

## **Raport – Projekt Zasadniczy HDiSED**

Hurtownia danych na potrzeby analizy konsumpcji treści cyfrowych

Prowadzący:  
Marcin Gorawski  
Krzysztof Pasterak

Autorzy:  
Cezary Pastor  
Jakub Kuk

## Wprowadzenie

W ramach projektu mieliśmy na celu stworzenie systemu do zbierania danych o firmach związanych z treściami cyfrowymi. Na przykład takiej analizy wzięliśmy przykład firmy NETFLIX. Dane wzięliśmy ze strony [Kaggle](#).

Dane dla hurtowni usługi streamingowej linki:

### 1. Dane o subskrybentach

<https://www.kaggle.com/datasets/mauryansshivam/netflix-ott-revenue-and-subscribers-csv-file>

### 2. Filmy i seriale

<https://www.kaggle.com/datasets/durgeshrao9993/netflix-shows-dataset>

### 3. Filmy i seriale

<https://www.kaggle.com/datasets/shivamb/netflix-shows>

### 4. Najnowsze filmy i seriale na rok 2022

<https://www.kaggle.com/datasets/senapatirajesh/netflix-tv-shows-and-movies>

### 5. Cena subskrypcji dla krajów

<https://www.kaggle.com/datasets/prasertk/netflix-subscription-price-in-different-countries>

### 6. Oryginalne produkcje NETFLIX oraz ich ocena na IMDB

<https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores>

### 7. Cena akcji NETFLIX 2002-2022

<https://www.kaggle.com/datasets/meetnagadia/netflix-stock-price-data-set-20022022>

Zacznijmy od struktury hurtowni danych, która reprezentuje osobne tabele.

