Analiza trendów w okresie okołoświątecznym na Spotify

Agata Kopyt, Nikola Miszalska

December 2023

1 Cel projektu

Celem projektu jest przećwiczenie umiejętności składowania danych oraz ich analizy przy pomocy takich narzędzi jak: Apache NiFi, Apache Hadoop, Apache HBase, Apache Spark oraz powłoki Bash. Dzięki temu dokonamy analizy najpopularniejszych utworów w Polsce według platformy Spotify. Zbadamy tendencje wśród najchętniej odtwarzanych przez użytkowników platformy utworów, artystów oraz gatunków. Korzyściami wykonania projektu będą zdobyte umiejętności, a także zrozumienie trendów istniejących na platformie Spotify. Analiza ta może być wykorzystana np. przez potencjalnych reklamodawców, artystów czy producentów muzycznych.

2 Wykorzystywane źródła danych

2.1 Dane statyczne-30000 Spotify Songs

Zbiór danych 30 000 Spotify Songs jest w formacie csv i pochodzi ze strony Kaggle. Zbiór zawiera informację o 30 000 najczęściej odtwarzanych utworów na Spotify i pochodzi z listopada 2023r. Kolumny charakteryzujące zbiór to

- `track_id` - ID utworu
- `track_name` - Tytuł utworu
- `track_artist` - Wykonawca
- `track_popularity` - Popularność utworu (0-100), gdzie 100 oznacza najwyższą najpopularność
- `track_album_id` - ID płyty, na którym znajduje się utwór
- `track_album_name` - Tytuł płyty
- `track_album_release_date` - Data wydania płyty
- `playlist_name` - Nazwa playlisty, na której znaleziono utwór
- `playlist_id` - ID playlisty
- `playlist_genre` - Gatunek playlisty

- `playlist_subgenre` - Podgatunek playlisty
- `danceability` - Taneczność utworu (0-1), gdzie 1 oznacza przekonanie, że utwór jest taneczny
- `energy` - Energiczność utworu (0-1), gdzie 1 oznacza przekonanie, że utwór jest energiczny
- `key` - Tonacja utworu (0 - C, 1 C#, itp.)
- `loudness` - Głośność utworu w decybelach
- `mode` - Modalność utworu (1 tonacja durowa, 0 tonacja molowa)
- `speechiness` - "Mówność"oznacza obecność języka mówionego w utworze (0-1), gdzie wartości ok. 1 przypisywane są np. podcastom
- `acousticness` - Akustyczność (0-1), gdzie 1 oznacza przekonanie, że utwór jest akustyczny
- `instrumentalness` - Instrumentalność (0-1), gdzie 1 oznacza przekonanie, że utwór nie zawiera wokalu (nie licząc wokaliz)
- `liveness` - Żywiołowość (0-1), gdzie 1 oznacza przekonanie, że utwór jest żywiołowy
- `valence` - Nastrojowość (0-1), gdzie 1 oznacza przekonanie, że utwór ma nastrój pozytywny, a 0 nastrój negatywny
- `tempo` - Tempo utworu w BPM
- `duration_ms` - Długość utworu w ms

Zbiór został przez nas zmodyfikowany:

- Usunięte kolumny:
  - `track_popularity` - kolumna zawierająca nieaktualną informację o popularności utworu
  - `playlist_name` - kolumna zawierająca nazwę playlisty
  - `playlist_id` - kolumna zawierająca ID playlisty
  - `playlist_genre` - kolumna zawierająca gatunek playlisty
  - `playlist_subgenre` - kolumna zawierająca podgatunek playlisty
- Dodane kolumny:
  - `artist_id` - kolumna zawierająca ID wykonawcy uzyskane przy pomocy Spotipy
  - `artist_genres` - kolumna zawierająca listę gatunków przypisanych do wykonawcy przez Spotify uzyskane przy pomocy Spotipy; gdy nie przypisano żadnego gatunku, kolumnie przypisano listę zawierającą "unknown"

Ponadto uzupełniłyśmy zbiór o ok. 140 nowych utworów, głównie charakterystycznych dla Polskich słuchaczy. Usunęłyśmy również duplikaty piosenek - utwór o danym tytule i artyście mógł znaleźć się w zbiorze wiele razy, ponieważ artysta umieścił go na wielu albumach.

