Data Engineering Bootcamp
Open-ended Capstone Step 5: Prototype Your Data Pipeline

Estimated Time: 12-18 Hours

Now that you've acquired, explored, and cleaned your data, it is time to prototype your data pipeline! This is one of the most important tasks that data engineers do in their daily work. You'll prototype the pipeline first, and then deploy it to the cloud in later phases.

For this step, please use a volume of data that your local computer can handle (2-3 GB). You'll likely need to slice the dataset in order to select the appropriate volume.

In this step of the project you will:

1. Use Python to automate the data acquisition, cleaning/transformation, and storage process
    a. Write Python scripts to acquire (download) data from source in an automated manner. Any authentication / authorization step should be automated as well. You'll want to schedule this step later to run without supervision.
    b. Create Python scripts for cleaning/transforming the data and trigger them once data acquisition is completed
    c. Then create a Python script which writes the data out to a desired location, in local storage or in the cloud, after data cleaning/transformation completes
    d. Implement logging for relevant information, warnings, and errors at each of the steps above
    e. Make sure your code follows the OOP concepts i.e. you have your functionalities divided into classes, functions and attributes.
2. Design the pipeline to run without any user intervention

    a. The goal is to have one master OOP script which can call the activities in order without any user input. This is similar to what you will do in the real world where the data pipelines are automated and the developer intervenes only when there is an issue.
3. Log the metrics of the pipeline (filename, num of columns, num of rows, errors, warnings, info, etc) to a local storage

Deliverables:

- Push your Python code to your project's GitHub repo
- Update the README Markdown file

Slide deck updates:

From this step, you should include:

- Choices you had to make about any cleaning/transformation of the data you perform in your prototype
- Choices you made about the automation of your data pipeline that impact its performance or reliability