**Select 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.

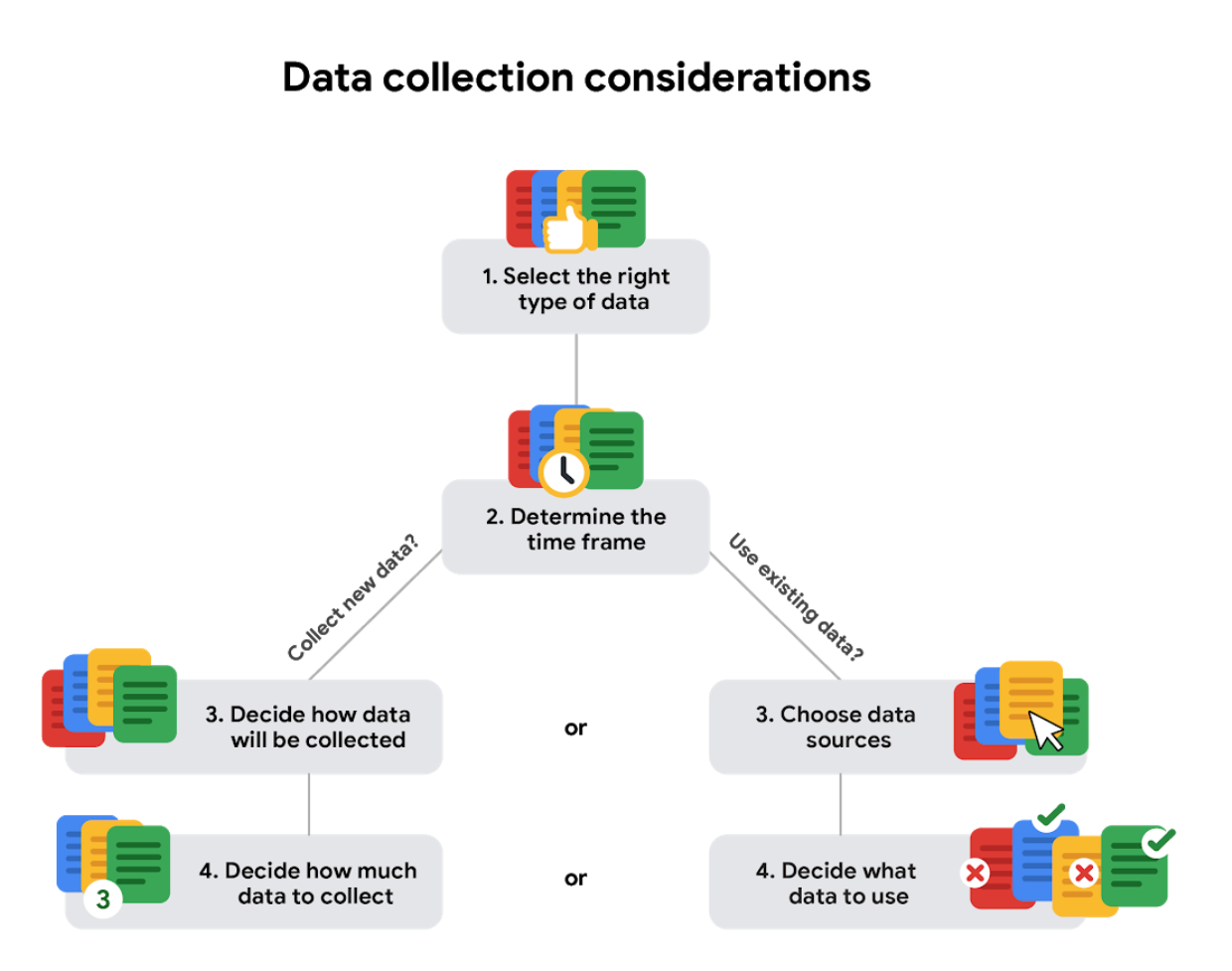
**How much data to collect**

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:



# 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.

## Data format examples

As with most things, it is easier for definitions to click when you can pair them with examples you might encounter on a daily basis. Review each data format’s definition first and then use the examples to lock in your understanding.

