**ETL Project - To Find Chinese Restaurants in New York City and San Francisco**

* The sources of data that you will extract from.

|  |
| --- |
| Two sources are used in this ETL project:  https://www.kaggle.com/datasets?search=New+york+restaurant  https://maps.googleapis.com/maps/api/place/nearbysearch/json |

* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

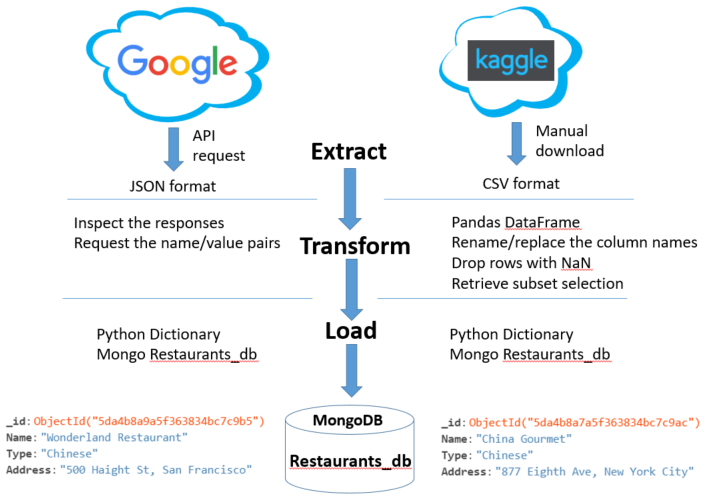
|  |
| --- |
| cleaning & filtering |

* The type of final production database to load the data into (relational or non-relational).

|  |
| --- |
| Relational |

* The final tables or collections that will be used in the production database.

|  |
| --- |
| MongoDB |



 **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

Manually downloaded the restaurant information in csv format for New York City from www.kaggle.com, and ran the Python requests against maps.googleapis.com to retrieve only Chinese restaurant information in json format for San Francisco.

 **T**ransform: what data cleaning or transformation was required.

In order to easily work with data in csv format, we used pandas to convert the comma-delimited file into a data frame for a table-like view. For the dataframe of NYC, we dropped any row whose type was NaN and replaced the space in between a column name with a underscore. In addition, we also renamed the column name from Compay\_Name to Name and Sub\_Subindustry to Type. Because we only wanted the restaurant name, type, and address, we filtered the New York City’ restaurants by the type that was Chinese while we used keyword of Chinese to retrieve San Francisco’s Chinese restaurants via an api call.

 **L**oad: the final database, tables/collections, and why this was chosen.

We chose MongoDB because it was easy to work with. Unlike the PostgreSQL database, you do not have to create a database and/or a table first by using pgAdmin4 before you can insert the data. In MongoDB, if the database/table does not exist, it will create it on the fly as the script kicks off. Another reason is when installing MongoDB software, a GUI program called MongoDB Compass Community becomes available to help the easy management of the database/table. We use Python dictionary to store each restaurant’s name, type, and address and use a collector to insert that record into the Mongo database called Restaurants\_db in a FOR loop.