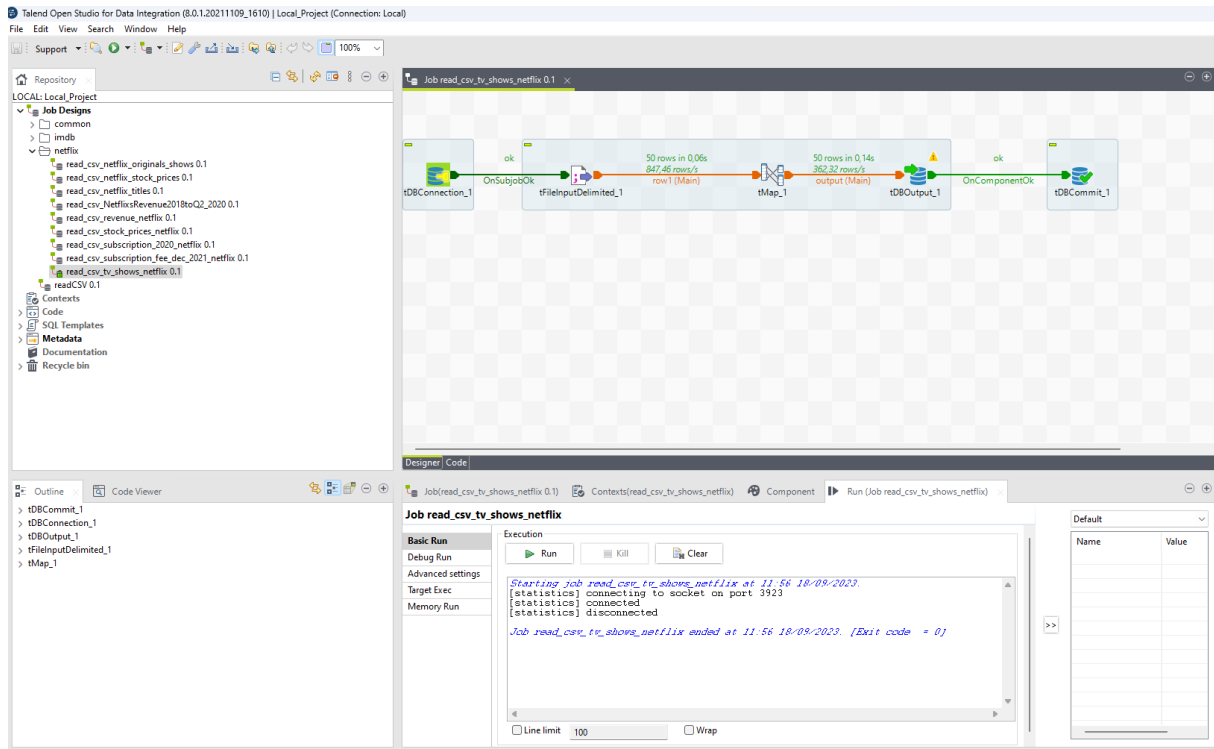
<b>public</b> <b>__EFMigrationsHistory</b> MigrationId character varying(150) ProductVersion character varying(32)	<b>public</b> <b>common_tv_shows</b> id bigint Title character varying Year timestamp without time zone Age character varying IMDB character varying Rotten_Tomatoes character varying Netflix boolean Hulu boolean Prime_Video boolean Disney boolean Type boolean	<b>public</b> <b>imdb_movies</b> id bigint Title character varying Date timestamp without time zone Score double precision Genre character varying Overview character varying Crew character varying Origin_Title character varying Status character varying Origin_Language character varying Budget double precision Revenue double precision Country character varying	<b>public</b> <b>netflix_films</b> show_id text title text director text casts text country text date_added timestamp with time zone release_year text rating text duration text listed_in text description text	<b>public</b> <b>netflix_originals_shows</b> id integer Title text Genre text Premiere timestamp with time zone Runtime integer IMDB_Score double precision Language text	<b>public</b> <b>netflix_titles</b> id integer type text title text director text casts text country text date_added timestamp with time zone release_year text rating text duration text listed_in text description text
<b>public</b> <b>netflix_revenue</b> id integer Date timestamp with time zone Global_Revenue bigint UACN_Revenue bigint EMEA_Revenue bigint LATM_Revenue bigint APAC_Revenue bigint UACN_Members bigint EMEA_Members bigint LATM_Members bigint APAC_Members bigint UACN_RPU double precision EMEA_RPU double precision LATM_RPU double precision APAC_RPU double precision Domestic_Members integer Domestic_Revenue integer International_Members integer International_Revenue integer Domestic_Free_Trialers integer International_Free_Trialers integer Netflix_Global_Users integer	<b>public</b> <b>netflix_revenue_2018_2020</b> id integer Area text Q1_2018 bigint Q2_2018 bigint Q3_2018 bigint Q4_2018 bigint Q1_2019 bigint Q2_2019 bigint Q3_2019 bigint Q4_2019 bigint Q1_2020 bigint Q2_2020 bigint	<b>public</b> <b>netflix_stock_prices</b> id integer Date timestamp with time zone Open double precision High double precision Low double precision Close double precision Adj_Close double precision Volume integer	<b>public</b> <b>netflix_subscription_2020</b> id integer Area text Years text Subscribers integer	<b>public</b> <b>netflix_subscription_fee_dec_2021</b> id integer Country_Code text Country text Total_Library_Size integer Number_of_TV_Shows integer Number_of_Movies integer Cost_Per_Month_Basic double precision Cost_Per_Month_Standard double precision Cost_Per_Month_Premium double precision	

- Tables (12)
- > \_\_EFMigrationsHistory
  - > common\_tv\_shows
  - > imdb\_movies
  - > netflix\_films
  - > netflix\_originals\_shows
  - > netflix\_revenue
  - > netflix\_revenue\_2018\_2020
  - > netflix\_stock\_prices
  - > netflix\_subscription\_2020
  - > netflix\_subscription\_fee\_dec\_2021
  - > netflix\_titles
  - > netflix\_tv\_shows