2.2 Dane dynamiczne

Danymi dynamicznymi w projekcie są dane systematycznie zbierane z API Spotify. Dane te dotyczą playlisty "Top 50 Polska". Aktualizacja danych następuje raz dziennie, ponieważ taka jest częstotliwość aktualizacji playlist "Top 50 ..."przez Spotify. Każdy pobrany zbiór zawiera 50 rekordów. Warto zaznaczyć, że utwory, które znalazły się na playliście danego dnia oznaczają najpopularniejsze utwory z dnia poprzedniego, co zostało uwzględnione przy analizie.

Dzienny zbiór danych posiadał następujące kolumny:

- `track_id` - analogiczna kolumna do ramki statycznej
- `rank` - miejsce w rakningu utworu danego dnia
- `track_name` - analogiczna kolumna do ramki statycznej
- `artist` - analogiczna kolumna do `track_artist` z ramki statycznej
- `date_added` - data dodania utworu do playlisty "Top 50 Polska", która jest równoznaczna z datą publikacji rankingu
- `popularity` - analogiczna kolumna do `track_popularity` z ramki statycznej

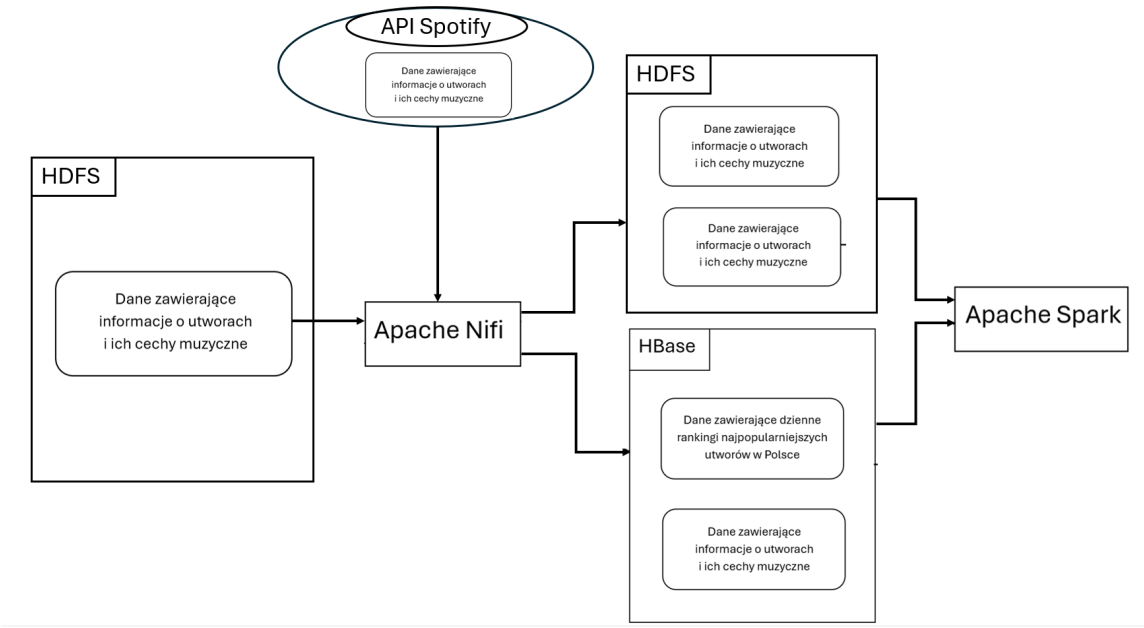
3 Diagram oraz opis architektury stworzonego systemu i wykorzystanych narzędzi

Architektura stworzonego przez nas systemu została przedstawiona na Rysunku 1.

1. **Warstwa Składowania Danych:**
   - **HDFS (Hadoop Distributed File System):**
     - Odpowiada za przechowywanie danych statycznych oraz dynamicznych w sposób rozproszony.
2. **Warstwa Przetwarzania Przepływów Danych:**
   - **Apache NiFi:**
     - Integruje się z API Spotify, pobiera dane dynamiczne.
     - Przeprowadza przetwarzanie ETL (Extract, Load).
     - Przekazuje zarówno dane statyczne, jak i dynamiczne do kolejnych warstw.
3. **Warstwa Składowania Danych NoSQL:**
   - **HBase:**
     - Baza danych NoSQL służąca do przechowywania dużych ilości danych.
     - Przechowuje zarówno dane statyczne, jak i dynamiczne po ich przetworzeniu przez NiFi.

