To do this, we specify

a certain column name in the query.

For example, maybe we only

want to see data from Pennsylvania.

We'd begin our query the same way we just learned.

SELECT, space, add an asterisk.

Then FROM our solar potential database

but this time we'll add where.

Play video starting at :4:46 and follow transcript4:46

Add a space and state underscore name

, the name of the column.

Play video starting at :4:55 and follow transcript4:55

Now because we only want to see data from Pennsylvania,

we add an equal sign and the word Pennsylvania with

single quotes around it.

In SQL, single quotes

indicate the beginning and ending of a string.

Finally, we run the query.

Play video starting at :5:18 and follow transcript5:18

Now we can review the data on

solar potential for only Pennsylvania.

Now we've got the data we want and

we're ready to start putting it to work,

which we'll cover later on.

But for now, let's celebrate finishing another module.

You've covered a lot of complex

and highly technical information.

As you keep practicing though,

things will start to feel a lot more natural.

For now, take a moment to sit

back and think about all you've learned.

You discovered metadata and how it keeps

data organized by describing what that data is all about.

You've seen how internal and

external data are accessed and how

data analysts use them to find

compelling insights to solve business problems.

You can sort and filter your data

to really pinpoint the information you need.

Finally, you just learned about

queries and you even practice writing some.

Coming up you'll have a few readings and

then a weekly challenge to test your knowledge.

This will help you confirm that you've understood what

we've worked on in these videos and as always,

if you're ever unsure about a question,

I highly encourage you to review

the videos and readings to find the answer.

You're the data detective now, so use those skills.

Keep up the great work and I'll

see you after the weekly challenge.

# In-depth guide: SQL best practices

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=1706832000&Signature=Ncd-YF~~g06aR1NWjFNONNz5e0UE9cYvyA3alA3bjB34a2dBjrpMT9nesNKDHiPGWXZfr5deM62u9W-FdpgUIM8FkFiKYiiBZi-6GUccnjvkt6NHwPqz6KWf4YWSy2rSR565EaUh0fpUbb0Q9szM9fzRTkMaASy4M--Y1tsiw1E_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[PDF File](https://d3c33hcgiwev3.cloudfront.net/UwaGyGQoRLu9Dw_8BLUTiQ_4d31c4c09dc54520835973c4ac8240f1_DAC3-In-depth-guide_-SQL-best-practices.pdf?Expires=1706832000&Signature=Ncd-YF~~g06aR1NWjFNONNz5e0UE9cYvyA3alA3bjB34a2dBjrpMT9nesNKDHiPGWXZfr5deM62u9W-FdpgUIM8FkFiKYiiBZi-6GUccnjvkt6NHwPqz6KWf4YWSy2rSR565EaUh0fpUbb0Q9szM9fzRTkMaASy4M--Y1tsiw1E_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

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 of favorite foods 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

2

3

4

5

6

SELECT

      Friend

FROM

      FavoriteFoods

WHERE

      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

2

3

4

5

6

SELECT

      Friend

FROM

      FavoriteFoods

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:

1

2

3

4

5

6

7

--This is an important query used later to join with the accounts table

SELECT

      rowkey, -key used to join with account\_id

      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:

1

2

3

4

5

6

7

SELECT

      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 rename SUM(tickets) as just 'total'. Because of the space, SQL will use 'total' as the name and won’t understand what you mean by '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 tickets\_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:

1

2

3

4

5

6

7

8

SELECT

CASE WHEN genre = 'horror' THEN 'Will not watch' WHEN genre = 'documentary'

THEN 'Will watch alone' ELSE 'Watch with others' END AS

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:

1

2

3

4

5

6

7

8

9

10

SELECT

    CASE

        WHEN genre = 'horror' THEN 'Will not watch'

        WHEN genre = 'documentary' THEN 'Will watch alone'

        ELSE 'Watch with others'

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

FROM

    MovieTheater

GROUP BY

    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:

1

2

3

4

5

6

7

-- Date: September 15, 2020

-- Analyst: Jazmin Cisneros

-- Goal: Count the number of rows in the table

SELECT

    COUNT(\*) number of rows -- the \* stands for all so count all

FROM

    table

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

1

2

3

4

5

6

7

8

/\* Date: September 15, 2020

Analyst: Jazmin Cisneros

Goal: Count the number of rows in the table

\*/

SELECT

    COUNT(\*) number of rows -- the \* stands for all so count all

FROM

    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:

A black screen with white text

Description automatically generated

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:

A black screen with white text

Description automatically generated

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:

A black screen with white text

Description automatically generated

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/)

### 1.

Question 1

Fill in the blank: When using SQL, the \_\_\_\_\_ clause can be used to filter a dataset of customers to only include people who have made a purchase in the past month.