Rysunek 1. Zdjęcie przedstawia powstałe tabele

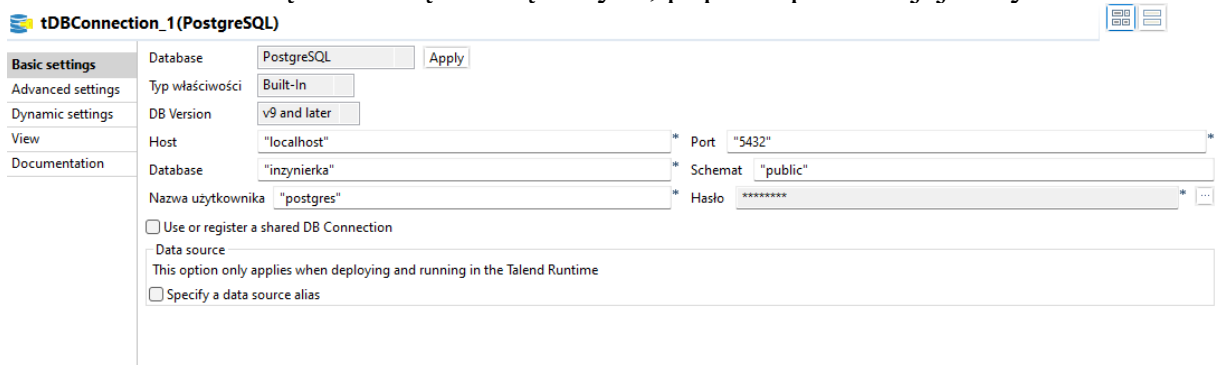
Działanie aplikacji:

W projekcie użyto oprogramowania Talend Open Studio for Data Integration 8.0.1 jako program ETL.



Możemy zauważyć że program posiada „jobs” są to zadania jakie program wykonuje w tym przypadku jest to ekstrakcja (extract) danych z plików typu excel a następnie transformowanie (transform) tych danych poprzez mapowanie zmiennych a następnie ładowanie (load) do bazy danych lub tworzenie nowych tabel z tymi danymi.


1. tDBConnection – łączenie się z bazą danych, poprzez podanie jej danych.












2. tFileInputDelimited – służy do odczytu plików, w tym przypadku pliki excel oraz dane jaki ma zostać użyty separator między danymi „;” oraz jakie są pola i jakie mają typy.

Schema of tFileInputDelimited\_1

tFileInputDelimited\_1

Column	Key	Type	<input checked="" type="checkbox"/>	N..	Date Patte...	Len...	Prec...	De...	Co...
 id	<input checked="" type="checkbox"/>	long	<input type="checkbox"/>						
Titles	<input type="checkbox"/>	String	<input type="checkbox"/>						
Year	<input type="checkbox"/>	Date	<input type="checkbox"/>		"yyyy"				
Rating	<input type="checkbox"/>	String	<input type="checkbox"/>						
IMDB_Rating	<input type="checkbox"/>	dou...	<input type="checkbox"/>						
Netflix	<input type="checkbox"/>	bool...	<input type="checkbox"/>						

OK
Cancel

tFileInputDelimited\_1

Basic settings
Advanced settings
Dynamic settings
View
Documentation

Property Type
Built-In

Schemat
Built-In
Edit schema

When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0.