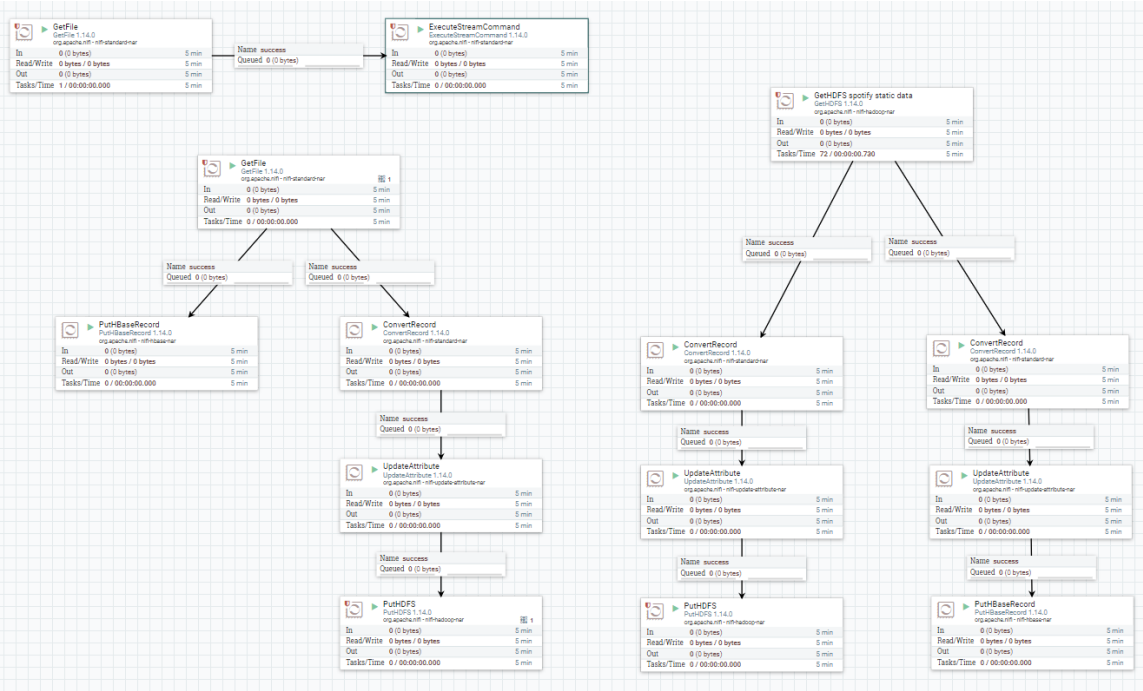
4. **Warstwa Analiz i Widoków Wsadowych:**

   - **Apache Spark:**
     - Silnik przetwarzania danych do analizy i generowania widoków wsadowych.
     - Integruje się z HBase, pobiera dane i przetwarza je w klastrze.



Rysunek 1: Diagram architektury systemu i wykorzystanych narzędzi

4 Opis sposobu pozyskiwania, przetwarzania i składowania danych źródłowych



Rysunek 2: Nifi Flow

4.1 Pobieranie i Przetwarzanie Danych Z Zewnętrznego API

W pierwszym etapie procesu, skrypt bashowy automatycznie pobierany jest z lokalnego folderu przy użyciu procesora GetFile. Kolejnym krokiem jest wykonanie go przez procesor ExecuteStreamCommand, co prowadzi do uruchomienia skryptu Pythonowego. Ten skrypt komunikuje się z zewnętrznym API, pobiera dane dotyczące dziennych trendów na spotify i zapisuje je w lokalnych katalogach 'archive' oraz 'stage', używając daty jako części nazwy pliku z rozszerzeniem JSON. System jest zaprojektowany tak, aby unikać ponownego pobierania danych, sprawdzając istnienie pliku z daną datą w katalogu 'archive'.

4.2 Przetwarzanie i Ładowanie danych do Hadoopa i HBase

W drugim etapie, procesor `GetFile` pobiera plik z katalogu 'stage'. Następnie, za pomocą procesora `PutHbaseRecord`, dane są ładowane do tabeli `'daily_data'` w HBase. Po tym jak procesor `GetFile` weźmie plik z katalogu 'stage', jest on z niego usuwany, co eliminuje ryzyko wielokrotnego ładowania tych samych danych. W rezultacie w lokalnym folderze 'stage' jest zawsze jeden plik zawierający świeże dane lub katalog ten jest pusty. Dalej dane są zapisywane w formacie Parquet w systemie plików HDFS przy użyciu procesora `PutHDFS`.