1 point

**FROM**

**WHERE**

**SELECT**

**FILTER**

### 2.

Question 2

Which cases are most often used for column names in a database table and represent a SQL best practice? Select all that apply.

1 point

Sentence case

Camel case

Lowercase

Snake case

### 3.

Question 3

A database table is named WebTrafficAnalytics. What type of case is this?

1 point

Sentence case

Camel case

Snake case

Lowercase

### 4.

Question 4

What can be removed from the following query without preventing it from running?

|  |
| --- |
| **SELECT \***  **FROM `Uni\_dataset.new\_table`**  **WHERE ID = 'Lawrence'** |

1 point

**WHERE**

**SELECT**

The asterisk (**\***)

Backticks (**`**)

# Glossary terms from module 3

## ****Terms and definitions for Course 3, Module 3****

**Administrative metadata:** Metadata that indicates the technical source of a digital asset

**CSV (comma-separated values) file:** A delimited text file that uses a comma to separate values

**Data governance:** A process for ensuring the formal management of a company’s data assets

**Descriptive metadata:** Metadata that describes a piece of data and can be used to identify it at a later point in time

**Foreign key:** A field within a database table that is a primary key in another table (Refer to primary key)

**FROM:** The section of a query that indicates where the selected data comes from

**Geolocation:** The geographical location of a person or device by means of digital information

**Metadata:** Data about data

**Metadata repository:** A database created to store metadata

**Naming conventions:** Consistent guidelines that describe the content, creation date, and version of a file in its name

**Normalized database:** A database in which only related data is stored in each table

**Notebook:** An interactive, editable programming environment for creating data reports and showcasing data skills

**Primary key:** An identifier in a database that references a column in which each value is unique (Refer to foreign key)

**Redundancy:** When the same piece of data is stored in two or more places

**Schema:** A way of describing how something, such as data, is organized

**SELECT:** The section of a query that indicates the subset of a dataset

**Structural metadata:** Metadata that indicates how a piece of data is organized and whether it is part of one or more than one data collection

**WHERE:** The section of a querythat specifies criteria that the requested data must meet

**World Health Organization:** An organization whose primary role is to direct and coordinate international health within the United Nations system

### 1.

Question 1

Fill in the blank: Data professionals use data \_\_\_\_\_ to handle issues related to data security and privacy while ensuring data assets are formally managed.

1 point

integrity

governance

strategy

mapping

### 2.

Question 2

A data analyst at a software company uses metadata to track the source of user bugs in order to troubleshoot the problem. They learn about the device, operating system, and browser on which each bug was encountered. What type of metadata are they using?

1 point

Administrative

General

Descriptive

Structural

### 3.

Question 3

Which of the following statements accurately describe primary and foreign keys in a relational database? Select all that apply.

1 point

A primary key references a column in which each value is unique.

Foreign keys are used to connect one table to another.

A foreign key uniquely identifies a record in a relational database table.

A table can only have one primary key, but it can have multiple foreign keys.

### 4.

Question 4

A data analyst runs the following query. What do they want to retrieve from the database?

|  |
| --- |
| **SELECT \***  **FROM Video\_Games**  **WHERE Creator = 'Maddox'** |

1 point

All fields of the video game creators

All fields of the video games

All fields of the video games created by Maddox

All fields of the video games not created by Maddox

### 5.

Question 5

A junior data professional prepares for an analysis project about a very broad and global topic. However, they will only have access to internal data. What are some potential limitations that they should be aware of? Select all that apply.

1 point

The data may not fully represent the facts.

The data is not owned by the company.

It may be difficult to gather data from multiple departments.

It will be more difficult to confirm the reliability of the data.

### 6.

Question 6

A data team at a trade school is sending a text alert to all students who have fewer than 10 credits. What spreadsheet tool will enable them to display only the students who meet that condition?

1 point

Sort the number of student credits in descending order

Filter out students with more than 10 credits

Sort the number of student credits in ascending order

Filter out students who have fewer than 10 credits

### 7.

Question 7

Which SQL statement will return only green cars from the Color column of the Cars database table?

