Wpływ aktywności na Twitterze na notowania spółek giełdowych

Konspekt projektu

Zespół *Chmurki*
Sebastian Deręgowski, Klaudia Gruszkowska, Bartosz Jamroży

1. Cel projektu i potencjalne korzyści z wdrożenia

Nasz projekt skupia się na analizie wpływu postów z serwisu społecznościowego Twitter na ceny akcji największych, światowych firm.

Why?
Zebrane informacje na temat aktywności wybranych użytkowników lub o postach oznaczonych poszczególnymi hashtagami na Twitterze wraz z kursami akcji kilku wybranych, największych spółek umożliwią ocenę wpływu jednego na drugi. Chcemy pozwolić na przeanalizowanie wielkości wpływów popularnego serwisu społecznościowego na inne dziedziny życia nawet na tak ważne jak ceny akcji.

How?
Docelowym rozwiązaniem prezentowanym końcowemu użytkownikowi jest raport zawierający analizę przetworzonych i połączonych danych.

What?
Projekt składa się faz: pozyskiwania, wstępnego przetwarzania danych i analiz wsadowych. Dane będą pochodzić z dwóch różnych źródeł. Oba pobieramy za pomocą udostępnionych API. Pierwszy zbiór dostarczy nam cen akcji kilku dużych, światowych spółek, a drugi informacji do postach na serwisie Twitter.

2. Opis zbiorów danych planowanych do wykorzystania w projekcie

Pierwszy zbiór danych dotyczy cen akcji kilku spółek notowanych na nowojorskiej giełdzie papierów wartościowych (NYCE). Planujemy analizować jak zmieniają się notowania Google'a, Microsoftu, Apple'a oraz Tesli w interwałach pięciominutowych. Dane będziemy pobierać przy pomocy API ze strony www.alpha vantage.co. Dane będą zwracane w postaci

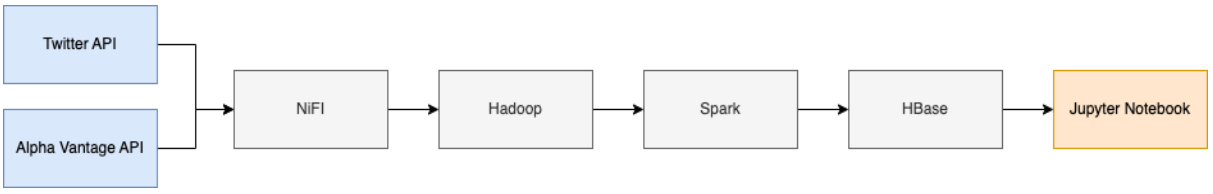
JSONa zawierającego informacje nt. kursu otwarcia, zamknięcia, a także najmniejszej i największej wartości w danym przedziale czasowym.


```
{
    "Meta Data": {
        "1. Information": "Intraday (5min) open, high, low, close prices and volume",
        "2. Symbol": "IBM",
        "3. Last Refreshed": "2022-11-15 16:15:00",
        "4. Interval": "5min",
        "5. Output Size": "Full size",
        "6. Time Zone": "US/Eastern"
    },
    "Time Series (5min)": {
        "2022-11-15 16:15:00": {
            "1. open": "144.3400",
            "2. high": "144.3400",
            "3. low": "144.3400",
            "4. close": "144.3400",
            "5. volume": "4464"
        },
        "2022-11-15 16:05:00": {
            "1. open": "144.3400",
            "2. high": "144.3400",
            "3. low": "144.3400",
            "4. close": "144.3400",
            "5. volume": "159775"
        },
```


Rys. 1. Przykładowy JSON zwrócony przez API

Drugi zbiór danych będzie pochodził z Twittera. Do każdej z analizowanych spółek dobierzemy kilka powiązanych z nią kont twitterowych (np. konto spółki i jej głównych udziałowców), których aktywność będziemy monitorować. Planujemy także zbierać dane z konkretnych hashtagów powiązanych z tymi spółkami. Dane będziemy zbierać poprzez omawiany na zajęciach Twitter API (https://developer.twitter.com/en/docs). Będą zwracane w postaci JSONa zawierającego informacje nt. daty, godziny, tekstu, hashtagów, ilości wyświetleń, udostępnień i polubień, co przyda nam się potem w analityce.

3. Planowany stos architektoniczny



Rys. 2. Architektura systemu

Na początku pobieramy dane z obydwu API przy pomocy Apache NIFI. Następnie dane są składowane w postaci zmerge'owanych plików na Apache Hadoop. Dalej przy pomocy Apache Spark dokonywana jest wsadowa analiza danych i dane są zapisywane na Apache HBase. Jako symulację frontendu/narzędzia BI planujemy wykorzystać Jupyter Notebook, gdzie będą przedstawione wyniki części analitycznej.

4. Planowany podział pracy w zespole

Sebastian Deręgowski:

- opracowanie pobierania danych z notowaniami spółek giełdowych;
- przygotowanie tabel w HBase;
- dwie analizy w Jupyter Notebook.

Klaudia Gruszkowska:

- wstępne przetworzenie danych i załadowanie ich na Hadoop;
- wsadowa analiza danych w Sparku;
- dwie analizy w Jupyter Notebook.

Bartosz Jamroży:

- opracowanie pobierania danych z Twittera;
- stworzenie systemu plików w Hadoop;
- dwie analizy w Jupyter Notebook.