## 1. Data warehouses and data lakes

Great job on these exercises! Now it's time to clarify some concepts.

## 2. Warehouses with stunning view on the lake

Remember the data pipelines lesson at the end of Chapter 1? We quickly mentioned data lakes. Along the course we also mentioned databases several times. In the first lesson of the course, we mentioned data warehouses. So what are these and what is the difference?

## 3. Data pipeline

That's the topic of this lesson. First, let's look at our data pipeline again.

## 4. Data lakes and data warehouses

As the data pipeline graph shows, the data lake is where all the collected raw data gets stored, just as it was uploaded from the different sources. It's unprocessed and messy. While the data lake stores all the data, the data warehouse stores specific data for a specific use. For example, users and their subscription type, or all the listening sessions for behavioral analysis. For this reason, a data lake can take petabytes of data, but warehouses are usually pretty small - small on the scale of big data, I mean. It can still way bigger than your external hard drive. A data lake can store any kind of data, whether it's structured, semi-structured or unstructured. This means that it does not enforce any model on the way to store the data. This makes it cost-effective. Data warehouses enforce a structured format, which makes them more costly to manipulate. However, this lack of structure also means it's very difficult to analyze. Some big data analytics using deep learning can be implemented to discover hidden patterns and trends, but that's about it, and should probably be last resort. The data warehouse, on the other hand, is optimized for analytics to drive business decisions. Because no model is enforced in data lakes and any structure can be stored, it is necessary to keep a data catalog up to date. Data lakes are used by data scientists for real-time analytics on big data, while data warehouses are used by analysts for ad-hoc, read-only queries like aggregation and summarization.

## 5. Data catalog for data lakes

A data catalog is a source of truth that compensates for the lack of structure in a data lake. Among other things, it keeps track of where the data comes from, how it is used, who is responsible for maintaining it, and how often it gets updated. It's good practice in terms of data governance (managing the availability, usability, integrity and security of the data), and guarantees the reproducibility of the processes in case anything unexpected happens. Or if someone wants to reproduce an analysis from the very beginning, starting with the ingestion of the data. Because of the very flexible way data lakes store data, a data catalog is necessary to prevent the data lake becoming a data swamp. It's good practice to have a data catalog referencing any data that moves through your organization, so that we don't have to rely on tribal knowledge, which makes us autonomous, and makes working with the data more scalable. We can go from finding data to preparing it without having to rely on a human source of information every time we have a question.

## 6. Database vs. data warehouse

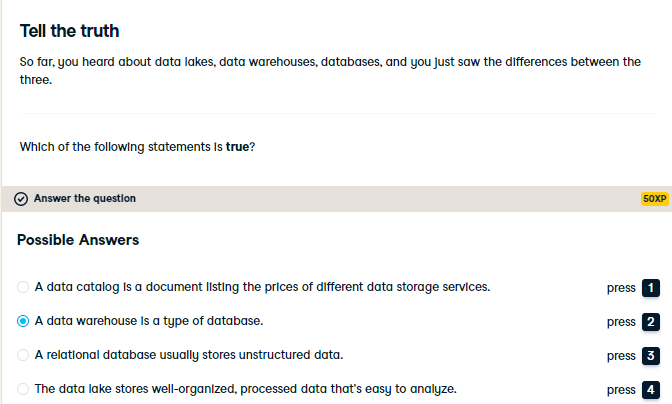
Let's take a step back. We've used the term database throughout the course several times. Where does it fit in? Database is a very general term that can be loosely defined as organized data stored and accessed on a computer. It's a general term and a data warehouse is a type of database.

## 7. Summary

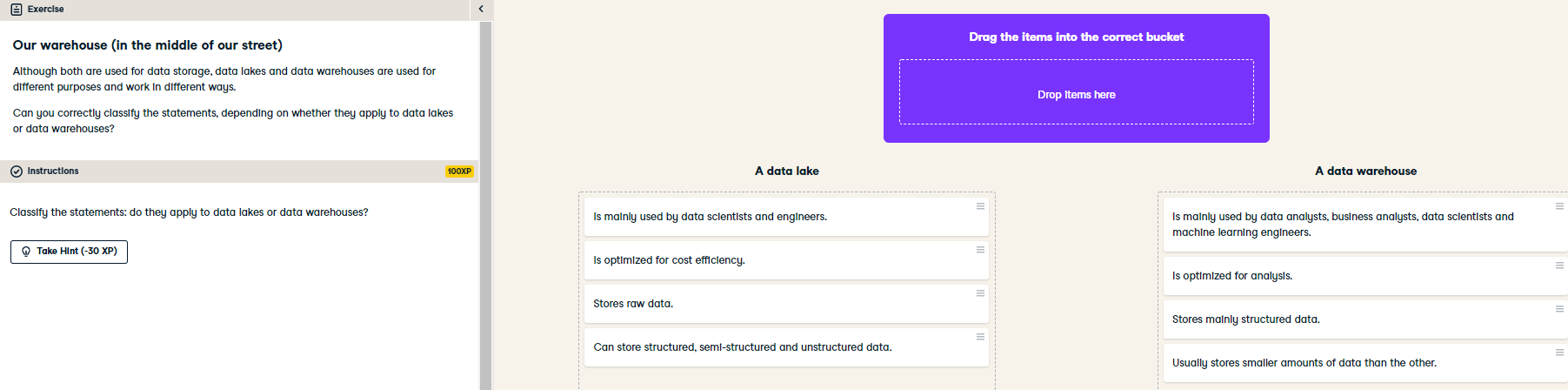
All right! Now you know the characteristics of data lakes, data warehouses and databases, how they differ, and why a data catalog is useful and necessary.

## 8. Let's practice!

Let's end this chapter by cementing your knowledge with a few exercises. Then, onwards to Chapter 3 to learn more about moving and processing data.



Right! A "database" is a very general term, and a data warehouse is a type of database. Data warehouses also have subsets, like data marts, which are highly curated for a particular community of users, such as a specific team Data marts are also much smaller, tens of gigabytes instead of the hundreds of gigabytes to petabytes of data that can be held in a data warehouse.



Excellent job! You definitely know how to distinguish a data lake from a data warehouse, when it's better to use one versus the other, what they are optimized for and who uses them.