1 point

|  |
| --- |
| **SELECT \***  **FROM Cars**  **WHERE 'Green'** |
| **SELECT \***  **FROM Color = 'Green'** |

|  |
| --- |
| **SELECT \***  **FROM Cars**  **WHERE Color = 'Green'** |
| **SELECT \***  **WHERE 'Green'** |

### 8.

Question 8

A data analyst at a retail company uses a tool to explore the data in its customer database. They learn the definition of each column, the data types contained, and the relationships between different tables. What does this scenario describe?

1 point

Designing a database

Combining data from more than one source

Performing data analysis

Using a metadata repository

Module 4

Hey, good to have you back.

Up until now, we've focused on preparing

your data for processing and analysis.

In these next videos,

we'll explore another big part of that process,

organizing and protecting your data.

Keeping your data organized is

important for a few reasons;

it makes it easier to find and use,

helps you avoid making mistakes during

your analysis and helps to protect it.

Coming up, we'll go over

the basics of organizing data for

personal and professional use

and file naming conventions.

Then we'll take a look at some

security features for spreadsheets.

By the end of these next few videos,

you'll be able to do all these things and you'll

be able to explain these steps to stakeholders,

so they can feel confident that

your data practices are safe and secure.

When you're ready to get started,

go ahead to the next video.

There we'll get started with

organizing data for personal use.

Hey, welcome back.

Whether you're organizing your personal data for

your own use or organizing project data for work,

there are certain procedures you want to follow to

make sure your data is easy to find and use.

In this video, we'll cover

some best organization practices

and also check out some different ways

project data can be organized.

There are plenty of best practices you

can use when organizing data,

including naming conventions, foldering,

and archiving older files.

We've talked about file naming before,

which is also known as naming conventions.

These are consistent guidelines

that describe the content,

date, or version of a file in its name.

Basically, this means you want to use

logical and descriptive names for

your files to make them easier to find and use.

Speaking of easily finding things,

organizing your files into folders helps

keep project-related files together in one place.

This is called foldering.

For example, all the files related to your vacation plan

might go in the Vacation2025 folder.

You might then break that folder down even further by

creating subfolders like itinerary or photos,

depending on what else you'd like to easily access.

It can also be useful to move old projects to

a separate location to create

an archive and cut down on clutter.

It's so much easier to find and use

my files when I name them

something meaningful and searchable

and when I organize them into folders.

It makes all my data more accessible and useful.

In addition to these three best practices,

there are two more things you'll want to consider

when organizing data for work use.

First, the project data you'll be using for

work could be accessed and used by multiple people.

It's important to align

your naming and storage practices with

your team to avoid any confusion.

Your team might also develop

metadata practices like creating

a file that outlines project

naming conventions for easy reference.

We'll get to talk more about naming conventions

for work files in more detail later.

Secondly, you want to think about how often you're

making copies of data and storing it in different places.

Most importantly, because if data is stored

in lots of different databases or spreadsheets,

it can contradict itself and lead to mistakes later on.

Also storing data in

multiple places takes up a lot of space.

Relational databases can help you avoid

data duplication and store your data more efficiently.

You can use these practices to organize

data in different ways according to your project.

Let's look at some examples of data organization.

I have some sample project folders here,

each organized in a slightly different way.

Let's open them up and see what they look like.

We'll start with the high-level Finances folder.

The Finances folder has been organized categorically.

There are subfolders like budget,

invoices, and payroll

that represent different categories.

Let's click on "Invoices" to see what's in there.

In the invoices folder,

you can see that we have another set

of subfolders labeled

by year, 2014, 2015....

Looks like these are in chronological order.

Sometimes the way files are organized can tell us

how the data within those files is also organized.

Let's open a file to see if that's right.

In the 2014 subfolder,

there's a file with invoices from June.

Play video starting at :3:18 and follow transcript3:18

If we open it, we

can see that they've been organized by date,

just like the folders.

There's different ways to organize

data depending on what you need it for.

The categorical organization of the subfolders and

finances made it easy

for me to go straight to the invoices,

but the chronological organization

of the invoices subfolder

can help us find financial data

from the exact date we're looking for.

There's other ways to organize data too:

in order of importance or even by location.

For example, a company might use

hierarchical organization so that employee data mirrors

the structure of their employee organization. Or a company

working with geographical data

might choose to organize by location.