File name/Stream
D:/Inzynierka/Data/netflix/TV Shows - Netflix.csv

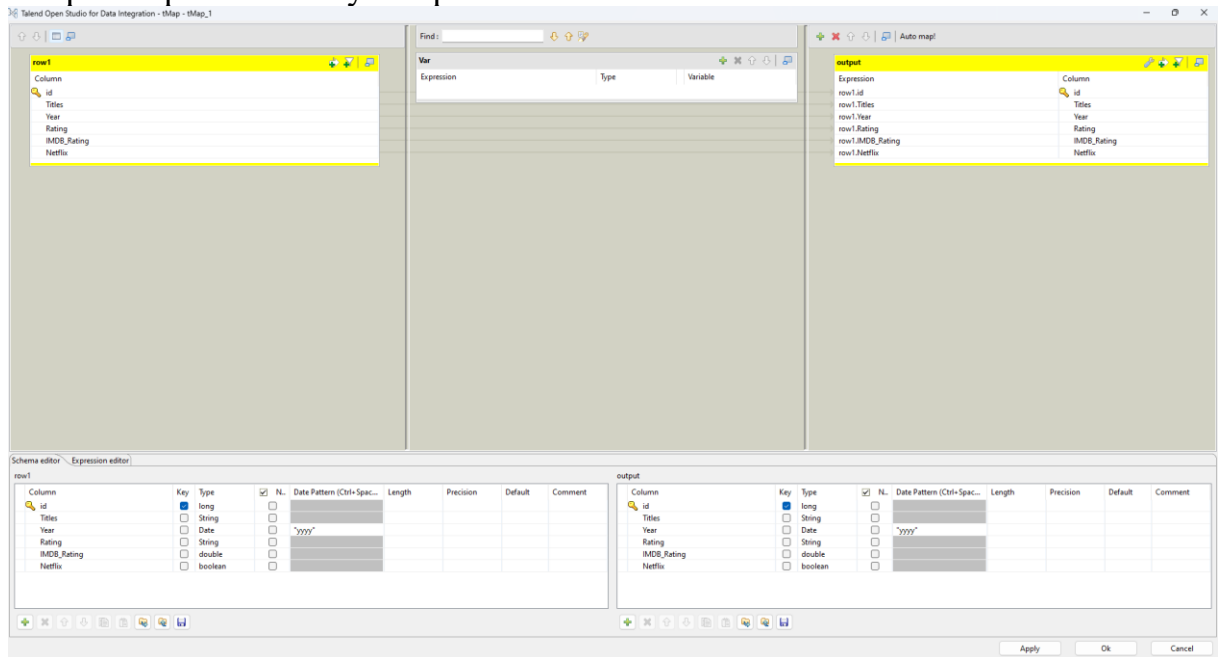
Row Separator
"\n"
Separator pól
";"

