Modelowanie wielowymiarowe i transformacja danych z linii komend

19 grudnia 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.
- 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).

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 Transformacja danych do nowego schematu △

Treść

Zaproponuj dla problemu Million Song Dataset (MSD) poprawiony schemat. Zastanów się, jakie elementy można poprawić w oryginalnym schemacie, tak aby zapewnić jego elastyczność (np. nowe informacje na temat użytkowników), dokładniejszą obsługę informacji o dacie, mniejszą złożoność pamięciową oraz lepszą wydajność czasową przy rozrastających się informacjach o poszczególnych obiektach w tym problemie.

3 Transformacja danych z linii komend *10p.*◇

Treść

Zrealizuj zadanie transformacji danych do przygotowanego schematu przy użyciu komend unixowej powłoki systemowej (np., bash). Wykorzystaj takie narzędzia jak `head`, `cat`, `paste`, `sort`, `grep`, `sed`, `wc`, `cut`, `comm`, `echo`, `date`, `join`, `awk` (wspierający również tabele asocjacyjne i przetwarzanie wielu plików), a także instrukcje powłoki (np., `for`, `while`).

Do pomiaru czasu trwania poszczególnych operacji można wykorzystać program `time`.

Użytkownicy systemu `MacOS` powinni zwrócić uwagę na różnice w implementacji różnych narzędzi. W celu ich zniwelowanie polecamy zainstalować wersję `GNU` narzędzi.

Punktacja:

- Przekształcenie plików wejściowych do schematu gwiazdy: 5 pkt.
- Wykonanie każdego zapytania: 1 pkt.