MapReduce, Hadoop Streaming, Hive
Zestaw 9 – discogs

Pochodzenie danych to
`https://www.kaggle.com/datasets/ofurkancoban/discogs-releases-dataset`
Uwaga! Dane pobieramy z miejsca wskazanego w opisie projektu

Dwa zbiory danych

1. `datasource1` – informacje na temat płyt (1)

Dane mają format CSV, pliki nie posiadają nagłówków.

Pola w pliku:

- `id`
- `release_id`
- `status`
- `title`
- `artist_id`
- `artist_name`
- `label_id`
- `format`
- `genre`
- `style`
- `country`
- `release_date`

2. `datasource4` – informacje wytwórni muzycznych (4)

Dane mają format CSV, każdy z plików posiada nagłówek.

Pola w pliku:

- `label_id` - identyfikator wytwórni
- `label_name` – nazwa wytwórni

Program MapReduce (2)

Działając na zbiorze `datasource1` należy dla każdej wytwórni, każdego artysty, w każdej dekadzie, wyznaczyć liczbę wydanych płyt oraz lista różnych gatunków, których te płyty dotyczyły.

W wynikowym zbiorze (3) powinny znaleźć się atrybuty:

- identyfikator wytwórni
- identyfikator artysty
- nazwa artysty
- dekada
- liczba wydanych płyt
- lista różnych gatunków na wydanych płytach

Krzysztof Jankiewicz

Program Hive (5)

Działając na wyniku zadania MapReduce oraz zbiorze danych `datasource4` należy wyznaczyć dla każdej dekady te trzy wytwórnie, które wydały największą liczbę płyt.

Ostateczny wynik (6) powinien zawierać następujące atrybuty:

- `label_name` – nazwa wytwórni
- `decade` – dekada
- `artists_count` – liczba artystów wydanych płyt
- `releases_count` – liczba wydanych płyt
- `genres` – lista różnych gatunków na wydanych płytach