## Primary versus secondary data

| **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 primary and secondary data and presents examples of each.

## Internal versus external data

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

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Internal data | Data that is stored 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 is stored outside of a company or organization | * National average wages for the various positions throughout your organization * Credit reports for customers of an auto dealership |

## Continuous versus discrete data

| **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) * Maximum capacity allowed in a room * Tickets sold in the current month |

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

## Qualitative versus quantitative data

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Qualitative | A subjective and explanatory measure of a quality or characteristic | * Favorite exercise activity * Brand with best customer service * Fashion preferences of young adults |
| Quantitative | A specific and objective measure, such as a number, quantity, or range | * Percentage of board certified doctors who are women |

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

## Nominal versus ordinal data

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Nominal | A type of qualitative data that is categorized without 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) * Satisfaction level measured in a survey (satisfied, neutral, dissatisfied) |

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

## Structured versus unstructured data

The following table highlights the differences between structured and unstructured data and presents 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 cannot be stored as columns and rows in a relational database. | * Social media posts * Emails * Videos |

**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.

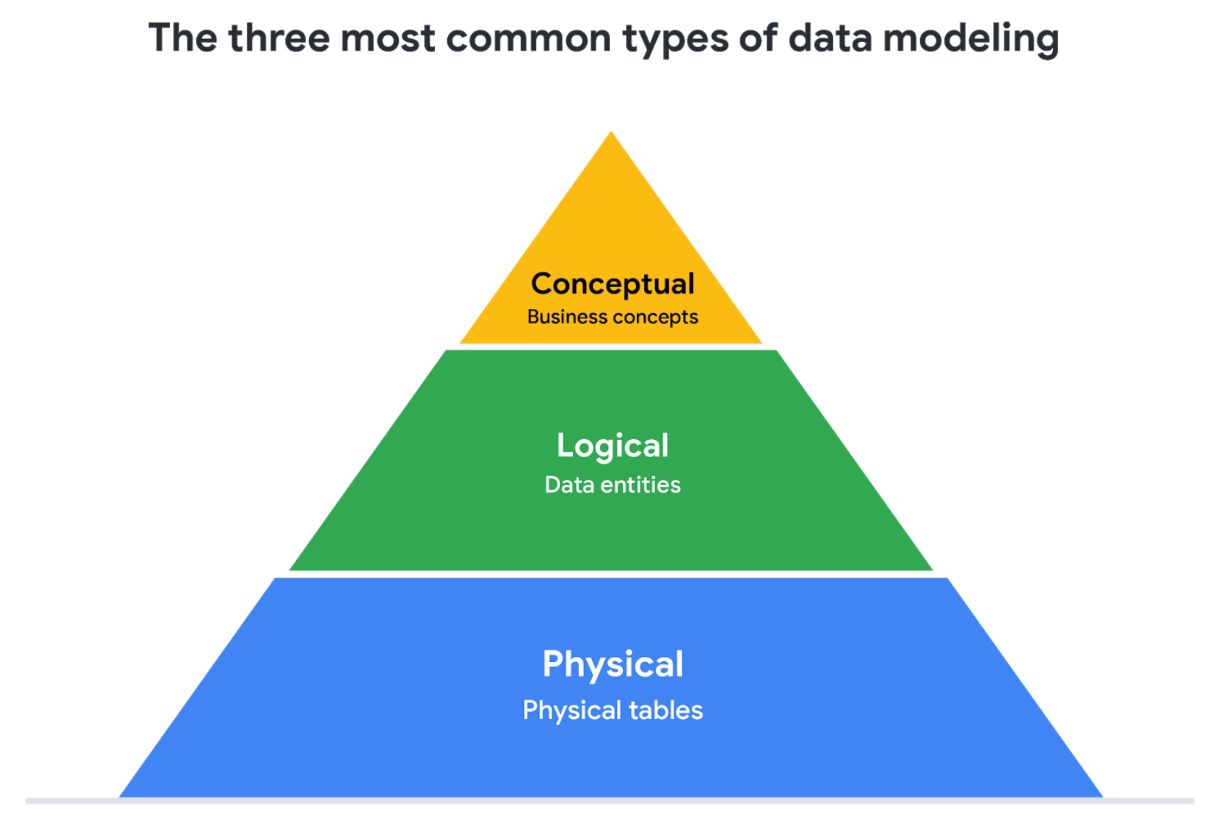
**Important note:** As a junior data analyst, you won't be asked to design a data model. But you might come across existing data models your organization already has in place.

