SCHOOL OF DATA
by Springboard

Data Engineering Career Track

Open-ended Capstone Step 3: Collect Your Data

Estimated time: **2 - 4 Hours**

Overview

It's time to collect the data you'll use for your Open-ended Capstone Project!

While working on this capstone, you'll apply your structured thinking skills to a **real-world** dataset. You've already selected on the dataset that you are going to use in this capstone. By now, your mentor should have approved your project proposal - if they have not, please wait until you have been approved before beginning this step.

Project Steps

Step One: Acquire the Dataset

In this step, you must write a Python script to acquire the data and store it locally on your machine (or on a harddrive). If you do not have enough local storage for the entire dataset, figure out a way to acquire and store a subset of the data on your machine. You will need it to model your pipeline later in the project.

This will build off of the skills you've been learning in the course. If you are unsure how to proceed, revisit some of the earlier units.

Step Two: Submit Your Dataset And Discuss It With Your Mentor

Once you've acquired the dataset that interests you, submit a link to the location of the dataset online, and push the Python script you used to acquire and store the data to your GitHub repo. Talk to your mentor about it during your next call.

Please note: It may be worthwhile to consider creating **more than one dataset** so that you can discuss the pros and cons of each during your mentor call.