It's a good idea to take time

early on in a project to consider what

the best organization methods will

be for you and your team to stick to.

Here's another way to think about it.

Unorganized data is like a messy room.

It's overwhelming, hard to find anything

in, and gets worse the longer you avoid cleaning it up.

But by making sure early

on you know where to put your files,

you can keep your work data organized,

easy to use, and error free.

Now that you see how important it is to keep

data organized for both personal and work use,

we'll take a closer look at file naming

conventions and how they carry over into your databases.

See you in the next video.

# File organization guidelines

Every data analyst’s goal is to conduct efficient data analysis. One way to increase the efficiency of your analyses is to streamline processes that help save time and energy in the long run. Meaningful, logical, and consistent file names help data analysts organize their data and automate their analysis process. When you use consistent guidelines to describe the content, date, or version of a file and its name, you’re using file naming conventions.

In this reading, you’ll learn more about best practices for file naming conventions and file organization.



## Best practices for naming files

File-naming conventions help you organize, access, process, and analyze data because they act as quick reference points to identify what’s in a file. One important practice is to decide on file naming conventions—as a team or company—early in a project. This will prevent you from spending time updating file names later, which can be a time-consuming process. In addition, you should align your project’s file names with your team’s or company’s existing file-naming conventions. You don’t want to spend time learning a new file-naming convention each time you look up a file in a new project!

It's also critical to ensure that file names are meaningful, consistent, and easy-to-read. File names should include:

* The project’s name
* The file creation date
* Revision version
* Consistent style and order

Further, file-naming conventions should act as quick reference points to identify what is in the file. Because of this, they should be short and to the point.

In the following sections, you’ll explore each part of a sales report file name that follows an established naming convention, **SalesReport\_20231125\_v02**. This example will help you understand the key parts of a strong file name and why they’re important.

### **Name**

Giving a file a meaningful name to describe its contents makes searching for it straightforward. It also makes it easy to understand the type of data the file contains.

In the example, the file name includes the text **SalesReport**, a succinct description of what the file contains: a sales report.

### **Creation date**

Knowing when a file was created can help you understand if it is relevant to your current analysis. For example, you might want to analyze only data from 2023.

In the example, the year is described as **20231125**. This reads as the sales report from November 25, 2023 following the year, month, and day (YYYYMMDD) format of the international date standard. Keep in mind that different countries follow different date conventions, so make sure you know the date standard your company follows.

### **Revision version**

Including a revision version helps ensure you’re working with the correct file. You wouldn’t want to make edits to an old version of a file without realizing it! When you include revision numbers in a file name, lead with a zero. This way, if your team reaches more than nine rounds of revisions, double digits are already built into your convention.

In the example, the version is described as **v02**. The v is short for the version of the file, and the number following the v indicates which round of revisions the file is currently in.

### **Consistent order and style**

Make sure the information you include in a file name follows a consistent order. For example, you wouldn’t want version three of the sales report in the example to be titled **20231125\_v03\_SalesReport**. It would be difficult to find and compare multiple documents.

When you use spaces and special characters in a file name, software may not be able to recognize them, which causes problems and errors in some applications. An alternative is to use hyphens, underscores, and capital letters. The example includes underscores between each piece of information, but your team could choose to use hyphens between year, month, and date, too: **SalesReport\_2023\_11\_25\_v02**.

## Ensure team consistency

To ensure all team members use the agreed-upon file naming conventions, create a text file as a sample that includes all of the naming conventions on a project. This can benefit new team members to help them quickly get up to speed or a current team member who just needs a refresher on the file naming conventions.

## File organization

To keep your files organized, create folders and subfolders—in a logical hierarchy—to ensure related files are stored together and can be found easily later. A hierarchy is a way of organizing files and folders. Broader-topic folders are located at the top of the hierarchy, and more specific subfolders and files are contained within those folders. Each folder can contain other folders and files. This allows you to group related files together and makes it easier to find the files you need. In addition, it’s a best practice to store completed files separately from in-progress files so the files you need are easy to find. Archive older files in a separate folder or in an external storage location.

## Key takeaways

Use consistent, meaningful file-naming conventions throughout your project to save you and your team time by making data easy to find and use. File-naming conventions should be agreed upon by all team members before starting a project and should describe the project by including its name, the date, and the revision version. Document this information in a location that team members can access.

