Data Cleanup & Analysis

Team ETL

AUDREY LAWRENCE, john shows, jerry tilenis

Volume Trading Data on August 7, 2020 for NASDAQ and NYSE

Aggregate High Volume Trades and Total Volume Trades for Select Stocks

# Extract

Team ETL pulled total volume activity for a specific set of identified stocks from the NASDAQ and instances of high volume activity (>80,000 shares per transaction) for the same set of stocks for all exchanges from Blackbox:

<https://www.blackboxstocks.com/>: All Exchanges data

<https://eoddata.com/products/default.aspx>: NASDAQ data

Blackbox and NASDAQ were retrieved as csv files.

# Transform

The following steps were taken to transform the selected data sets:

* The Blackbox data set contained multiple entries of high volume (>80,000) trades for 72 unique stocks completed on August 7, 2020.
* Multiple entries for each unique stock were grouped and summed to produce one high volume activity total for August 7, 2020.
* Details from the original Blackbox dataframe were then merged with the resulting summary data.
* Multiple instances of block purchases for unique stocks were counted and added to the dataframe.
* Column names for the Blackbox data set did not import cleanly and, therefore, returned errors when used for functions. Necessary columns were renamed.
* The Blackbox data set contained several columns of information that were not germane to the final desired data set, therefore, drop column was used to drop these columns from the data set.
* The NASDAQ data set contained total volume transactions for unique entries of over 3000 stocks. Additionally, the data set contained columns that were not germane to our objective.
* Drop column was used to eliminate unneeded columns: “open”, “close”, “high”, “low”.
* The Blackbox data set and the NASDAQ data set were each checked for duplicate entries of unique stock ticker symbols. It was confirmed that neither data set contained any remaining duplicate entries of unique stock tickers symbols.
* The cleaned NASDAQ dataframe and Blackbox data sets were merged on the “Symbol” column matching like variables in the “Symbol” column.
* The resulting data set, “final\_df”, contained the total volume of all transactions for August 7, 2020, for the selected stocks as well as the total of high-volume transactions for the selected stocks on August 7, 2020.
* The columns for the resulting data set, “final\_df” were reordered using “movecol” for more visual appeal.
* Link was established to pgAdmin to the “final\_df” as a SQL table utilizing “create engine”.

Data Cleanup & Analysis

# Load

* The final production database “final\_df” is a relational database.
* The final database was linked to pgAdmin utilizing “create engine”.
* The final table used in the production database is: BlockOrderDailyVolume
* The database “final\_df” was created in Python. Because primary keys cannot be created in Python, the primary key was added in pgAdmin.
* Run the table schema BlockOrderDailyVolume to create the primary key.
* To add the primary key to the database, we used “alter table”.
* pgAdmin was selected as the application for final publication because of it’s ease of use in working with relational databases.
* A relational database was appropriate for use because of the common data found in both the NASDAQ and Blackbox database tables. As a result, the data can easily be sorted based on any field and reports can be generated that report selected desired filed from each record.
* An additional benefit of a relational database is the option to generate new tables from a single query.

# To recreate this process for any given day:

* Visit <https://eoddata.com/products/default.aspx> and request the end of day NASDAQ csv.
* Visit <https://www.blackboxstocks.com/> and request the DarkPool data csv. (subscription required)
* Read csv files to ETL\_Main notebook using pandas
* Run the ETL\_Main notebook and follow the comments in the notebook.
* Within Postgres create a database titled etl-project.
* Install sqlalchemy and psycopg2 to create database connection between Jupyter and Postgres.
* Engine connection will create the BlockOrderDailyVolume table within etl-project database within Postgres.
* If desired, run BlockOrderDailyVolume.sql file to add primary key for “Symbol” and to rename the “Volume” column to “Daily Volume”.

Data Cleanup & Analysis