Zestaw 4 – Stock Data

Pochodzenie danych to `https://www.kaggle.com/jacksoncrow/stock-market-dataset`

Dane zawierają informacje dotyczące notowań spółek giełdowych począwszy od roku 1962

Zbiór danych

Wykorzystywane są dwa zbiory danych.

Pierwszy, główny, "strumieniowy" to zbiór plików mających format csv i następujące pola:

- `Date` – określa datę notowania
- `Open` – cena otwarcia
- `High` – maksymalna cena w ciągu dnia
- `Low` – cena minimalna w ciągu dnia
- `Close` – cena zamknięcia skorygowana o splity (`https://pl.wikipedia.org/wiki/Split_akcji`)
- `Adj Close` – cena zamknięcia skorygowana zarówno o dywidendy, jak i splity.
- `Volume` - liczba akcji, które zmieniły właściciela w ciągu danego dnia
- `Stock` - symbol spółki giełdowej

Drugi zbiór statyczny `symbols_valid_meta.csv` zawiera następujące pola:

- `Nasdaq Traded` – wskazanie, czy akcje wliczane są do indeksu giełdowego NASDAQ
- `Symbol` – symbol spółki giełdowej
- `Security Name` – nazwa spółki
- `Listing Exchange`,
- `Market Category`, - kategoria przypisana do akcji przez NASDAQ na podstawie wymagań dotyczących notowań. Wartości: Q = NASDAQ Global Select MarketSM, G = NASDAQ Global MarketSM, S = NASDAQ Capital Market
- `ETF` - określa, czy papier wartościowy jest funduszem typu ETF (*exchange traded fund*).
- `Round Lot Size` – wskazuje liczbę udziałów, które składają się na okrągłą partię.
- `Test Issue` – wskazuje, czy zabezpieczenie jest zabezpieczeniem testowym.
- `Financial Status` – wskazuje, kiedy emitent nie złożył w odpowiednim czasie wniosków regulacyjnych, nie spełnił standardów NASDAQ dotyczących ciągłych notowań i / lub złożył wniosek o upadłość. Wartości obejmują: D = wadliwy: emitent nie spełnił wymagań NASDAQ dotyczących kontynuacji notowań, E = zaległe należności: wystawca nie wywiązał się w terminie z należności, Q = bankrut: emitent złożył wniosek o ogłoszenie upadłości, N = normalny (domyślny), G = wadliwy i bankrut, H = wadliwy i zaległe należności, J = zaległe należności i bankrut, K = wadliwe, zaległe należności i bankrut
- `CQS Symbol` – identyfikator zabezpieczenia używanego do rozpowszechniania danych za pośrednictwem źródeł danych SIAC *Consolidated Quotation System* (CQS) i *Consolidated Tape System* (CTS).
- `NASDAQ Symbol` – identyfikator zabezpieczeń używany w różnych protokołach łączności NASDAQ i kanałach danych rynkowych NASDAQ.
- `NextShares`

ETL

Utrzymywanie agregacji na poziomie miesiąca oraz symbolu i nazwy spółki giełdowej. Wartości agregatów to:

- średnia wartość kursu zamknięcia (Close)
- najmniejsza wartość akcji (Low)
- największa wartość akcji (High)
- sumaryczny obrót (Volume)

Częstotliwość aktualizacji danych ma być parametryzowana:

- co jeden dzień – w przypadku działania na bieżącym strumieniu danych,
- co 10 sekund – w przypadku korzystania z historycznego zbioru danych.

Wykrywanie "anomalii"

Wykrywanie "anomalii" ma polegać wykrywaniu znaczących wahań kursu akcji danej spółki, który zostanie zarejestrowany w ciągu podanego czasu.

Program ma być parametryzowany przez:

- D – długość okresu czasu wyrażoną w dniach
- P – stosunek (minimalny) różnicy pomiędzy najwyższym kursem akcji danej spółki a najniższym do najwyższego kursu

Wykrywanie anomalii ma być dokonywane każdego dnia.

Przykładowo, dla parametrów D=7, P=40 program każdego dnia będzie raportował te spółki, dla których w ciągu ostatnich 7 dni stosunek różnicy pomiędzy najwyższym kursem akcji danej spółki a najniższym do najwyższego kursu przekroczył 40%.

Raportowane dane mają zawierać

- analizowany okres - okno (start i stop)
- nazwę spółki
- najwyższy kurs akcji
- najniższy kurs akcji
- stosunek różnicy pomiędzy najwyższym kursem akcji danej spółki a najniższym do najwyższego kursu

Załóż, że dane mogą być nieuporządkowane – mogą być opóźnione o jeden dzień.