Data Engineering Capstone Project

Overview

The purpose of the data engineering capstone project is to give you a chance to combine what you've learned throughout the program. This project will be an important part of your portfolio that will help you achieve your data engineering-related career goals.

In this project, you can choose to complete the project provided for you, or define the scope and data for a project of your own design. Either way, you'll be expected to go through the same steps outlined below.

Udacity Provided Project

In the Udacity provided project, you'll work with four datasets to complete the project. The main dataset will include data on immigration to the United States, and supplementary datasets will include data on airport codes, U.S. city demographics, and temperature data. You're also welcome to enrich the project with additional data if you'd like to set your project apart.

Open-Ended Project

If you decide to design your own project, you can find useful information in the Project Resources section. Rather than go through steps below with the data Udacity provides, you'll gather your own data, and go through the same process.

Instructions

To help guide your project, we've broken it down into a series of steps.

Step 1: Scope the Project and Gather Data

Since the scope of the project will be highly dependent on the data, these two things happen simultaneously. In this step, you'll:

- Identify and gather the data you'll be using for your project (at least two sources and more than 1 million rows). See Project Resources for ideas of what data you can use.
- Explain what end use cases you'd like to prepare the data for (e.g., analytics table, app back-end, source-of-truth database, etc.)

Step 2: Explore and Assess the Data

- Explore the data to identify data quality issues, like missing values, duplicate data, etc.

- Document steps necessary to clean the data

Step 3: Define the Data Model

- Map out the conceptual data model and explain why you chose that model
- List the steps necessary to pipeline the data into the chosen data model

Step 4: Run ETL to Model the Data

- Create the data pipelines and the data model
- Include a data dictionary
- Run data quality checks to ensure the pipeline ran as expected
    - Integrity constraints on the relational database (e.g., unique key, data type, etc.)
    - Unit tests for the scripts to ensure they are doing the right thing
    - Source/count checks to ensure completeness

Step 5: Complete Project Write Up

- What's the goal? What queries will you want to run? How would Spark or Airflow be incorporated? Why did you choose the model you chose?
- Clearly state the rationale for the choice of tools and technologies for the project.
- Document the steps of the process.
- Propose how often the data should be updated and why.
- Post your write-up and final data model in a GitHub repo.
- Include a description of how you would approach the problem differently under the following scenarios:
    - If the data was increased by 100x.
    - If the pipelines were run on a daily basis by 7am.
    - If the database needed to be accessed by 100+ people.

Rubric

In the Project Rubric, you'll see more detail about the requirements. Use the rubric to assess your own project before you submit to Udacity for review. As with other projects, Udacity reviewers will use this rubric to assess your project and provide feedback. If your project does not meet specifications, you can make changes and resubmit.