16.8.1

Database Versus Data Storage

Your client has all the data already, but they do not have it uploaded to a database. They maintain all their raw data in AWS's S3. You explain to Jennifer that data storage is a great place for data collection, because it allows many types and formats of data without having to slow down the intake for processing.

Data storage holds raw data such as CSVs, Excel files, and JavaScript Object Notation (JSON) files. Think of your own computer file system where you keep a ton of files as data storage. This data doesn't need to be queried and analyzed for business decisions. The files still have structure and can be reviewed, but not nearly as efficiently as a database.

A database contains cleaned, related information in tabular form. This database has been carefully planned and structured so that data can be analyzed efficiently through queries. Doing so comes at a cost of processing data to fit all the rules and structures.

Data storage is a place where large amounts of raw data can be kept without any wrangling or curating. Data storage allows us to keep data of different types or data we might want to parse in the future.

The benefit of having dedicated data storage is that nothing limits the intake of data. Data can flow in constantly and be saved without having to worry if it fits the criteria of the database. We have seen this with our extract, transform, and load (ETL) process—the data storage can hold raw files, such as CSV or JSON, for different needs.

AWS's S3 is a popular data storage service that we'll cover next.