☐ Opcje CSV

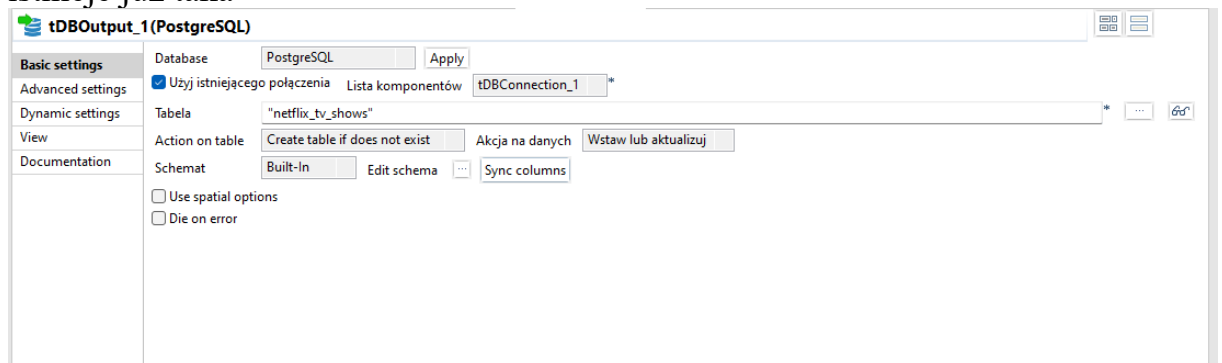
Header
1
Footer
0
Limit

☒ Skip empty rows
☐ Rozpakuj jako zip
☐ Die on error

### 3. tMap – mapowanie danych z plików na zmienne tabeli.

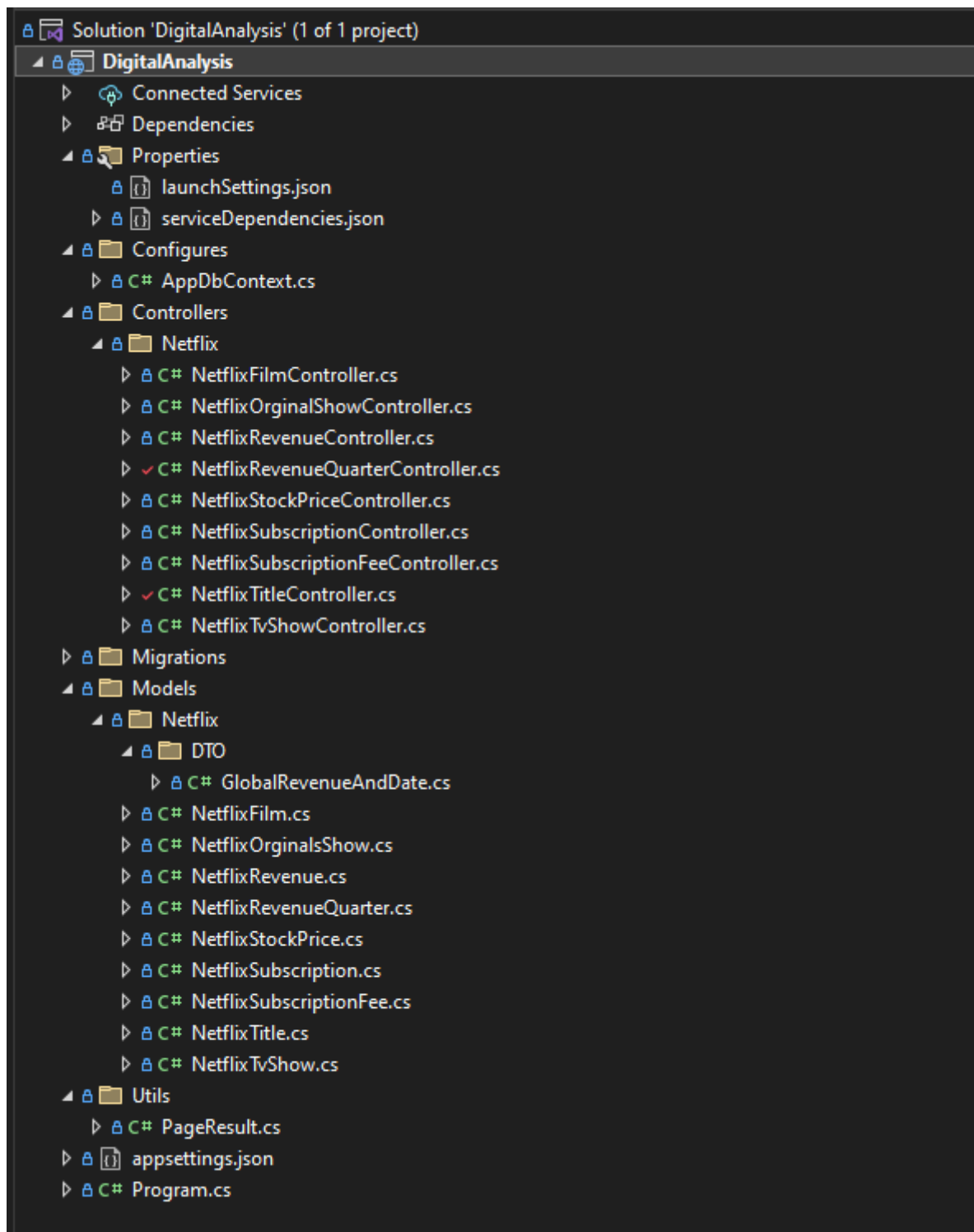


### 4. tDBOutput – akcja jak ma być wykonana na podłączonej bazie, w tym przypadku jest to tworzenie tabeli o nazwie „netflix\_tv\_shows” o ile nie istnieje już taka



### 5. tDBCommit – komitowanie zmian na bazę.

Struktura aplikacji back-end:



1. Controllers – w tym folderze znajdują się kontrolery, dzięki którym możliwa jest interakcja z bazą danych postgres poprzez REST API. Dla każdej tabeli powstał osobny kontroler z możliwością pobrania danych. Ze względu na to, że hurtownie danych są tylko do odczytu jedynie jest możliwość pobierania danych poprzez GET np.

```

[HttpGet]
[Route("/get-netflix-stock-price/{page}/{page_size}")]
0 references
public async Task<IActionResult> GetNetflixStockPrice([FromRoute(Name = "page")] int page, [FromRoute(Name = "page_size")] int pageSize)
{
    var totalCount = await _dbContext.NetflixStockPrices.CountAsync();

    var netflixStockPrices = await _dbContext.NetflixStockPrices
        .OrderBy(p => p.Id)
        .Skip((page - 1) * pageSize)
        .Take(pageSize)
        .ToListAsync();

    var pageResult = new PageResult<NetflixStockPrice>(netflixStockPrices, totalCount, pageSize, page);

    return Ok(pageResult);
}

