Data storage and retrieval

1. Data storage and retrieval

00:00 - 00:00

Previously in this chapter, you learned about different data sources and data types.

2. The data science workflow

00:00 - 00:00

Now, let's discuss efficient ways of storing and retrieving the data that was collected. As you can see this is still part of the first step in the data science workflow we defined before.

3. Things to consider when storing data

00:00 - 00:00

When storing data there are multiple things to take into consideration. First, we need to determine where we want to store the data. Then, we need to know what kind of data we are storing. And lastly, we need to consider how we can retrieve our data from storage. Let's take a closer look.

4. Location: Parallel storage solutions

00:00 - 00:00

Data science projects could require large amounts of data. At this point the data probably can't be stored on a single computer anymore. In order to make sure that all data is saved and easy to access, it is stored across many different computers. Large companies often have their own set of storage computers, called a “cluster” or a “server”, on premises.

5. Location: The cloud

00:00 - 00:00

Alternatively, you could pay another company to store data for you. This is referred to as “cloud storage”. Common cloud storage providers include Microsoft Azure, Amazon Web Services, or AWS, and Google Cloud. These services provide more than just data storage; they can also help your organization with data analytics, machine learning, and deep learning. For now, we’ll just focus on data storage.

6. Types of data storage

00:00 - 00:00

Different types of data require different storage solutions. Some data is unstructured, like email, text, video and audio files, web pages, and social media messages. This type of data is often stored in a type of database called a Document Database.

7. Types of data storage

00:00 - 00:00

More commonly, data can be expressed as tables of information, like what you might find in a spreadsheet. A database that stores information in tables is called a Relational Database. Both of these types of databases can be found on the cloud storage providers that were mentioned earlier.

8. Retrieval: Data querying

00:00 - 00:00

Once data has been stored in a Document Database or a Relational Database, we'll need to access it. At a basic level, we'll want to be able to request a specific piece of data, such as "All of the images that were created on March 3rd" or "All of the customer addresses in Montana". In addition, we might even want to do some analysis, such as summing, counting, or averaging data.

9. Retrieval: Data querying

00:00 - 00:00

Each type of database has its own query language; Document Databases mainly use NoSQL, while Relational Databases mainly use SQL. SQL stands for "Structured Query Language" and NoSQL stands for "Not only SQL".

10. Putting it all together: Location

00:00 - 00:00

Storing your data is like building a library. First, you need to decide where to build your library. That corresponds to choosing a location: either an on-premises cluster or one of the cloud providers we discussed before: Azure, AWS, or Google Cloud.

11. Putting it all together: Data type

00:00 - 00:00

Next, you need to decide what types of shelves to install to store your books. The types of shelves will depend on the types of books.

12. Putting it all together: Data type

00:00 - 00:00

This is analogous to choosing between a Document Database for unstructured data or a Relational Database for tabular data. Just like a library might have multiple types of shelves, you might need to have some data stored in a Document Database and other data stored in a Relational Database.

13. Putting it all together: Queries

00:00 - 00:00

Finally, you'll need a system for referencing and checking out books. The way you locate and retrieve each book depends on how that book is stored.

14. Putting it all together: Queries

00:00 - 00:00

Similarly, you need a query language to speak to the database. For Document Databases, we generally use NoSQL, and for Relational Databases, we generally use SQL.

15. Let's practice!

00:00 - 00:00

Now that you understand different ways of storing data, let's practice!