**ETL Project**

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Kaggle has data on this virus that originates from a dashboard put together by Johns Hopkins University.

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

Our data sources include 4 CSV files from Kaggle that include daily level information on the number of 2019-nCoV affected cases, deaths and recovery from 2019 novel coronavirus across the globe.

We started the data extraction of these files by first reading the CSV Files using Pandas (pd.read\_CSV()).

**Transform**: *what data cleaning or transformation was required.*

Next, we transformed the data in Pandas by dropping the columns we did not want to keep as well as by making sure the dates were in a consistent format as it was mentioned in the Kaggle discussion board that the date format is different for certain dates. In addition, we added an Id column and set it to index. This will be beneficial when setting primary keys for future queries and joins.

**Load**: *the final database, tables/collections, and why this was chosen.*

Finally, we uploaded our transformed data table to pgAdmin database by using Postgres SQL within Pandas. We chose pdAdmin since this is a relational database and we have a SQL compatible dataset. Our 5 tables were created as follows:

Corona\_data: Daily level information on the number of 2019-nCoV affected cases across the globe

Confirmed\_data: Time series data of confirmed cases

Deaths\_data: Time series data of death cases

Recovered\_data: Time series data of recovered cases

Total\_cdr\_country: Groupby Country sum of Confirmed cases, Deaths cases, and Recovered cases

**Analytics Opportunity from Tables**:

These tables will allow a user to evaluate the trends of infections over time while being able to calculate the case fatality rate (Death Case volume divided by Confirmed Case Volume). It appears the case fatality rate of nCoV is about 2-4% which means the infection is NOT as deadly as Ebola or Nipah at this time.

Analyzing these tables will also reveal % change in Confirmed case overtime, so that we can see if the virus is spreading at an increasing rate and if certain locations of a higher spread rate.