4.3 Przetwarzanie i Ładowanie Danych do Hadoopa i HBase

Dane statyczne znajdują się w pliku `spotify_songs.csv` w katalogu `/nifi_in/spotify/` w HDFS. Procesor `GetHDFS` umożliwia pobranie tego pliku, a następnie są one konwertowane. Skonwertowane dane są zapisywane w dwóch formatach: Parquet, które jest ładowane do HDFS za pomocą procesora `PutHDFS`, oraz JSON, które trafiają do tabeli 'spotify_songs' w HBase.

5 Opis sposobu analizy danych i rodzaju generowanych widoków wsadowych

Zdefiniowałyśmy okresy:

- Pora roku
- Wydarzenie sezonowe
- Miesiąc
- Dzień tygodnia

a następnie dla każdej podgrupy przeprowadzałyśmy 6 analiz:

- Porównanie wpływu "wieku"utworu – Czy nowości mają przewagę?
  - Wydane w ciągu miesiąca
  - Wydane w ciągu roku
- Podsumowanie cech muzycznych utworów z danego okresu
- Najpopularniejsze utwory z danego okresu
- Najpopularniejsi artyści z danego okresu
- Najpopularniejsze gatunki z danego okresu

Zanim przeprowadzenie analiz było możliwe, musiałyśmy dokonać połaczenia zbiorów, przetworzenia ich (np. nie branie pod uwagę utworów, których cech brakuje w zbiorze statycznym) oraz zdefiniowania dodatkowych kolumn takich jak: pora roku, wydarzenia sezonowe, miesiąc oraz dzień tygodnia w którym utwór pojawił się w rankingu. Ponadto dodałyśmy kolumny analityczne takie jak: główny gatunek, wiek utworu (na podstawie daty wydania albumu) w dniu rankingu, indykator czy gdy utwór pojawił się w rankingu był młodszy niż miesiąc, indykator czy gdy utwór pojawił się w rankingu był młodszy niż rok. Następnie zdefiniowałyśmy agregacje w celu utworzenia widoków wsadowych, które zapisywane są w formacie parquet w Apache Hadoop w odpowienich katalogach: `\analysis\nazwa grupy okresów\dzień utworzenia analiz\nazwa analizy`.
Podczas agregacji utworzyłyśmy takie kolumny jak: group_popularity, average_percent_of_top_50, days_on_chart, track_popularity, average_track_age, artist_albums, total_days (liczba dni w danym okresie), number_of_tracks, artist_popularity, artist_subgenres, genre_popularity, average_duration_min i analogicznie average_cecha muzyczna.

Popularność (grupy, artysty, utworu, gatunku) została obliczona utworzoną przez nas metryką, między innymi dlatego, że zebrane `popularity` odnoszą się do popularności utworu na całym świecie, podczas gdy nasze analizy dotyczą Polski.

Przykładowo popularość artysty została obliczona wzorem:

s - suma odwrotności pozycji w rankingu
l - liczba wystąpień w rankingach artysty w danym okresie (z uwzględnieniem, że dany artysta może mieć wiele utworów w rankingu)
t - liczba dni danego okresu
u - liczba utworów w jednym rankingu
m - maksymalna suma odwrotności pozycji w rankingu (artysta mógłby zająć je wszystkie)

$$\frac{s}{m} \cdot \frac{l}{u \cdot t^2}$$

W ramach przykładu, rysunki 3,4,5,6,7,8 przedstawiają widoki wsadowe dla okresu okołoświątecznego.

	seasonal_event	released_within_a_month	group_popularity	average_percent_of_top_50
0	Christmas	0	0.854012	90.769231
1	Christmas	1	0.005459	9.230769

Rysunek 3: Porównanie wpływu "wieku"utworów z okresu okołoświątecznego - wydane w ciągu miesiąca

	seasonal_event	released_within_a_year	group_popularity	average_percent_of_top_50
0	Christmas	1	0.230144	44.923077
1	Christmas	0	0.268606	55.076923

