**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**](https://review.udacity.com/#!/rubrics/2497/view), 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.

**Datasets**

Gathering the right data is one of the most important task for data engineers. If you are looking for data for your own, self-directed project, or for data to enrich the data for the Udacity provided project, below are some good resources to find interesting datasets.

**Analytics**

* [**Google: Dataset Search**](https://toolbox.google.com/datasetsearch)
* [**Kaggle Datasets**](https://www.kaggle.com/datasets)
* [**Github: Awesome Public Datasets**](https://github.com/awesomedata/awesome-public-datasets)
* [**Data.gov**](https://catalog.data.gov/dataset)
* [**Dataquest: 18 places to find data sets for data science projects**](https://www.dataquest.io/blog/free-datasets-for-projects/)
* [**KDnuggets: Datasets for Data Mining and Data Science**](https://www.kdnuggets.com/datasets/index.html)
* [**UCI Machine Learning Repository**](https://archive.ics.uci.edu/ml/datasets.html)
* [**Reddit: r/datasets/**](https://www.reddit.com/r/datasets/)

**Software Development**

* [**Github: Public APIs**](https://github.com/toddmotto/public-apis)
* [**Last Call: Top 50 Most Popular APIs on RapidAPI (2018)**](https://blog.rapidapi.com/most-popular-apis/)
* [**Facebook: Graph API**](https://developers.facebook.com/docs/graph-api)

**Example**

[**This Kaggle dataset**](https://www.kaggle.com/hugomathien/soccer/home) is a good example of someone who found disparate datasets and combined them to provide an even more valuable dataset for others to analyze.

**Udacity Provided Project**

If you decide not to create your own capstone project, you can use the one we've provided. For this project, we have some datasets available to you and some ideas for the project. However, it's still open-ended in nature and you'll have to define the data model and the corresponding use cases for your final deliverable.

**Datasets**

The following datasets are included in the project workspace. We purposely did not include a lot of detail about the data and instead point you to the sources. This is to help you get experience doing a self-guided project and researching the data yourself. If something about the data is unclear, make an assumption, document it, and move on. Feel free to enrich your project by gathering and including additional data sources.

* **I94 Immigration Data:** This data comes from the US National Tourism and Trade Office. A data dictionary is included in the workspace. [**This**](https://travel.trade.gov/research/reports/i94/historical/2016.html) is where the data comes from. There's a sample file so you can take a look at the data in csv format before reading it all in. You do not have to use the entire dataset, just use what you need to accomplish the goal you set at the beginning of the project.
* **World Temperature Data:** This dataset came from Kaggle. You can read more about it [**here**](https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data).
* **U.S. City Demographic Data:** This data comes from OpenSoft. You can read more about it [**here**](https://public.opendatasoft.com/explore/dataset/us-cities-demographics/export/).
* **Airport Code Table:** This is a simple table of airport codes and corresponding cities. It comes from [**here**](https://datahub.io/core/airport-codes#data).

**Accessing the Data**

Some of the data is already uploaded to the workspace, which you'll see in the navigation pane within Jupyter Lab. The immigration data and the global temperate data is in an attached disk.

**Immigration Data**

You can access the immigration data in a folder with the following path: ../../data/18-83510-I94-Data-2016/. There's a file for each month of the year. An example file name is i94\_apr16\_sub.sas7bdat. Each file has a three-letter abbreviation for the month name. So a full file path for June would look like this: ../../data/18-83510-I94-Data-2016/i94\_jun16\_sub.sas7bdat. Below is what it would look like to import this file into pandas. Note: these files are large, so you'll have to think about how to process and aggregate them efficiently.

**fname** = '../../data/18-83510-I94-Data-2016/i94\_apr16\_sub.sas7bdat'

df = pd.read\_sas(fname, 'sas7bdat', encoding="ISO-8859-1")

The most important decision for modeling with this data is thinking about the level of aggregation. Do you want to aggregate by airport by month? Or by city by year? This level of aggregation will influence how you join the data with other datasets. There isn't a right answer, it all depends on what you want your final dataset to look like.

**Temperature Data**

You can access the temperature data in a folder with the following path: ../../data2/. There's just one file in that folder, called GlobalLandTemperaturesByCity.csv. Below is how you would read the file into a pandas dataframe.

fname = '../../data2/GlobalLandTemperaturesByCity.csv'

df = pd.read\_csv(fname)