Katedra
Informatyki i Automatyki
Politechniki Rzeszowskiej

Bazy Danych

2. DDL i ETL

Opracował: Maciej Penar

Spis treści

1. (Mniej niż) garść informacji

SQL jako język dzieli się na kilka obszarów:

- DQL – Data Query Language – czyli SELECT-y
- DML – Data Manipulation Language – czyli INSERT / UPDATE / DELETE
- DDL – Data Definition Language – czyli definiowanie metadanych

O ile DML i DQL są częściami języków które dość dobrze dają się przenosić pomiędzy BD różnych dostawców, to niestety nie można tego powiedzieć odnośnie naszego bohatera czyli DDL.

Bardzo często przy budowie tzw. Hurtowni Danych (ang. Data Warehouses) należy przerzucić dane z systemów źródłowych (lub plików) do docelowej BD. Programowanie takiego przerzutu nosi nazwę procesu Ekstrakcji-Transformacji-Ładowania (ETL, Extract-Transform-Load).

2. Zadanie

Należy:

1. pobrać zbiór danych MovieLens (link) – zawierający 250MB
2. zapoznać się z opisem danych: link
3. zainstalować wybraną Bazę Danych (info poniżej)
4. załadować zbiór danych do BD
5. posadzić pomocniczy widok (info poniżej)

Sprawozdanie które oczekuję do godziny 8:00 dnia 2020-04-20 ma formę freestylu – zadanie daje się wykonać na kilka sposobów np.:

- można napisać program w dowolnym języku który zaczytuje pliki i wykonuje odpowiednie DML-e
- może wybrana BD posiada jakieś narzędzie pomocnicze do ładowania tabel
- …. Pliki można zgrepować na INSERT INTO … to głupi pomysł
- Można skorzystać z któregoś z narzędzi ETL: Talend, Pentaho

Niezależnie od wybranej metody należy ją szczegółowo opisać:

- Opisać zbiór danych
- Narysować docelowy Diagram ERD w 3PN
- **Napisać DDL-ki które:**
  - **obejmują więzy integralności (tj. klucze obce z racjonalnie dobranymi akcjami ON DELETE/ON UPDATE)**
  - Obejmują ograniczenia typu DEFAULT
- jeśli powstał kod to dołączyć źródła (nie dołączać binariów)

Ocenie będzie podlegać:

1. używalność zaproponowanego rozwiązania ETL. Rozwiązanie powinno być:
   a. albo szybkie (warto poczytać o np. ładowaniu hurtowym (ang. Bulk load) link link)
   b. albo przystępne – np. procesy programowane graficznie
2. sam fakt zainstalowania BD

INFO

Musicie uważać, bo plik z filmami jest denormalizowany wg. gatunków.

Dodatkowo chciałbym żeby w ramach DDL-ki znalazł się widok eksponujący trzy kolumny

- Tytuł filmu
- Id użytkownika
- Ocenę

INFO DOT. BAZ DANYCH

Polecam SQL Server Developer Edition – czyli edycja zawierająca wszystkie feature'y do celów demonstracyjnych. Ogólnie wszystkie bazy danych posiadają edycje Developerskie za które nie trzeba płacić. Warto też zaznaczyć, że BD na ogół pracują jako procesy systemowe – tj. oddzielamy klienta GUI od samej BD. Jeśli potrzebujecie GUI to poniżej zamieszczam jak się nazywają w ramach różnych BD:

Baza Danych	**Klient**
SQL Server	SQL Server Management Studio (SSMS)
Oracle Database	SQL Developer
IBM DB2	Data Studio
Postgres	PgAdmin
MariaDB / MySQL	SQLWorkbench