**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.

**Data-modeling techniques**

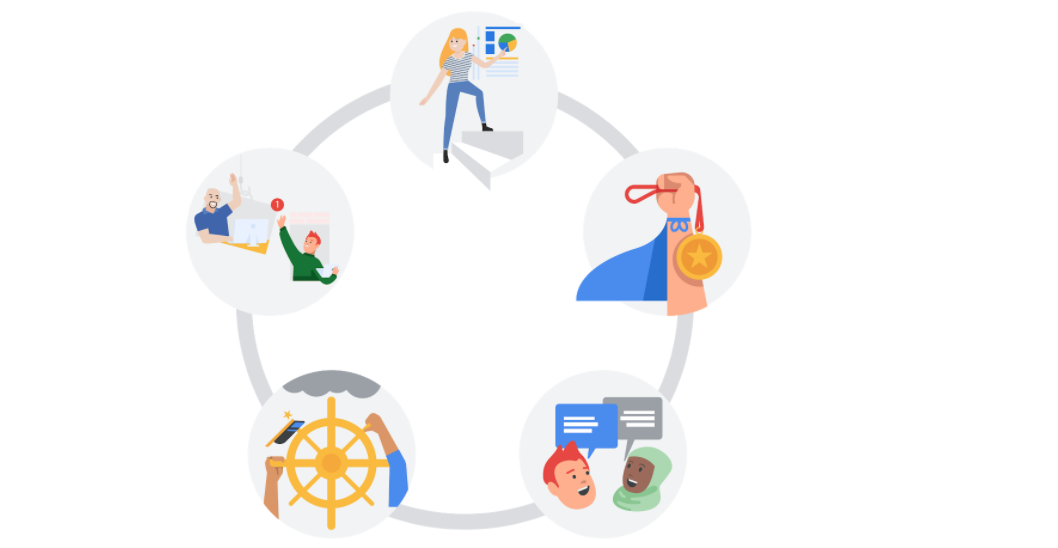
There are a lot of approaches when it comes to developing data models, but two common methods are the **Entity Relationship Diagram (ERD)** and the **Unified Modeling Language (UML)** diagram. ERDs are a visual way to understand the relationship between entities in the data model. UML diagrams are very 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.

**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!

# 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**

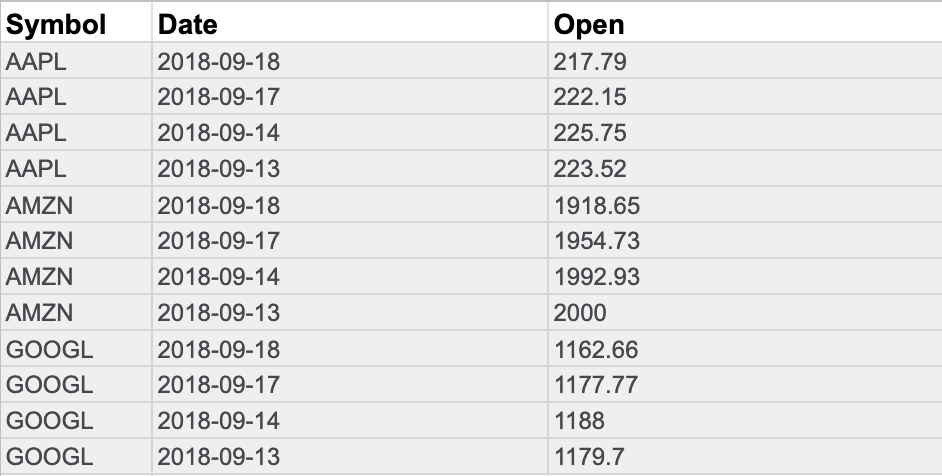
Mario is a plumber who owns a plumbing company. After years in the business, he buys another plumbing company. Mario wants to merge the customer information from his newly acquired company with his own, but the other company uses a different database. So, Mario needs to make the data compatible. To do 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.

