Grace Brown and Emma Mills

20 October 2024

DS 2002 Data Project One-Pager

During the development of our ETL pipeline, we faced a few challenges regarding API connectivity, authentication, processing of both the JSON and CSV formats, and handling large datasets. With both datasets deriving from Kaggle, it was important for us to ensure Kaggle API was set up correctly and functioned smoothly. Secondly, handling both JSON and CSV formats posed a challenge as converting formats and maintaining consistency in the naming conventions were extremely important. We had to ensure that all column names and types aligned with the CSV, JSON, or SQL format to reduce any errors or misrepresentation of data. Finally, the largest challenge we encountered was using a large data set. In the JSON to CSV data, the original database used was a powerful 10 GB COVID data source that required significant run time and processing. While the data was very interesting, it was necessary to switch to a smaller dataset in order to improve the project’s overall performance and efficiency.

In terms of easier aspects, the integration of both pandas and SQLite made interaction with the SQL databases very intuitive and easy to structure. This made switching between formats easier and allowed the columns to be modified fairly quickly. Going forward, a utility like this is very applicable to future data projects we may encounter. This adds flexibility to not be stuck working with one set format of data, and have the ability to switch between JSON formats to CSV and CSV to JSON. This provides us with the knowledge necessary to re-integrate data to be in a system that is identically formatted, making it easy to add, remove, and modify columns within one singular pipeline.