Project 2 ETL Technical Report by Team 3

Anthony Stevens, Cecilia Zhang, Neda Mehdizadeh, and Sahar Jamal

Project Theme : Covid-19 in the USA

Step 1. Extract: Data Sources

- Covid-19 in USA-Number of Novel CoronaVirus 2019 cases in USA from Kaggle.com 3 csv files (primary)
  Reference Link:
  https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_counties_covid19_daily.csv
- State, County, CBSA, Country API Data from Covid Act Now (primary)
  Reference Link: https://apidocs.covidactnow.org/data-definitions/

Step 2. Transform: Data Wrangling

- Covid daily data is retrieved from “us_counties_covid19_daily”csv file, then encoded the fips (Federal Information Processing System) column to an int instead of a float. After that the null values were dropped to keep only usable data.

- The original data is scraped from json file “counties. timeseries” and is cleaned by selecting columns with useful information. Then dropped columns with NaN values such as “lat” or contained dictionaries which were not useful.. Leaving us with a handful of useful data. We ran into an issue with two of the columns throwing errors for “undefined columns”. The solution was to rename and lowercase the names of the columns.

Step 3. Loading:

We chose PostgreSQL as our database to load our data. The reason for this was due to this database being a common sql database, so we knew the expertise in this area would definitely challenge us to perfect our skills. After tackling the portion of data we decided to keep in pandas, we then lined up a table within postgres. To do so we used the <CREATE TABLE> statement, and then we identified each column we wanted in the table. The next step was to run the table and check back for the table to upload. We then entered into pandas and connected the engine so that the data would load into the database and align nicely. To confirm our work was done correctly, we utilized the <SELECT *> of the specific table to print what had loaded.