**META-B ETL Project Report**

**Sources for Data Extraction**

* Yahoo Finance historical stock price data : [LOBLAW CO (L.TO) Stock Historical Prices & Data](https://ca.finance.yahoo.com/quote/L.TO/history?p=L.TO)
* European Centre for Disease Prevention and Control (ECDC) Covid-19 cases data: <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>
* Identified the Bank of Canada Valet API: <https://www.bankofcanada.ca/valet/docs> as a reliable dataset to look at the financial impact of the COVID-19 pandemic through the indicator of the US – Canadian Exchange Rate.

**Method of Data Extraction**

* Yahoo Finance Data:
  + We used pandas to read the html tables of the Yahoo finance website
  + A function was created to generate historical daily stock data by month in a table to mitigate Yahoo Finance’s limit of 100 data points when scraping.
* Covid-19 Cases Data:
  + We used pandas to read csv file downloaded from ECDC website
* USD x Canadian Exchange Rates:
  + Used Jupyter Notebook to call the Bank of Canada Valet API
  + Used documentation to give the API call a specific query to pull the daily observations of the US-Canadian Exchange Rate in JSON format

**Data Transformation Process**

* Yahoo Finance Data:
  + We converted our list of monthly tables into dataframes
  + These data frames were appended to final dataframe for the stock that contains pricing information of the stock for 2020
  + Filtered the stock dataframe to select the date, the open price, the close price.
  + Converted the data types for the filtered dataframe
  + Created a new column that for the daily percentage changes
  + Used an inner join to consolidate all the individual stock dataframes
* Covid-19 Cases Data:
  + Collected a list of all columns we need
  + Filtered data for only Canada & USA ,reorganize and rename the columns
  + Grouped the dataframe by Data to get the final dataframe
* USD x Canadian Exchange Rates:
  + Converted the exchange rate JSON into a DataFrame
  + Renamed the columns
  + Filtered by selecting the specific timeframe desired (filtered to only show from 2020-01-01 to 2020-08-13)
  + Set the “Date” column as the Index
  + Normalized the column holding the exchange rate values as it was set to the dtype of object
  + Replaced the non-normalized Exchange Rate Column with the normalized Exchange Rate column by adding and dropping columns
  + Achieved data in desired transformed state

**Data Loading Process**

* Chose to use SQL (pgAdmin4), a relational database, to store our data
* Created a relationship diagram for our combined database (titled “covid\_finance\_db” (see script.sql)) through [https://app.quickdatabasediagrams.com](https://app.quickdatabasediagrams.com/#/) and exported the DB schema diagram to PostgreSQL
* Adjusted the column names in SQL to ensure they matched the column names for their associated tables that we had created previously in the Jupyter Notebook file
* Used Jupyter Notebook to connect our engine and uploaded data to SQL: COVID-Cases, USDxCAD\_ExchangeRate and Stock\_Dataframe into the three existing tables in a SQL database.

Data has now been **E**xtracted, **T**ransformed and **L**oaded in!