16.1.1

What Is Big Data?

For this project you'll be partnering with Jennifer, an account manager at BigMarket. SellBy, your client, loves to talk about the power of big data, but Jennifer isn't a data expert. So you start off the project by giving her a quick overview of what big data actually is.

What exactly constitutes big data? At what point does data become "big"? Is it just the size? A good rule of thumb to apply is this: Data is considered big data when it exceeds the capacity of operational databases.

For example, a retail item might have hundreds of reviews but not be big data because a local machine can parse it. However, consider that there are tens of thousands of items for sale on Amazon and each has hundreds of reviews. In this case, you have big data.

Some additional examples of big data are listed below:

- **Financial Industry Regulatory Authority (FINRA)** stores **37 billion financial records (https://aws.amazon.com/solutions/case-studies/finra/)** daily and analyzes trends over a period of days, weeks, and months.
- **McDonald's** collects transactional data as it serves 69 million people each day in more than **100 countries (https://corporate.mcdonalds.com/corpmcd/about-us/around-the-world.html)**.
- **Netflix** collects ratings, searches, watch dates, device information, pause and skip data, repeat views, and additional information for more than **158 million subscribers (https://aws.amazon.com/solutions/case-studies/netflix/)**.

↻ Retake

Four Vs of Big Data

There are four characteristics of big data:

- **Volume** refers to the size of data (e.g., terabytes of product information). For instance, a year's worth of stock market transactions is a large amount of data.
- **Velocity** pertains to how quickly data comes in (customers across the world purchasing every second). As an example, McDonald's restaurants are worldwide with customers buying food at a constant rate, so the data comes in fast.
- **Variety** relates to different forms of data (e.g., user account information, product details, etc.). Consider the breadth of Netflix user information, videos, photos for thumbnails, and so forth.
- **Veracity** concerns the uncertainty of data (e.g., reviews might not be real and could come from bots). As an example, Netflix would want to verify whether users are actively watching the shows, falling asleep, or just playing them in the background.

The four Vs of big data will help you determine when to migrate from regular data to big data solutions.

Big Data Problems

Working with datasets of this size creates unique challenges. How will we store all of this data? How can we access it quickly? How do we back up this type of data?

You and Jennifer certainly have your work cut out for you, but you know that learning how to work with datasets of this size will help you give the best recommendations to your client.

We'll cover all of these issues in the module.