```

2. Migrations – folder zawierający wszystkie migracje jakie zostały wprowadzone przez aplikację, w tym utworzenie wszystkich tabel.
3. Models – folder zawierający wszystkie modele danych jakie są używane w aplikacji

Swagger:

The screenshot displays the Swagger UI for an application named "DigitalAnalysis" (version 1.0, OAS3). The URL shown is <https://localhost:7115/swagger/v1/swagger.json>. The "Select a definition" dropdown is set to "DigitalAnalysis v1".

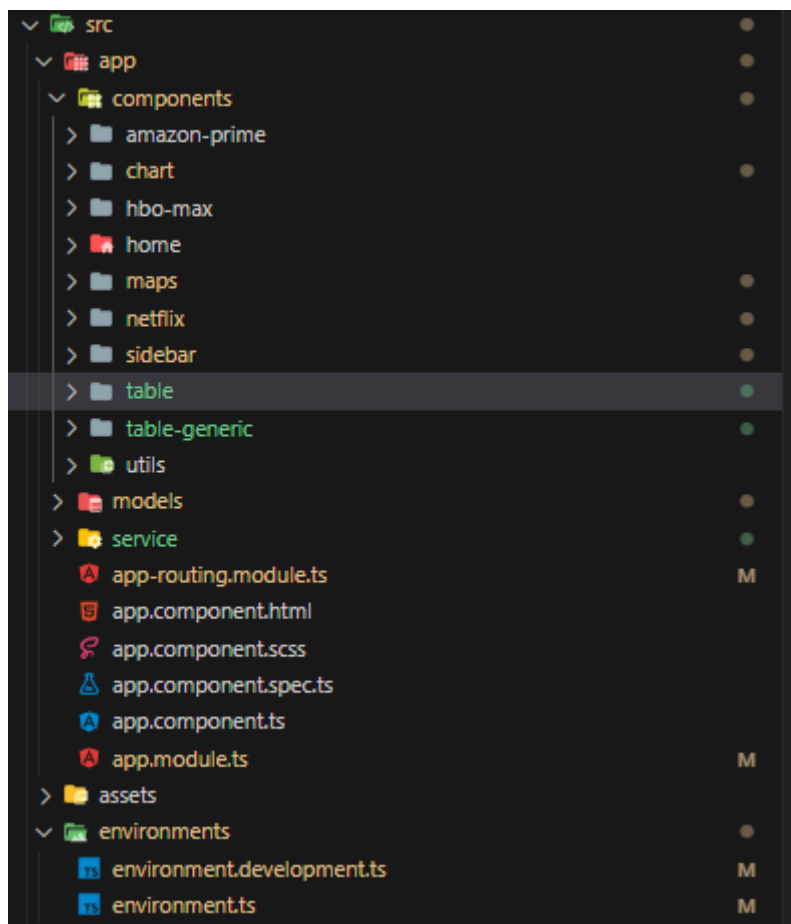
The API endpoints are organized into sections, each with a resource name and a list of endpoints:

- NetflixFilm**
  - GET /get-netflix-films/{page}/{page\_size}
- NetflixOriginalShow**
  - GET /get-netflix-original-show/{page}/{page\_size}
- NetflixRevenue**
  - GET /get-netflix-revenues/{page}/{page\_size}
  - GET /get-all-global-revenue-and-dates
- NetflixRevenueQuarter**
  - GET /get-netflix-revenue-quarter/{page}/{page\_size}
  - GET /get-netflix-revenue-quarter/all
- NetflixStockPrice**
  - GET /get-netflix-stock-price/{page}/{page\_size}
  - GET /get-netflix-stock-prices/for-year/{year}