### 1.

Question 1

Fill in the blank: Naming \_\_\_\_\_ are consistent guidelines used to describe the content, date, or version of a file.

1 point

references

attributes

conventions

descriptors

### 2.

Question 2

What are the key goals of foldering in data analytics? Select all that apply.

1 point

Organize files into subfolders

Keep project-related files together

Assign metadata about the folders

Transfer files from one place to another

### 3.

Question 3

What is the process of structuring folders broadly at the top, then breaking down those folders into more specific topics?

1 point

Assigning naming conventions

Developing metadata

Producing a backup

Creating a hierarchy

### 4.

Question 4

Which of the following examples would be the most effective file name?

1 point

Data\_519

CampaignData\_03

AirportCampaign\_2013\_10\_09\_V01

May30-2019\_AirportAdCampaignResults\_Terminals3-5\_InclCustSurveyResponses\_Ideas

You're back. Okay, now that

our data's organized and easy to find,

it's time to start thinking about how to protect it.

The good news is that spreadsheets come with

security features already built in.

In this video, we'll look at

different spreadsheet programs and how

their security features, like

sheet protections and access control, are similar.

When I say "security features,"

you might be imagining ways to

protect data from other people.

But that's just one kind of security.

Security features can be designed to keep

unauthorized users from viewing certain files,

or just lock your worksheets so that you

don't accidentally break your formulas.

This is called data security.

Data security means protecting data

from unauthorized access or corruption

by adopting safety measures.

Whatever spreadsheet program you're using will

have similar security measures built in.

As a data analyst,

you'll run into Google Sheets and Excel a lot.

Let's talk about what they have in common.

First, both programs have features

that let you protect your spreadsheets or parts

of your spreadsheets from being edited, from

the entire worksheet down to single cells in a table.

If you're collaborating with other users,

you can easily lock down

your formulas so that they aren't accidentally broken.

Speaking of collaborating,

Excel and Google Sheets both have

access control features like

password protection and user permissions.

This gives you more control

over who can do what to your spreadsheet.

Because these programs are located in different places,

these features are slightly different.

For Excel spreadsheets, you can encrypt

files and worksheets with passwords

before emailing them to other users. In Google Sheets,

these settings are found under the sharing menu,

which allows you to control who can

see or edit the sheet online.

Google Sheets can also be copied so that users

can work with that data without altering the original.

Tabs can also be hidden and unhidden in Sheets and Excel,

allowing you to change what data is being viewed.

But remember, even hidden tabs

can be unhidden by someone else,

so be sure you're okay with those

tabs still being accessible.

As a data analyst, data security will be a priority.

But no matter which program you use to create spreadsheets,

there's security features to help you

keep your work safe and secure.

There are some other basic best practices you can

take to keep your data more secure overall,

which we'll cover later in a reading.

You've made it to the end of this module. Congrats.

In these videos, we've covered strategies for

organizing data for personal and work use,

how to develop functional file naming conventions,

and some security measures you can

take advantage of in spreadsheets.

Before you move on to the next step

in the data analysis lifecycle.

It's important that you make sure your data is

prepared, and that includes organizing and securing it.

As usual after this video,

you'll have your weekly challenge.

I know you've got this.

Then after the weekly challenge,

there's some optional material all about

connecting to the online data community.

As you start building your career in data analytics,

it'll be really valuable to connect with others,

learn about new trends in

the field and share your own work.

I think you'll get a lot out of those videos.

That'll help you develop a professional online presence

and find ways to communicate with people in your field,

which is key as networking becomes

more and more online and

remote work opportunities become the norm.

But if you feel pretty confident

about your online presence,

you can move into the course challenge instead.

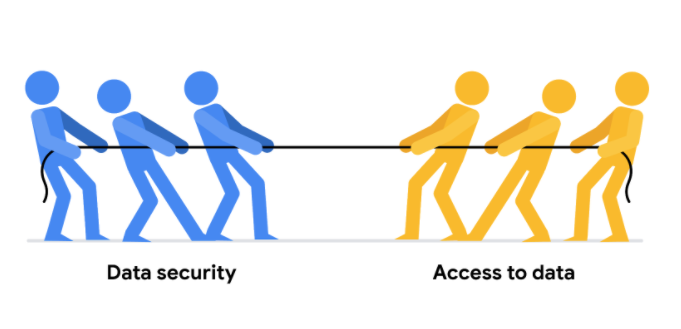
Good luck on this weekly challenge,

and I'll see you soon!