Rysunek 4: Porównanie wpływu "wieku"utworów z okresu okołoświątecznego - wydane w ciągu roku

	summary	danceability	energy	duration_min	loudness	mode	speechiness	k
0	count	650	650	650	650	650	650	6
1	mean	0.6117107692307686	0.628907692307692	3.068070512820507	-7.775503076923069	0.6353846153846154	0.0813015384615384	5.1892307692307
2	stddev	0.14046777248630957	0.1839401509717348	0.7249517290997012	2.996310369649548	0.4816928191575751	0.08178874925418064	3.56833458829587
3	min	0.206	0.106	1.146	-22.507	0.0	0.0253	(
4	max	0.885	0.96	4.66825	-2.018	1.0	0.351	1

Rysunek 5: Podsumowanie cech muzycznych utworów z okresu okołoświątecznego

	seasonal_event	track_name	artist	track_album_name	track_popularity	artist_main_genres	days_on_chart	average_track_age
0	Christmas	HIPNOZA	NIKOŚ	HIPNOZA	17.402182	[unknown]	13	161.0
1	Christmas	Last Christmas - Single Version	Wham!	The Singles: Echoes from the Edge of Heaven	16.735632	[pop, rock]	13	168.0
2	Christmas	All I Want for Christmas Is You	Mariah Carey	Merry Christmas	9.557143	[pop]	13	10643.0
3	Christmas	Snowman	Sia	Everyday Is Christmas (Deluxe Edition)	5.731183	[pop]	13	1877.0
4	Christmas	Rockin' Around The Christmas Tree	Brenda Lee	Merry Christmas From Brenda Lee	5.309890	[rock]	12	21612.5

Rysunek 6: Najpopularniejsze utwory z okresu okołoświątecznego

	seasonal_event	artist_popularity	artist	number_of_tracks	average_track_age	artist_main_genres	artist_albums	average_duration_min	average_danceabi
0	Christmas	0.002975	NIKOŚ	1	161.000	[unknown]	[HIPNOZA]	3.241200	0.745
1	Christmas	0.002861	Wham!	1	168.000	[pop, rock]	[The Singles: Echoes from the Edge of Heaven]	4.432667	0.733
2	Christmas	0.001634	Mariah Carey	1	10643.000	[pop]	[Merry Christmas]	4.018450	0.335
3	Christmas	0.001341	Blacha 2115	2	65.000	[hip hop, electronic]	[Kevin sam w domu, Jolie Jolie]	2.697061	0.706
4	Christmas	0.001239	Sia	2	1877.375	[pop]	[Everyday Is Christmas (Deluxe Edition)]	2.888117	0.684

Rysunek 7: Najpopularniejsi artyści z okresu okołoświątecznego

	seasonal_event	artist_main_genres	artist_subgenres	genre_popularity	number_of_tracks
0	Christmas	pop	[classic uk pop, polish pop, polish viral po...	0.231787	44
1	Christmas	other	[classic girl group, australian dance, urban...	0.143381	35
2	Christmas	rock	[heartland rock, classic canadian rock, roc...	0.043504	15
3	Christmas	hip hop	[hip hop, polish hip hop, polish hip hop]	0.042621	30
4	Christmas	unknown	[unknown]	0.024884	14

Rysunek 8: Najpopularniejsze gatunki z okresu okołoświątecznego

Na podstawie przeprowadzonych analiz można określić:

- Jak ważne jest wydawanie nowych utworów,
- Jakie cechy muzyczne i długość powinien mieć potencjalnie nowy popularny utwór,
- Z jakimi utworami konkurowałby nowy popularny utwór,
- Jakich utworów używać do promocji produktu,
- Z jakimi artystami konkurowałby nowy popularny utwór,
- Z jakimi artystami warto nawiązać współpracę,

- W jakim gatunku powinien być nowy utwór,
- Utwory z jakich gatunków warto odtwarzać np. w sklepach.

w danym okresie.

6 Podział pracy

Zadanie	**Członkowie zespołu**
Skrypt pobierający dane z API Spotify	Agata Kopyt
Przetwarzanie i integracja danych	Agata Kopyt, Nikola Miszalska
Automatyzacja przepływów danych Apache NiFi	Nikola Miszalska
Składowanie danych w Apache Hadoop	Nikola Miszalska
Składowanie danych w platformie NoSQL	Nikola Miszalska
Analiza danych i generowanie widoków wsadowych	Agata Kopyt
Testy funkcjonalne	Nikola Miszalska
Raport i prezentacja	Agata Kopyt, Nikola Miszalska