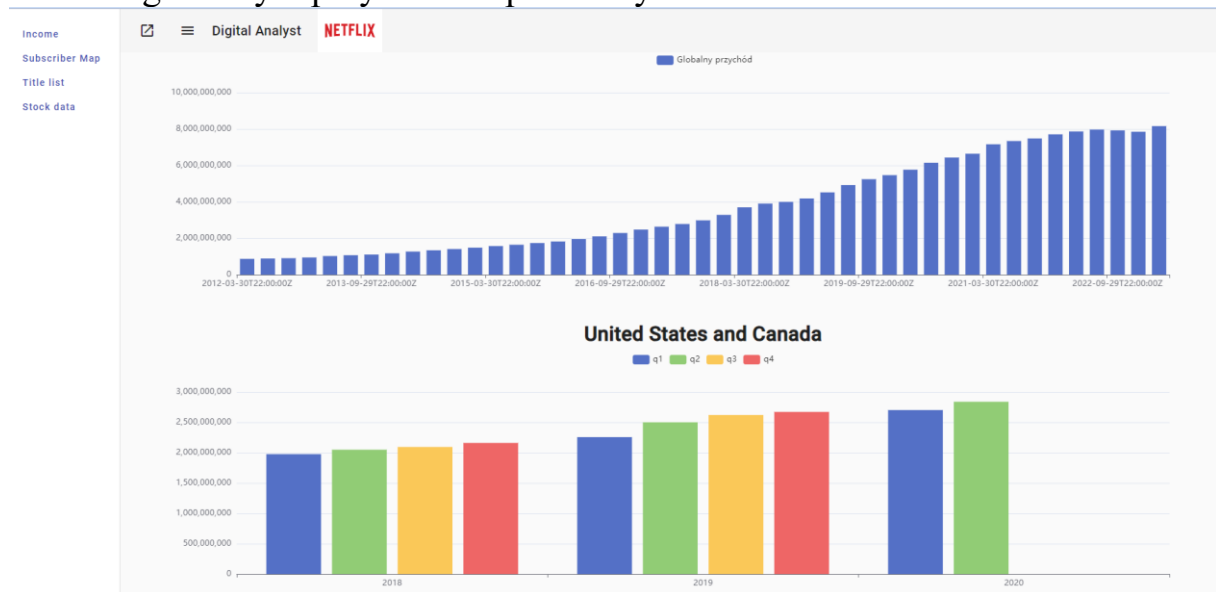
NetflixSubscription		^
GET	/get-netflix-subscription/{page}/{page_size}	▼
NetflixSubscriptionFee		^
GET	/get-netflix-subscription-fee/all	▼
GET	/get-netflix-subscription-fee/{page}/{page_size}	▼
GET	/get-netflix-subscription-fee/by-name/{country}	▼
NetflixTitle		^
POST	/api/NetflixTitle	▼
GET	/get-netflix-titles/{page}/{page_size}/{sortBy}/{sortDirection}	▼
NetflixTvShow		^
GET	/get-netflix-tv-shows/{page}/{page_size}	▼

Aplikacja front-end struktura:

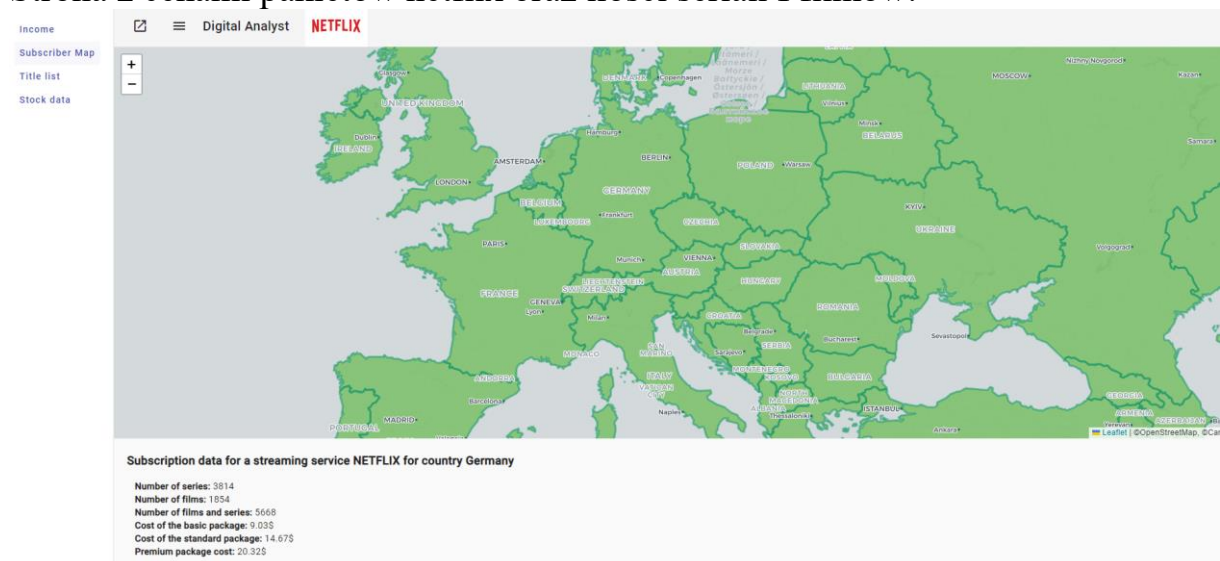


Wygląd aplikacji front-end:

Strona z globalnym przychodem platformy NETFLIX:



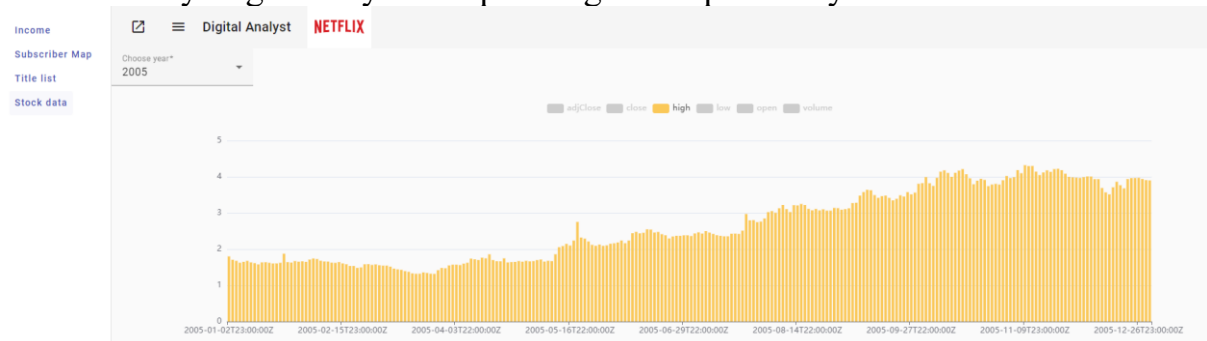
Strona z cenami pakietów netflix oraz ilości seriali i filmów:



## Strona z tytułami na platformie NETFLIX:

Income	🔍	☰	Digital Analyst	NETFLIX
Subscriber Map				
Title list				
Stock data				
Title	Type	Duration	Releaseyear	
Dick Johnson Is Dead	Movie	90 min	2020	▼
Blood & Water	TV Show	2 Seasons	2021	▼
Ganglands	TV Show	1 Season	2021	▼
Jailbirds New Orleans	TV Show	1 Season	2021	▼
Kota Factory	TV Show	2 Seasons	2021	▼
Midnight Mass	TV Show	1 Season	2021	▼
My Little Pony: A New Generation	Movie	91 min	2021	▼
Sankofa	Movie	125 min	1993	▼
The Great British Baking Show	TV Show	9 Seasons	2021	▼
The Starling	Movie	104 min	2021	▼
				Items per page: 10 1 - 10 of 8806 < >

## Strona z danymi giełdowymi dla podanego roku platformy NETFLIX:



## W ten sposób omówiliśmy całą architekturę aplikacji.

W ramach naszego systemu, na tę chwilę gromadzimy dane dotyczące danych statystycznych platformy NETFLIX z myślą o powiększenie o inne platformy jak HBO, Spotify, Disney oraz przy użyciu kostek wielowymiarowych przeprowadzać porównanie poszczególnych platform w konkretnych dziedzinach oraz tworzenie z tego raportów.

W dzisiejszym czasach wiedza o działaniu spółek oraz ich danych są nie do oceny dla działań na giełdzie.