16.8.3

PySpark and S3 Stored Data

Jennifer now understands the difference between databases and data storage and is really enjoying using S3 for storage. She's now wondering how to load in your data from storage. You remind her where we began with big data and explain that PySpark is the perfect tool for loading in stored data.

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)

```
template_url = "https://<bucket-name>.s3.amazonaws.com/<folder-name>/<file-name>"

example_url = "https://dataviz-curriculum.s3.amazonaws.com/data-folder/data.csv"
```

For other regions

```
template_url = "https://<bucket-name.s3-<region>.amazonaws.com/<folder-name>/<file-name>"

example_url =" https://dataviz-curriculum.s3-us-west-1.amazonaws.com/data-folder/data.csv"
```

SKILL DRILL

Load in a dataset to one of your S3 buckets, and then, using Google Colab, load in your dataset to a DataFrame.