Team Delta – ETL Process

Our team decided to pursue the financial track of data analysis for the ETL project. We found a dataset using Kaggle that essentially has CSV files of the top 100 technology companies stock history. The dataset can be found at:  
<https://www.kaggle.com/datasets/tomasmantero/top-tech-companies-stock-price> . We took a look at the available tech companies and decided that we were going to use Microsoft and Apple for our project.

We started the ETL process by reading in our respective CSV files for Apple as well as Microsoft. We then used the .info() command to get the column names as well as the respective data types of each column. Once done, we put our columns as well as the values of our columns into a list. This allowed us to create a data frame for each respective stock; Apple and Microsoft.

Once we created the DataFrame, we were able to fully see our dataset and decided that we needed to transform it to make it more readable. We decided that our first order of business was going to be dropping the “Adj Close” column. We decided to delete this column because we did not feel that it really provided much context into stock values. After removing the adjusted close column, we felt that the stock values were hard to read as they each were rounded 6 decimal places. To remedy this, we decided to round those respective values to 2 decimal places, making them much easier to read.

After creating the DataFrame, one problem we found is that we did not have a “Ticker” column which showed which stock we were currently looking at. We decided to add a “Ticker” column not only for readability but also as a primary key to perform a join on the two datasets.

When it came time to load the data into a database, we initially decided that we would use MongoDB. After careful consideration, we decided that would not make much sense as Mongo is used for non-relational databases. We felt that our data was all relational so we made the decision to load our data into PostGreSQL. We then created our classes as well as our connection to the Database. Since we were reviewing technology stocks, we decided to name our database “Tech\_Stock”. We then committed the Apple as well as Microsoft DataFrames to the “Tech\_Stock” database. Lastly, we were able to join the two DataFrames in PostGreSQL using an inner join on the “Ticker” column that we added.