Katedra
Informatyki i Automatyki
Politechniki Rzeszowskiej

Bazy Danych

2. DDL i ETL

Opracował: Maciej Penar

Spis treści

1. (Mniej niż) garść informacji 3

2. Zadanie 3

Info 4

Info dot. baz danych 4

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

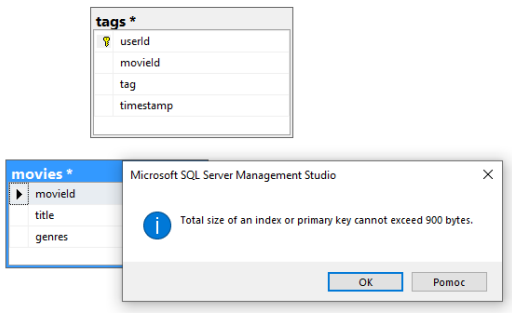
Polecam SQL Server Developer Edition – czyli edycja zawierająca wszystkie feature'y do celów demonstracyjnych. Ogólnie wszystkie bazy danych posiadają edycje Developerskie za które nie trzeba płacić. Warto też zaznaczyć, że BD na ogół pracują jako procesy systemowe – tj. oddzielamy klienta GUI od samej 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

3. FAQ

OGRANICZENIA I TYPY DANYCH

Dobry wieczór, mam problem z przypisaniem klucza głównego do encji movies w SQL Server. Próbowałem znaleźć rozwiązanie w Internecie, ale żadne ze znalezionych nie pomogło. Czy mógłby Pan wskazać jak rozwiązać ten problem?



W międzyczasie odpowiadając na pytanie bardziej szczegółowo -> widzę że próbujesz dodać PK na diagramach. Jest ryzyko że SQL który się generuje wybiera jakiś niepokojący zestaw kolumn jako PK np. (movieId,title,genres) - i dlatego SQL Server narzeka: ograniczenia indeksatora to 900 bajtów (u Ciebie varchary to 90 +1 + 600 + 1 + 600 + 1 > 900). To rodzi dwa sposoby radzenia sobie z problemem

1) dopasować typy danych racjonalnie: 600 znaków na tytuł to bardzo (na gatunek też) dużo + wydaje mi się, że

movieId wcale nie jest typu varchar

2) druga solucja polega na ulepszeniu definicji tabeli w CREATE TABLE - ograniczenia kluczy głównych i obcych można definiować na poziomie tej komendy

np.

CREATE TABLE movies(movieId VARCHAR(90) PRIMARY KEY, title VARCHAR(600), genres VARCHAR(600));

albo

CREATE TABLE movies(movieId VARCHAR(90) PRIMARY KEY NONCLUSTERED, title VARCHAR(600), genres VARCHAR(600));

albo to samo co pierwsze:

CREATE TABLE movies(movieId VARCHAR(90) PRIMARY KEY CLUSTERED, title VARCHAR(600), genres VARCHAR(600));

TAGI

Czy w drugim zadaniu usertagi też powinno się znormalizować? to znaczy, unikalne tagi przenieść do osobnej tabeli, a w tags zostawić zamiast nazwy tagu jego identyfikator?

Si, to jest najtrudniejszy problem do rozwiązania

TRANSFEROWALNOŚĆ I PRZYNALEŻNOŚĆ DO KLUCZA

Czy na diagramie erd trzeba uwzględniać transferowalność i przynależność do klucza?

Dosłownie pierwszy raz słysze te terminy

Chodzi mi o "transferable" i "identifying relationship"

Transferable - dosłownie mam puste echo w głowie, a identifing relationship - niby tak, ale raczej nie będzie istnieć taka potrzeba bo to tzw. encja słaba. Tj. encja słaba musi mieć słaby związek Tak naprawdę to jest na odwrót: byt który jest połączony słabym związkiem staje się encją słabą.

Odpowiem pod kątem 90 stopni - na zupełnie inne pytanie: jak masz możliwość wyboru pomiędzy ERD, a UML to preferowałbym UML (diagram klas) jako reprezentację schematu BD.

PRZECINKI W TYTUŁACH

Mam jeszcze pytanie dot. tej tabeli. Cały tytuł jest zamknięty w cudzysłowie, a w środku niektórych tytułów są przecinki. FIELDTERMINATOR wykrywa mi je i przenosi dalszą część tytułu do kolejnej kolumny. Czy są jakieś sposoby, aby w przypadku otoczenia tekstu "" nie brało pod uwagę przecinków w środku między znakami?

Nie wiem czy w narządku SQL Servera jest taka opcja w narzędziach ETL jest - bo to czesty case. Zależy co chcesz osiągnąć: jak trzeba 'one-shotować' ładowanie bazy danych, to ten jeden rekord poprawiasz w notatniku i fix.

Jeśli ładowanie jest powtarzającym sie procesem, to trzeba albo grepować, albo haczyć np. łądować paczke do pośredniej tabeli i z pośredniej tabeli przenosić dane - a tu już masz moc SQL/CLR.

Na potrzeby tego zadania założyłbym, że to one-shot tj. jak wejdziesz w notatnik i usuniesz sobie przecinek to nikomu korona z głowy nie spadnie

Rozumiem, w ponad 200000 rekordach może ich być sporo.

To zapuść grepa w notepad++ pewnie jakieś: .+?,.+?,.+?,.+