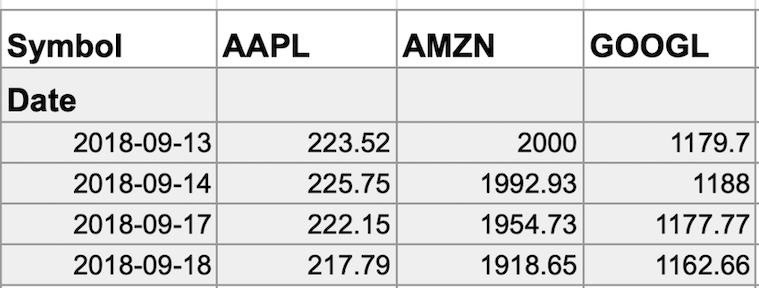
**Long data** is data where **each row contains a single data point** for a particular item. In the long data example below, individual stock prices (data points) have been collected for Apple (AAPL), Amazon (AMZN), and Google (GOOGL) (particular items) on the given dates.

**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 |

# Data anonymization

## What is data anonymization?

You have been learning about the importance of privacy in data analytics. Now, it is time to talk about **data anonymization** and what types of data should be anonymized. **Personally identifiable information**, or **PII**, is information that can be used by itself or with other data to track down a person's identity.

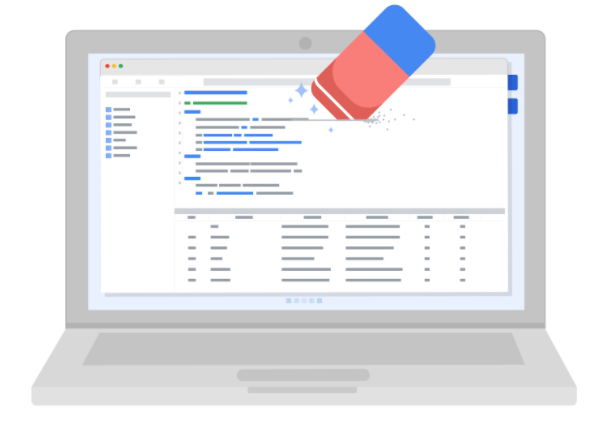
Data anonymization is the process of protecting people's private or sensitive data by eliminating that kind of information. Typically, data anonymization involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values.

### **Your role in data anonymization**

Organizations have a responsibility to protect their data and the personal information that data might contain. As a data analyst, you might be expected to understand what data needs to be anonymized, but you generally wouldn't be responsible for the data anonymization itself. A rare exception might be if you work with a copy of the data for testing or development purposes. In this case, you could be required to anonymize the data before you work with it.

## What types of data should be anonymized?

Healthcare and financial data are two of the most sensitive types of data. These industries rely a lot on data anonymization techniques. After all, the stakes are very high. That’s why data in these two industries usually goes through **de-identification**, which is **a process used to wipe data clean of all personally identifying information**.



