Modelowanie wielowymiarowe

29 października 2018

Opis pliku z zadaniami

Wszystkie zadania na zajęciach będą przekazywane w postaci plików `.pdf`, sformatowanych podobnie do tego dokumentu. Zadania będą różnego rodzaju. Za każdym razem będą one odpowiednio oznaczone:

- Zadania do wykonania na zajęciach oznaczone są symbolem $\triangle$ – nie są one punktowane, ale należy je wykonać w czasie zajęć.
- Punktowane zadania do wykonania na zajęciach oznaczone są symbolem $\diamond$ – należy je wykonać na zajęciach i zaprezentować prowadzącemu, w wypadku nie wykonania zadania w czasie zajęć lub nieobecności, zadania staje się zadaniem do wykonania w domu ($\star$).
- Zadania do wykonania w domu oznaczone są symbolem $\star$ – są one punktowane, należy je dostarczyć w sposób podany przez prowadzącego i w wyznaczonym terminie (zwykle przed kolejnymi zajęciami).
- Zadania programistyczne można wykonywać w dowolnym języku programowania, używając jedynie biblioteki standardowej dostępnej dla tego języka.

1 Serwis aukcyjny △

Treść

Firma internetowa prowadząca serwis aukcyjny postanowiła wdrożyć hurtownię danych w celu analizy zawieranych transakcji i zysków przez nie generowanych. Serwis posiada dużą bazę użytkowników. Każdy użytkownik może być zarazem sprzedawcą, jak i kupującym. Każdy użytkownik musi podać dokładne dane personalne, aby mógł korzystać z serwisu. Sprzedawca chcąc wystawić przedmiot na sprzedaż musi go najpierw opisać podając jego nazwę, kategorię, oraz dodatkowe cechy i otworzyć dla niego aukcję. Zamknięcie aukcji sukcesem odbywa się w momencie, gdy kupujący wyśle odpowiednie potwierdzenie dokonania zakupu. Aukcje, których czas upłyną bez dokonania zakupu, są traktowane jako zakończone bez sukcesu. Firma zarabia na prowizji od sprzedaży. Funkcja wyliczająca procent prowizji zależna jest od dotychczasowych wyników sprzedawcy.

Zadanie polega na zaprojektowaniu hurtowni danych, która będzie wspomagała decyzje kierownictwa w zakresie prowadzenia przedsiębiorstwa. Kierownictwo chce otrzymać przede wszystkim odpowiedzi na następujące pytania:

1. Ile aukcji zostało otwartych/zakończonych w podziale na jednostki czasu (np. w zestawieniu dziennym, tygodniowym lub miesięcznym)?
2. Ile aukcji nie zostało zakończonych sukcesem (tzn. zakupem)?
3. Jak wygląda ranking sprzedawców pod względem kwot sprzedaży?
4. W jakich miastach kupuje się sumarycznie najwięcej przedmiotów?
5. Jakie kategorie produktów są najchętniej kupowane rano, w południe, wieczorem?

W celu zaprojektowania hurtowni wykonaj następujące zadania:

1. Narysuj schemat hurtowni danych w postaci gwiazdy. Wskaż tabelę faktów oraz tabele wymiarów. W każdej tabeli uwzględnij wszystkie konieczne atrybuty. W tabeli faktów wskaż atrybuty, które są miarami. Krótko opisz i uzasadnij wybór ziarna tabeli faktów. Krótko opisz tabele wymiarów.
2. Czy w zaproponowanym przez Ciebie modelu występują naturalne hierarchie wymiarów? Pokaż dwie takie hierarchie.
3. Przerysuj schemat normalizując jeden z wymiarów. Nie musisz przepisywać nazw atrybutów pozostałych wymiarów.

2 Przetwarzanie dużych danych (aktualizacja) + transformacja danych do nowego schematu *10p.* ◇

Treść

Pobierz następujące dwa pliki pochodzące ze zbioru Million Song Dataset (MSD) ze strony przedmiotu:

- `unique_tracks.txt` — zawiera informacje takie jak identyfikator utworu, identyfikator wykonania, nazwę artysty oraz tytuł utworu,
- `triplets_sample_20p.txt` — zawiera identyfikator użytkownika, identyfikator utworu oraz datę odsłuchania.

Przeprowadź transoformacje danych z bieżącej postaci do nowego schematu w postaci gwiazdy. Po przeprowadzeniu transformacji wykonaj odpowiednie zapytania lub procedury, aby otrzymać odpowiedzi na następujące zapytania:

1. Ranking popularności piosenek,
2. Ranking użytkowników ze względu na największą liczbę odsłuchanych unikalnych piosenek,
3. Artysta z największą liczbą odsłuchań,
4. Sumaryczna liczba odsłuchań w podziale na poszczególne miesiące,
5. Wszyscy użytkownicy, którzy odsłuchali wszystkie trzy najbardziej popularne piosenki zespołu Queen.

