1. **Temat projektu**

Tematem projektu była implementacja aplikacji importującej dane z dowolnego statycznego źródła do systemu HDFS. Aplikacja ta miała za zadanie dokonać statystycznej analizy danych przy użyciu platformy Hadoop.

1. **Opis danych**

Wybrane przez nas dane opisują najpopularniejsze filmy na platformie YouTube. Dane zawarte są w plikach CSV i w plikach JSON. Do określenia filmu mianem popularności wykorzystano kombinację czynników takich jak: liczba wyświetleń, udostępnień, komentarzy i polubień. Popularność nie oznacza w tym przypadku najczęściej oglądanych filmów w roku kalendarzowym. Do najpopularniejszych na liście należą teledyski, występy gwiazd, reality show oraz losowe filmy. Zestaw danych zawiera dane z kilku miesięcy pewnego okresu na temat popularnych filmów z YouTube. Dane uwzględniają różne regiony, dla każdego regionu dane zapisane są w osobnym pliku. Dane zawierają takie informacje, jak: tytuł filmu, tytuł kanału, czas publikacji, tagi, wyświetlenia, polubienia i nielubienia, opis i liczbę komentarzy. Dane zawierają również pole category\_id, które różni się w zależności od regionu. Aby pobrać kategorię dla konkretnego filmu, należy odszukać go w powiązanym pliku JSON.

Ten zestaw danych został zebrany przy użyciu interfejsu API YouTube.

Dane zajmują 514 MB.

Struktura plików

|  |  |
| --- | --- |
| video\_id | Id filmu – string |
| trending\_date | Data np. 2017-11-10T17:00:03.000Z |
| title | Tytuł filmu – string |
| channel\_title |  |
| category\_id |  |
| publish\_time |  |
| tags |  |
| views |  |
| likes |  |
| dislikes |  |
| comment\_count |  |
| thumbnail\_link |  |
| comments\_disabled |  |
| ratings\_disabled |  |
| video\_error\_or\_removed |  |
| description |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Link do pobrania danych: <https://www.kaggle.com/datasnaek/youtube-new>

1. **Przeprowadzona statystyka**

W naszym projekcie przeprowadzona została następująca statystyka: dla każdej kategorii, zsumowano liczbę wszystkich interakcji użytkowników, takich jak: kliknięcia „lubię to”, „nie lubię”, czy liczba komentarzy dla danego filmu, należącego do danej kategorii filmu. Stąd można wnioskować, jakiej kategorii film jest najbardziej popularny na platformie YouTube.

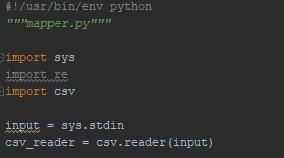
1. **Przygotowanie środowiska**

Naszą aplikację uruchomiliśmy na Dockerze.

W kontenerze zainstalowaliśmy obraz Hadoopa.

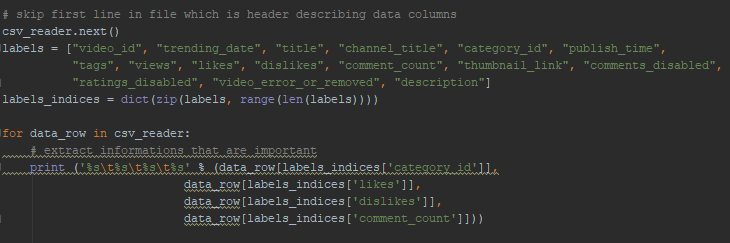
1. **Implementacja**
2. Mapper

Zadaniem mappera było wczytanie danych z plików i wydobycie z nich najbardziej interesujących informacji, czyli kategorii filmu, liczbę kliknięć „lubię to” i liczbę kliknięć „ nie podoba mi się” oraz liczbę komentarzy.



Rys. 1 Wczytanie plików

Do wypisania wyników dla każdej kategorii filmu, użyty został słownik, przechowujący klucz- wartość, gdzie klucz to cechy opisujące dany film a wartość to odpowiednia wartość (np. liczba polubień).

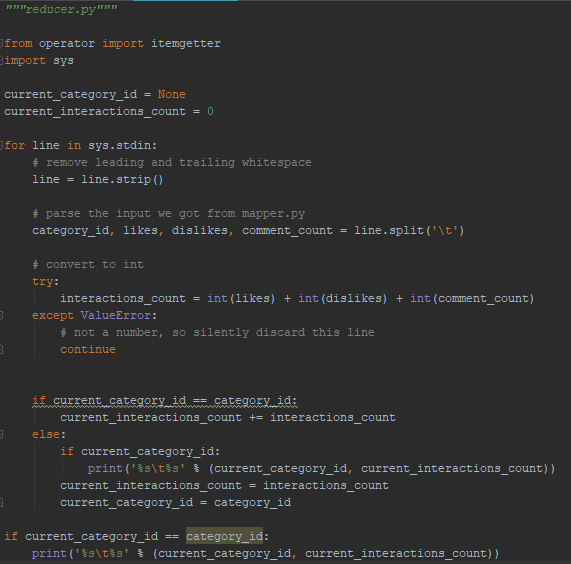


Rys. 2 Wynikiem mappera są rekordy przechowujące informacje o filmie:kategoria filmu, liczba kliknięć ‘Lubię to”, liczba kliknięć „nie lubię”, liczba komentarzy

1. Reducer

Reducer na wejściu przyjmuje wyniki otrzymane z mappera, przegląda wszystkie pojedyncze rekordy posortowane wg kategorii, w których znajdują się liczby: kliknięć ‘lubię to’, kliknięć ‘nie lubię’ oraz komentarz. Dla każdej kategorii sumuje liczbę wszystkich interakcji użytkowników.

Wynikiem Reducera jest nazwa kategorii i liczba wszystkich interakcji użytkownika dla danej kategorii filmu.



Rys. 3 Reducer zwracający liczbę wszystkich interakcji użytkowników dla danej kategorii filmów

1. Skrypt uruchamiający