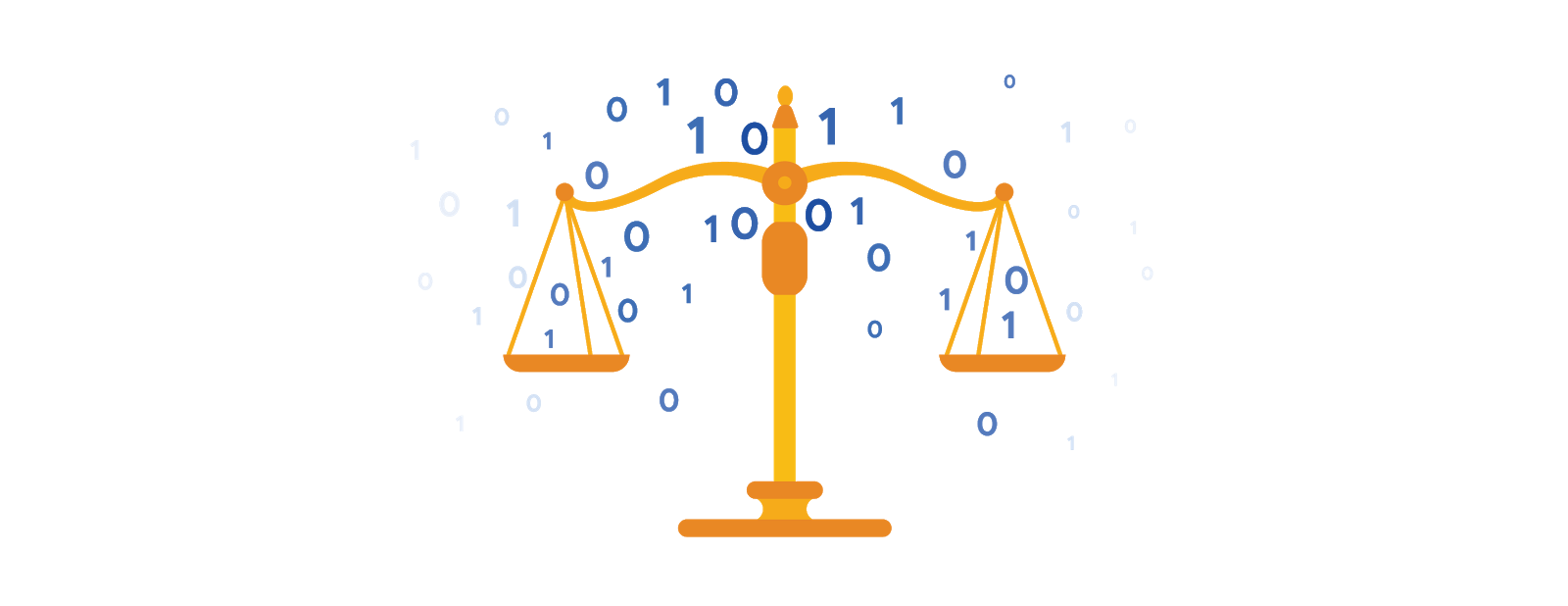
Data anonymization is used in just about every industry. That is why it is so important for data analysts to understand the basics. Here is a list of data that is often anonymized:

* Telephone numbers
* Names
* License plates and license numbers
* Social security numbers
* IP addresses
* Medical records
* Email addresses
* Photographs
* Account numbers

For some people, it just makes sense that this type of data should be anonymized. For others, we have to be very specific about what needs to be anonymized. Imagine a world where we all had access to each other’s addresses, account numbers, and other identifiable information. That would invade a lot of people’s privacy and make the world less safe. Data anonymization is one of the ways we can keep data private and secure!

**The open data debate**

Just like data privacy, open data is a widely debated topic in today’s world. Data analysts think a lot about open data, and as a future data analyst, you need to understand the basics to be successful in your new role.



**What is open data?**

In data analytics, **open data** is part of **data ethics,** which has to do with using data ethically. **Openness** refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

* Be available and accessible to the public as a complete dataset
* Be provided under terms that allow it to be reused and redistributed
* Allow universal participation so that anyone can use, reuse, and redistribute the data

Data can only be considered open when it meets all three of these standards.

**The open data debate: What data should be publicly available?**

One of the biggest benefits of open data is that credible databases can be used more widely. Basically, this means that all of that good data can be leveraged, shared, and combined with other data. This could have a huge impact on scientific collaboration, research advances, analytical capacity, and decision-making. But it is important to think about the individuals being represented by the public, open data, too.

**Third-party data** is collected by an entity that doesn’t have a direct relationship with the data. You might remember learning about this type of data earlier. For example, third parties might collect information about visitors to a certain website. Doing this lets these third parties create audience profiles, which helps them better understand user behavior and target them with more effective advertising.

**Personal identifiable information (PII)** is data that is reasonably likely to identify a person and make information known about them. It is important to keep this data safe***.*** PII can include a person’s address, credit card information, social security number, medical records, and more.

Everyone wants to keep personal information about themselves private. Because third-party data is readily available, it is important to balance the openness of data with the privacy of individuals.