**Balance security and analytics**

**Data security** means protecting data from unauthorized access or corruption by putting safety measures in place. Usually the purpose of data security is to keep unauthorized users from accessing or viewing sensitive data. Data analysts have to find a way to balance data security with their actual analysis needs. This can be tricky-- we want to keep our data safe and secure, but we also want to use it as soon as possible so that we can make meaningful and timely observations.

In order to do this, companies need to find ways to balance their data security measures with their data access needs.



Luckily, there are a few security measures that can help companies do just that. The two we will talk about here are encryption and tokenization.

**Encryption** uses a unique algorithm to alter data and make it unusable by users and applications that don’t know the algorithm. This algorithm is saved as a “key” which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.

**Tokenization** replaces the data elements you want to protect with randomly generated data referred to as a “token.” The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

Encryption and tokenization are just some of the data security options out there. There are a lot of others, like using authentication devices for AI technology.

As a junior data analyst, you probably won’t be responsible for building out these systems. A lot of companies have entire teams dedicated to data security or hire third party companies that specialize in data security to create these systems. But it is important to know that all companies have a responsibility to keep their data secure, and to understand some of the potential systems your future employer might use.

However, one thing you absolutely can do to help strike the right balance is to use **version control** best practices. Version control enables all collaborators within a file to track changes over time. You can understand who made what changes to a file, when they were made, and why.

Here's a simple example: Perhaps you're working on a project with a team of other people. You are all collaborating within the same set of files, but each person is responsible for a different part of the project. Without version control, it would be very difficult to keep track of who made what changes to the files and when. This would lead to confusion and, even worse, people accidentally overwriting each other's work! Version control is essential for data analytics professionals because it allows users to effectively collaborate with others and experiment with new ideas without fear of losing their work.

### 1.

Question 1

Fill in the blank: Data security involves adopting \_\_\_\_\_ in order to protect data from unauthorized access or corruption.

1 point

foldering procedures

data validation

safety measures

metadata strategy

### 2.

Question 2

What data-security measure uses a unique algorithm to alter data and make it inaccessible without the algorithm?

1 point

Password-protection

Encryption

AI authentication

Hidden tabs

### 3.

Question 3

When using tokenization as a safety measure, what is replaced as a randomly generated token?

1 point

The user’s location

The user’s search history

The line of code from a query

The data elements to be protected

### 4.

Question 4

What data-security practice enables all collaborators within a file to track changes, such as who made what edits to the file, when they were made, and why?

1 point

Version control

Foldering

Authentication

Password-protection

**Access control:** Features such as password protection, user permissions, and encryption that are used to protect a spreadsheet

**Data security:** Protecting data from unauthorized access or corruption by adopting safety measures

**Inbox:** Electronic storage where emails received by an individual are held

### 1.

Question 1

You are in charge of your company’s weekly accounting spreadsheet. It has 15 sheets, each containing a different employee’s purchases. You add restrictions to the spreadsheet to make sure employees can only edit their own sheets. What practice does this scenario describe?

1 point

Data preservation

Data hygiene

Data integrity

Data security

### 2.

Question 2

What aspects of a file do file-naming conventions typically describe? Select all that apply.

1 point

Version number

Creation date

Collaborator names

Content description

### 3.

Question 3

A data analyst team revisits an old project and wants to understand how the file-naming conventions are structured. Where does the team locate this information?

1 point

In the metadata

In folder hierarchies

In aggregated data

In SQL

### 4.

Question 4

A grocery store collects inventory data about its produce section. What is an appropriate naming convention for this file?

1 point

Todays\_Produce

Inventory\_Produce 2022-09-15 V01

Todays Produce 2022-15-09

Produce\_Inventory\_2022-09-15\_V01

### 5.

Question 5

A data analyst uses \_\_\_\_\_ to organize multiple files for a given project so they can be found and accessed in an efficient manner.

1 point

data hygiene

foldering

version control

data grouping

### 6.

Question 6

A vendor asks a data team at their partner company to share a spreadsheet. The spreadsheet contains three tabs. Tabs 1 and 2 are meant for the vendor to review, but tab 3 contains sensitive internal information. Which of the following tactics will enable the data team to keep tab 3 private? Select all that apply.

1 point

Rename tab 3 “Sensitive,” then share the spreadsheet with the vendor.

