# 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

The following table highlights the differences between primary and secondary data and presents 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<br>• 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<br>• Demographic data collected by a university<br>• Census data gathered by the federal government |

## 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<br>• Sales data by store location<br>• 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<br>• Credit reports for customers of an auto dealership |

## Continuous versus discrete data

The following table highlights the differences between continuous and discrete data and presents 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)<br>• Runtime markers in a video<br>• 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)<br>• Maximum capacity allowed in a room<br>• Tickets sold in the current month |

## Qualitative versus quantitative data

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

| Data format classification | Definition | Examples |
| --- | --- | --- |
| Qualitative | A subjective and explanatory measure of a quality or characteristic | • Favorite exercise activity<br>• Brand with best customer service<br>• 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<br>• Population size of elephants in Africa<br>• Distance from Earth to Mars at a particular time |

## Nominal versus ordinal data

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

| 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<br>• New job applicant, existing applicant, internal applicant<br>• 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)<br>• Ranked-choice voting selections (1st, 2nd, 3rd)<br>• Satisfaction level measured in a survey (satisfied, neutral, dissatisfied) |

## 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<br>• Tax returns<br>• Store inventory |
| Unstructured data | Data that cannot be stored as columns and rows in a relational database. | • Social media posts<br>• Emails<br>• Videos |

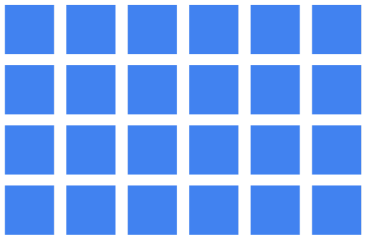
# The effects of different structures

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.

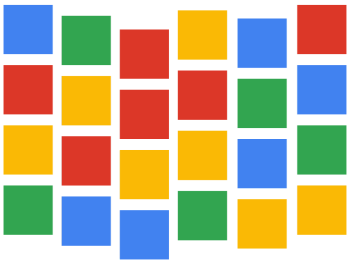
For example, when you rate your favorite restaurant online, you're creating structured data. But when you use Google Earth to check out a satellite image of a restaurant location, you're using unstructured data.
Here's a refresher on the characteristics of structured and unstructured data:

## Structured data



- Defined data types
- Most often quantitative data
- Easy to organize
- Easy to search
- Easy to analyze
- Stored in relational databases & data warehouses
- 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, data warehouses, 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:
- 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.

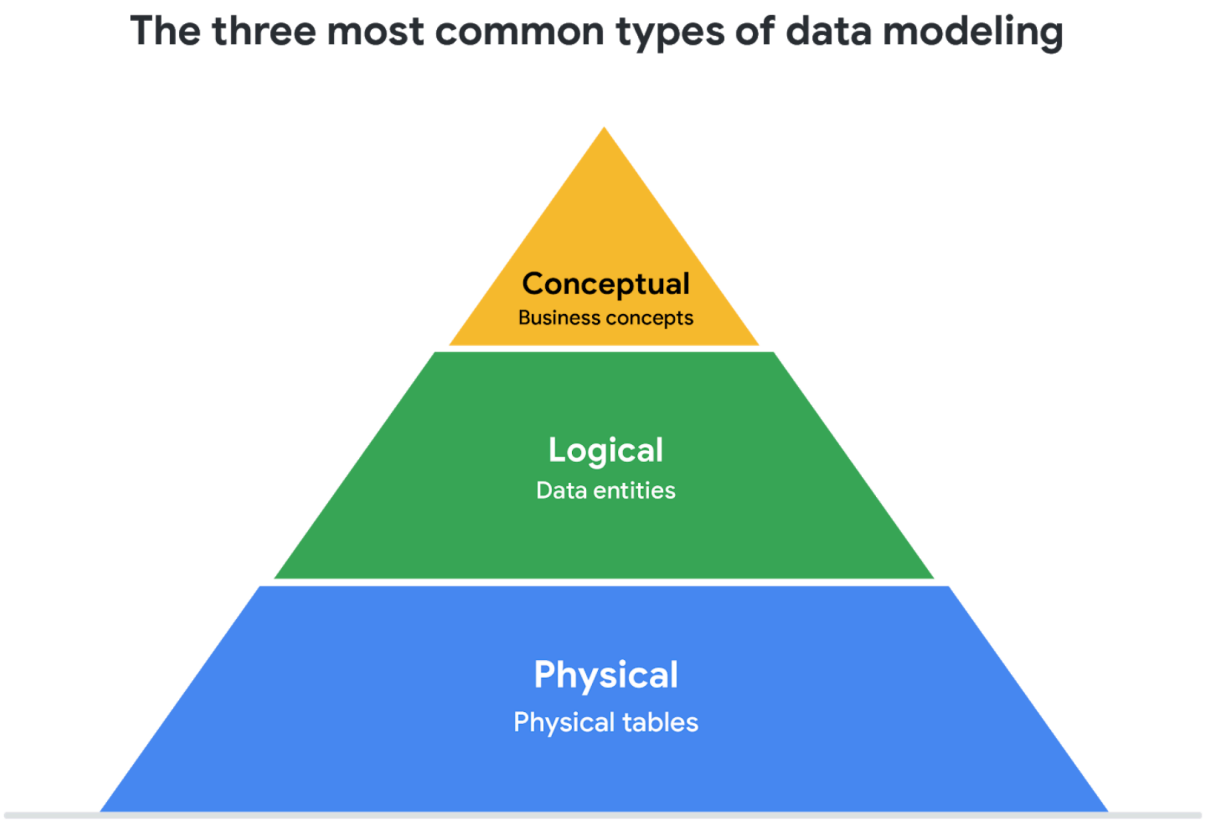
**Important note:** As a junior data analyst, you won't be asked to design a data model. But you might come across existing data models your organization already has in place.

## 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.

## 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.

## 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!