## ****Terms and definitions for Course 3, Module 2****

**Bad data source:** A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

**Bias:** A conscious or subconscious preference in favor of or against a person, group of people, or thing

**Confirmation bias:** The tendency to search for or interpret information in a way that confirms pre-existing beliefs

**Consent:** The aspect of data ethics that presumes an individual’s right to know how and why their personal data will be used before agreeing to provide it

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

**Currency:** The aspect of data ethics that presumes individuals should be aware of financial transactions resulting from the use of their personal data and the scale of those transactions

**Data anonymization:** The process of protecting people's private or sensitive data by eliminating identifying information

**Data bias:** When a preference in favor of or against a person, group of people, or thing systematically skews data analysis results in a certain direction

**Data ethics:** Well-founded standards of right and wrong that dictate how data is collected, shared, and used

**Data interoperability:** A key factor leading to the successful use of open data among companies and governments

**Data privacy:** Preserving a data subject’s information any time a data transaction occurs

**Ethics:** Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues

**Experimenter bias:** The tendency for different people to observe things differently (also called observer bias)

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

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

**General Data Protection Regulation of the European Union (GDPR):** Policy-making body in the European Union created to help protect people and their data

**Good data source:** A data source that is reliable, original, comprehensive, current, and cited (ROCCC)

**Interpretation bias:** The tendency to interpret ambiguous situations in a positive or negative way

**Observer bias:** The tendency for different people to observe things differently (also called experimenter bias)

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

**Openness:** The aspect of data ethics that promotes the free access, usage, and sharing of data

**Sampling bias:** Over representing or underrepresenting certain members of a population as a result of working with a sample that is not representative of the population as a whole

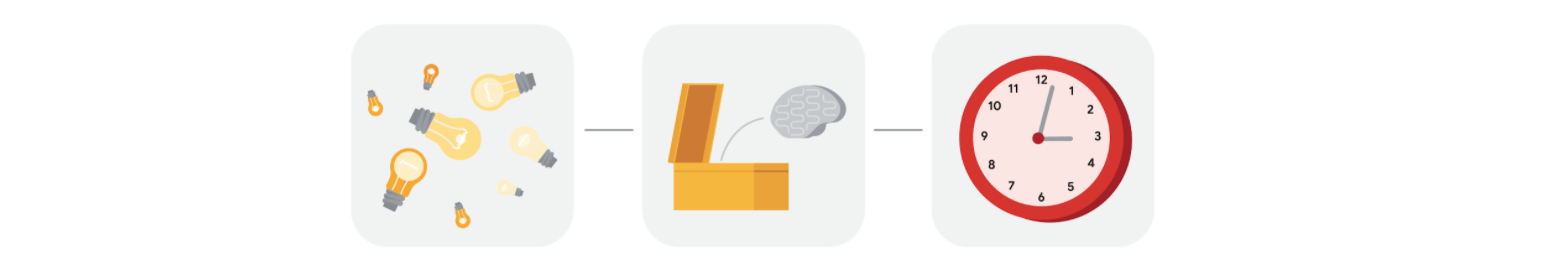
**Transaction transparency:** The aspect of data ethics that presumes all data-processing activities and algorithms should be explainable and understood by the individual who provides the data

