"In the cloud" is an oft-repeated phrase these days, and nearly every technology used promises cloud access. What does it mean? It's not technology "floating in the clouds," but rather computing hosted on a shared virtual environment, interconnected to a massive storage facility with contained servers, storage, and different sites.

Amazon Web Services (AWS) is the largest cloud provider in the market today. It's not the only one—many companies, such as Google and Microsoft, also have their own cloud platforms. We'll focus on AWS, but the services offered will be similar across all solutions.

Cloud services, such as AWS, allow a company to scale easily. Before cloud services, you would need to buy and store your own servers, which is costly, time-consuming, and requires physical space. A few considerations: How much to buy? What if you miscalculated your growth rate and spent too much or too little on servers? Cloud services allow the flexibility to adjust on the fly for budget and resources.

Before we begin, understand we'll be using AWS's set of free tier options. This way, we can learn and practice on the platform without incurring costs. AWS and other cloud services charge sneaky costs that quickly add up, so it's important to always follow the directions and use the free tier services. At the end of this module, we'll perform a full cleanup of all the services we used to ensure nothing is left in the cloud.

AWS offers a wide variety of storage options on its platform, including both structured and unstructured databases. We'll go through setting up a Postgres database using AWS's relational database service (RDS), but note that there are many databases to choose from.



Since PySpark is a big data tool, it has many ways of reading in files from data storage so that we can manipulate them. We have decided to use S3 as our data storage, so we'll use PySpark for all data processing.

Using PySpark is how we've been reading in our data into Google Colab so far. The format for reading in from S3 is the S3 link, followed by your bucket name, folder by each folder, and then the filename, as follows:

For US East (default region)



