Let's run through a mock scenario using two different types of raw data stored in S3. Our goal is to get this raw data from S3 into an RDS database. Let's start with uploading files to your own S3 bucket:

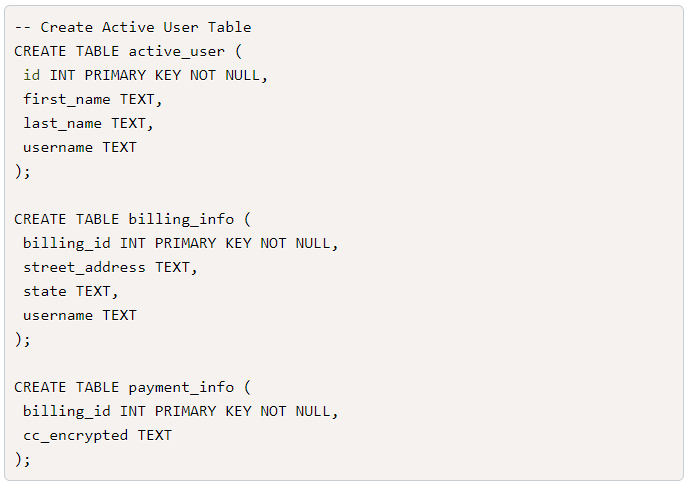
Create an S3 bucket, and then load the following files into the bucket:

* [user\_data.csv](https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/module_16/user_data.csv)
* [user\_payment.csv](https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/module_16/user_payment.csv)

**IMPORTANT**

Remember to make the bucket and files public.

Assume your company already has three tables set up in the RDS database and would like to get the raw data from S3 into the database. Create a new database in pgAdmin called "my\_data\_class\_db." We'll have it represent the company database by first running the following schema in pgAdmin for our RDS:



**NOTE**

Table creation is not part of the ETL process. We're creating the tables to represent a pre-established database you need for the raw data. In a real-life situation, databases will already have a well-defined schema and tables for you, as the engineer, to process data into.

Start with creating a new notebook, installing Spark: See Spark and Postgress.