How to load data into a relational database

There are 3 types of data in this exercise, however they are all in csv format after transformation.

There are 101 constitutes in FTSE100 according to LSE website, their pricing information has been scraped from yahoo finance and saved in Folder: stock\_pricing\_info, with a name of {stock\_code}.csv

Market cap and sector information are also scraped from LSE website and saved in as stock\_info.csv and sector\_info.csv.

In this case, we only have one type of file to import, therefore we will only need to validate csv, however if we have different type information to import, we will need to do more checks.

I have opt for a much complicated task than the original request, therefore with the time limit, I have found it is difficult to build all the functions I initially designed, such as using apache airflow to schedule this job to run daily to capture daily price change, different graphs with sector comparison, a function that can graph different stock/fund against ftse100 performance etc.

Step 1: Data quality check

         First, create a dataframe out of the csv files

         Second, drop unnecessary column, such as High, Low from stock pricing.csv

         Third, check if there is any blank (ghost) line, data type, and null data

         Fourth, create meta data with line numbers, max, min date/price for each file

Step 2: Schema design, primary key is stock\_code, each file and column should be unique, and all columns should be not\_null.

|  |
| --- |
| stock\_info.csv |
| stock\_code |
| market\_cap |



|  |
| --- |
| sector\_info.csv |
| stock\_uniquecode |
| ftse\_industry |
| ftse\_sector |

|  |
| --- |
| {Stockcode}.csv |
| Filename() |
| Date |
| Closing |

Step 3: Select the file to import and preview the data

Step 4: Viewing meta data and compare with meta data generated in step 1 to validate the amount of data being imported