Przygotuj `Dockerfile` w którym przy pomocy dowolnej technologi programistycznej/bazodanowej otrzymasz odpowiedzi na powyższe zapytania.

W celu przygotowania `Dockerfile` zapoznaj się z:

1. https://www.docker.com/
2. https://docs.docker.com/get-started/
3. https://docs.docker.com/get-started/part2/

Oraz przykładem dostępnym na stronie przedmiotu.

Rozwiązanie powinno działać zgodnie z poniższą specyfikacją techniczną:

1. Załóż, że pliki `unique_tracks.txt` oraz `triplets_sample_20p.txt` znajdują się w katalogu razem z Twoim `Dockerfile` i możesz skopiować je do wnętrza swojego obrazu.

2. Rozwiązania dla wszystkich podpunktów wypisz w standardowe wyjście jedno po drugim.

3. Dla podpunktu 1. wypisz na standardowe wyjście 10 najpopularniejszych piosenek posortowanych w kolejności od najpopularniejszej, każdą piosenkę w jednej osobnej linii w formacie:

```
<tytuł piosenki 1> <nazwa wykonawcy piosenki 1> <ilość odsłuchań piosenki 1>
<tytuł piosenki 2> <nazwa wykonawcy piosenki 2> <ilość odsłuchań piosenki 2>
...
<tytuł piosenki 10> <nazwa wykonawcy piosenki 10>  <ilość odsłuchań piosenki 10>
```

4. Dla podpunktu 2. wypisz na standardowe wyjście 10 użytkowników z największą liczbą odsłuchanych unikalnych piosenek posortowanych w kolejności malejącej po liczbie odsłuchanych unikalnych piosenek, każdego użytkownika w jednej osobnej linii w formacie:

```
<id użytkownika 1> <ilość odsłuchanych unikatowych piosenek przez użytkownika 1>
<id użytkownika 2> <ilość odsłuchanych unikatowych piosenek przez użytkownika 2>
...
<id użytkownika 10> <ilość odsłuchanych unikatowych piosenek przez użytkownika 10>
```

5. Dla podpunktu 3. wypisz na standardowe wyjście nazwę najpopularniejszego wykonawcy w formacie:

```
<nazwa wykonawcy> <sumaryczna ilość odsłuchań jego piosenek>
```

6. Dla podpunktu 4. wypisz na standardowe wyjście miesiące, każdy w jednej osobnej linii w formacie:

```
1 <sumaryczna ilość odsłuchań w miesiącu 1>
2 <sumaryczna ilość odsłuchań w miesiącu 2>
3 <sumaryczna ilość odsłuchań w miesiącu 3>
...
12 <sumaryczna ilość odsłuchań w miesiącu 12>
```

   lub

```
01 <sumaryczna ilość odsłuchań w miesiącu 1>
02 <sumaryczna ilość odsłuchań w miesiącu 2>
03 <sumaryczna ilość odsłuchań w miesiącu 3>
...
12 <sumaryczna ilość odsłuchań w miesiącu 12>
```

7. Dla podpunktu 5. wypisz na standardowe wyjście 10 pierwszych id użytkowników spełniających warunek zgodnie z porządkiem alfabetycznym (posortowanych rosnąco po id), każdego użytkownika w jednej osobnej linii w formacie:

```
<id użytkownika 1>
<id użytkownika 2>
...
<id użytkownika 10>
```

8. Obraz powinien wykonać zadanie po uruchomieniu bez potrzeby podawania żadnych dodatkowych parametrów (zawierać polecenie CMD/ENTRYPOINT).

9. Zadbaj o efektywny czas przetwarzania – całość nie powinna zająć więcej niż 20 min.

10. **Do projektu dołącz plik README w którym opiszesz wybraną techonologię i uzasadnienie swojego wyboru oraz schemat danych po transformacji.**

11. **Swoje rozwiązanie wyślij w archiwum zip, zawierającym folder o nazwie:** `<numer indeksu>_<imię>_<nazwisko>` **na adres mwydmuch@cs.put.poznan.pl (tytule maila umieść najpierw prefiks [PMD]) Bezpośrednio wewnątrz tego folderu powinien znajdować się plik** `Dockerfile`.

12. Twoje rozwiązanie zostanie sprawdzone sprawdzarką, dlatego postaraj się by Twoje rozwiązanie było zgodne ze specyfikacją!

13. Jak być może zdążyłaś/zdążyłeś zauważyć, w niektórych wypadkach możliwe jest więcej niż jedno rozwiązanie, sprawdzarka powinna akceptować wszystkie poprawne warianty.