**Unbiased sampling:** When the sample of the population being measured is representative of the population as a whole

# Module 3

# 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.

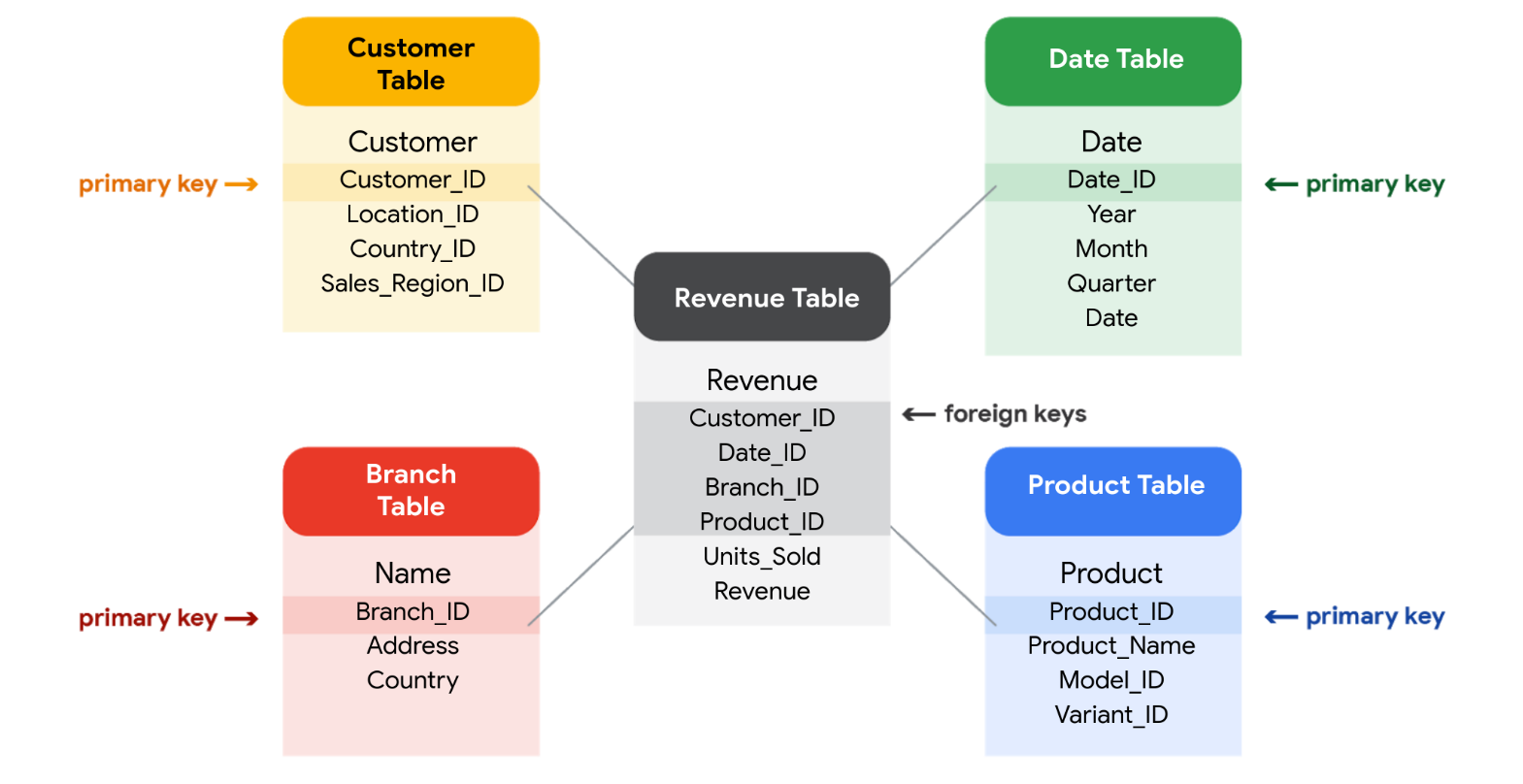
**Normalization** is a process of organizing data in a relational database. For example, creating tables and establishing relationships between those tables. It is applied to eliminate data redundancy, increase data integrity, and reduce complexity in a database.

## ****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.



# 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

* **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.

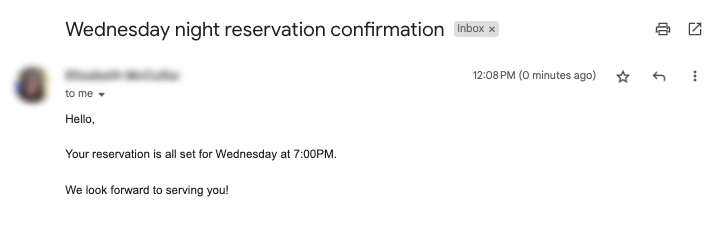
### **Photos**

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.

