Pipeline Documentation

This data pipeline is extracting Stock Market Data per trigger. Each trigger generates 100 days historical data on a selection of stocks. It is advised to trigger the pipeline in the morning to retrieve data until d-1 included.

Data Extraction

Data extraction process can be found in *src* > *api* > *connectors.py*.

A loop iterates through each symbol stored locally in a list at the initialization of the *src* > *main.py* file.

The loop generates a JSON payload for each symbol. Each payload is stored in the *data* > *raw* folder as JSON files with the following naming convention: *SYMBOL_date_raw.json*

Before populating the raw data folder, the folder is emptied.

File Mangement System

src > *file_management.py* centralizes all the functions to move files through different folders during the processing of the data.

Data Processing

The main file then loops through the initial list again. Each raw data file goes through a series of transformation defined in src > processing > transformers.py.

Here, all the shaping of the data and addition of the KPI's we want to track are detailed.

Each processed payload is now stored in the *data* > *processed* folder as JSON files with the following naming convention: *SYMBOL_date_processed.json*

Before populating the processed data folder, the folder is emptied.

Storage in a MongoDB time series table (local)

The main file loops one last time through the initial list to insert into a collection created with in Mongo Compass all the processed data files after emptying the collection.

All the procedures are defined in *src* > *storage* > *mongo.py*.

NB: To connect Power BI to Mongo DB locally, you will need an ODBC Driver for Mongo DB. I personally used CData MongoDB, but the free trial expires after a month.