Copy tabs 1 and 2 into a separate spreadsheet, then share the new file with the vendor.

Hide tab 3, then share the spreadsheet with the vendor.

Make a copy of the spreadsheet, delete tab 3, then share the new file with the vendor.

### 7.

Question 7

A junior data analyst finishes an employee performance analysis project, so they delete the related files. A few months later, their supervisor assigns them another project related to employee performance, for which the deleted files could have been very useful. What would have been a better course of action?

1 point

Keep the project files on their local drive

Print out the project files

Email the project to their supervisor

Archive the project files

### 8.

Question 8

Fill in the blank: In an effective \_\_\_\_\_, subfolders are stored within broader-topic folders in order to help organize files into categories.

1 point

encryption

permission

file extension

hierarchy

Module 5

Hey, it's great to have you back! So far we've covered

everything from using SQL to

the key aspects of data ethics.

You've developed a huge range

of skills, and they're all going to

help you on your journey to a career in data analytics.

But you don't have to do everything on your own.

As a data analyst,

you'll be part of a growing data community.

By building a consistent

and professional online presence,

you'll be able to connect to others in

your field and expand your network.

Coming up, you'll learn how you can get

started building your online presence.

Or if you're already part of the community,

how you can take your online network even further.

With remote, online work becoming more and more common,

so is online networking.

That means having and maintaining

a well-developed online presence could

open the door to so many new opportunities.

I find myself reaching out to people I've worked

with throughout my career to stay in touch,

ask them questions about their experiences,

and just see what interesting things they're doing,

and that's only possible because I

keep up my online presence.

Join me in the next video to get started

building your online presence, and get connected!

Hey again. Today, a lot of us spend

a lot of time connecting with people online.

We stay in touch with family and

friends we can't see everyday,

or post about what we're doing,

eating, and watching on social media.

But our presence online goes beyond the personal.

A consistent and professional online presence is

an important tool in building a career in data analytics.

A professional online presence is

important for a few key reasons.

First, it can help potential employers find you.

Second, it lets you make

connections with other data analysts in your field,

learn and share data findings,

and maybe even participate in community events.

Keep in mind that a lot of networking happens online now.

If you aren't keeping up your online presence,

you might be missing out on

great opportunities without even knowing it.

There are lots of different

professional sites that you can take

advantage of as you start

building your own online presence.

For now though, we'll focus on LinkedIn and GitHub.

LinkedIn is specifically designed to help

people make connections with other people in their field.

It's a great way to follow trends in your industry,

learn from industry leaders,

and stay engaged with the wider professional community.

And if you're actively looking for a new job,

LinkedIn has job boards that you can search.

You can even narrow down

your location to see who's hiring near you.

Plus, job recruiters frequently use LinkedIn to

find potential data analysts for new projects.

It's always a good idea to keep

your LinkedIn profile up to date with your resume.

You might find yourself being recruited.

LinkedIn also lets you

connect with people and build a network.

You can share exciting things happening in

your professional life and

keep up with where your connections go.

You never know when you might end

up working with someone again.

With LinkedIn, you can be endorsed for having

job skills or endorse other people.

If you impress someone at a previous job,

they can let other people know

just how awesome you are to work with.

GitHub, the other website I

mentioned earlier, is a little different.

GitHub is part code-sharing site, part social media.

It has an active community

collaborating and sharing insights to build resources.

You can talk with other GitHub users on the forum,

use the community-driven wikis,

or even use it to manage team projects.

GitHub also hosts community events where you can

meet other people in the field and learn some new things.

GitHub has a lot of features for you to check out.

The best way to learn more about

it is to check it out for yourself.

We'll also be talking more about

GitHub later in the program.

Sometimes if you're looking for a new career,

finding someone who has something in common with you,

like shared interests or the same hometown,

and reaching out to them, can help a lot.

Just a 15-minute conversation with

someone could set you on the path to a new career,

whether that's on a professional

networking site like LinkedIn,

or at a community event hosted by GitHub.

LinkedIn has become one of

the standard professional social media sites,

so it's a good starting place for

building your online presence.

GitHub offers a lot of

really great tools for data analysts in the community.

If you don't already have accounts on these sites,

challenge yourself to set them up now.

Connect with other people.

Share some updates about what

you're working on right now.

If you're already using LinkedIn and GitHub, great news:

we're going to talk more about how to enhance

your existing social media presence

next time. See you soon.