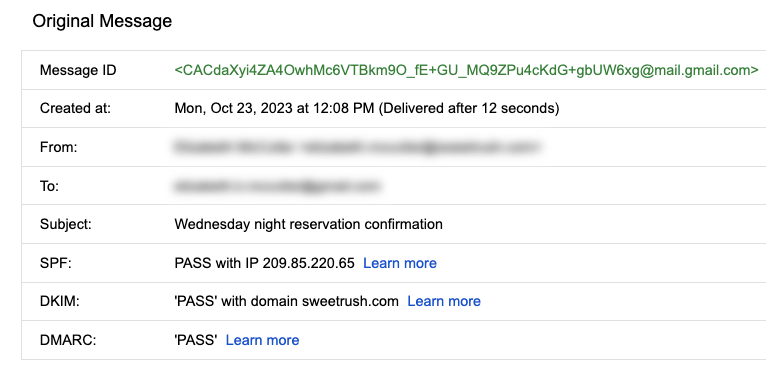
Image with accompanying Information pop-up that displays its description, the date and time the image was taken, its size, the device on which it was taken, and an option to add the geolocation of the image.

### **Emails**

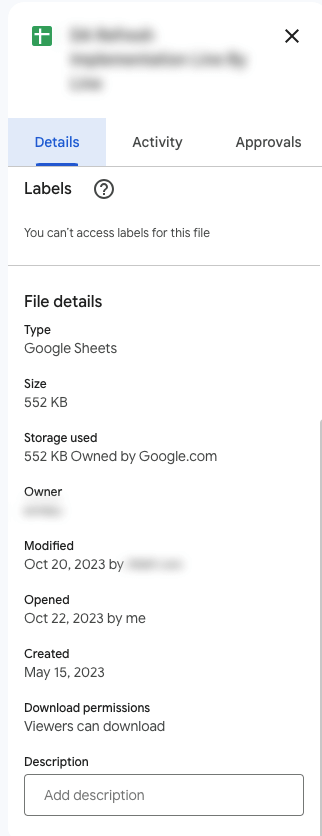
When an email is sent or received, it contains metadata such as subject line, sender, recipient, date sent, and time sent.



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.

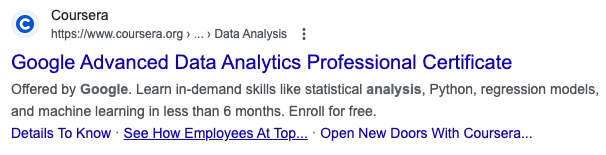
Hidden metadata from an email that includes Message ID, creation date, the recipient, the sender, the subject line, the SPF, DKIM, and DMARC.

Spreadsheets and 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.

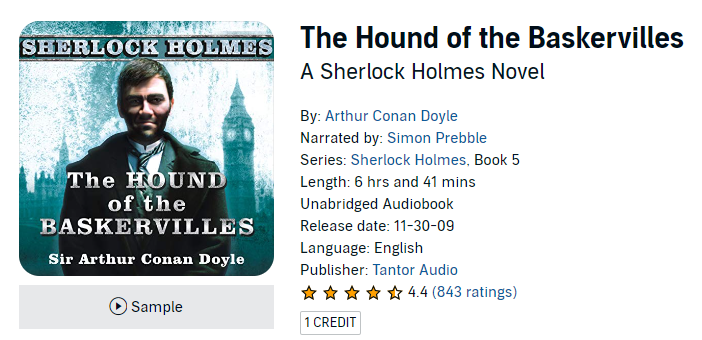
### **Websites**

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.

# Module 4

# 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.

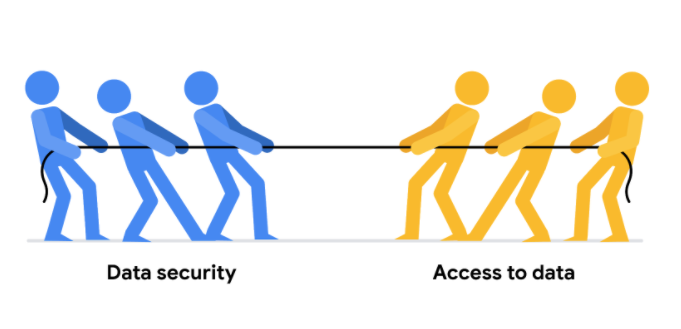
## File organization

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.

# 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.

## ****Terms and definitions for Course 3, Module 4****

**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