Analiza YELP podatkovne zbirke s tehnologijo Microsoft Analysis Services

Seminarska naloga pri predmetu TEHNOLOGIJE UPRAVLJANJA PODATKOV

Primož Hrovat  
Jakob Gaberc Artenjak

Mentor: Luka Šajn

december 2017, januar 2018

Povzetek

Kazalo vsebine

[1 Uvod 1](#_Toc502667579)

[2 Priprava podatkovne baze 1](#_Toc502667580)

[2.1 Podatki 1](#_Toc502667581)

[2.2 Namestitev podatkovne baze 2](#_Toc502667582)

[2.3 Priprava podatkov in migracija 2](#_Toc502667583)

[2.4 Kreiranje projekta 3](#_Toc502667584)

[3 Analiza podatkov 3](#_Toc502667585)

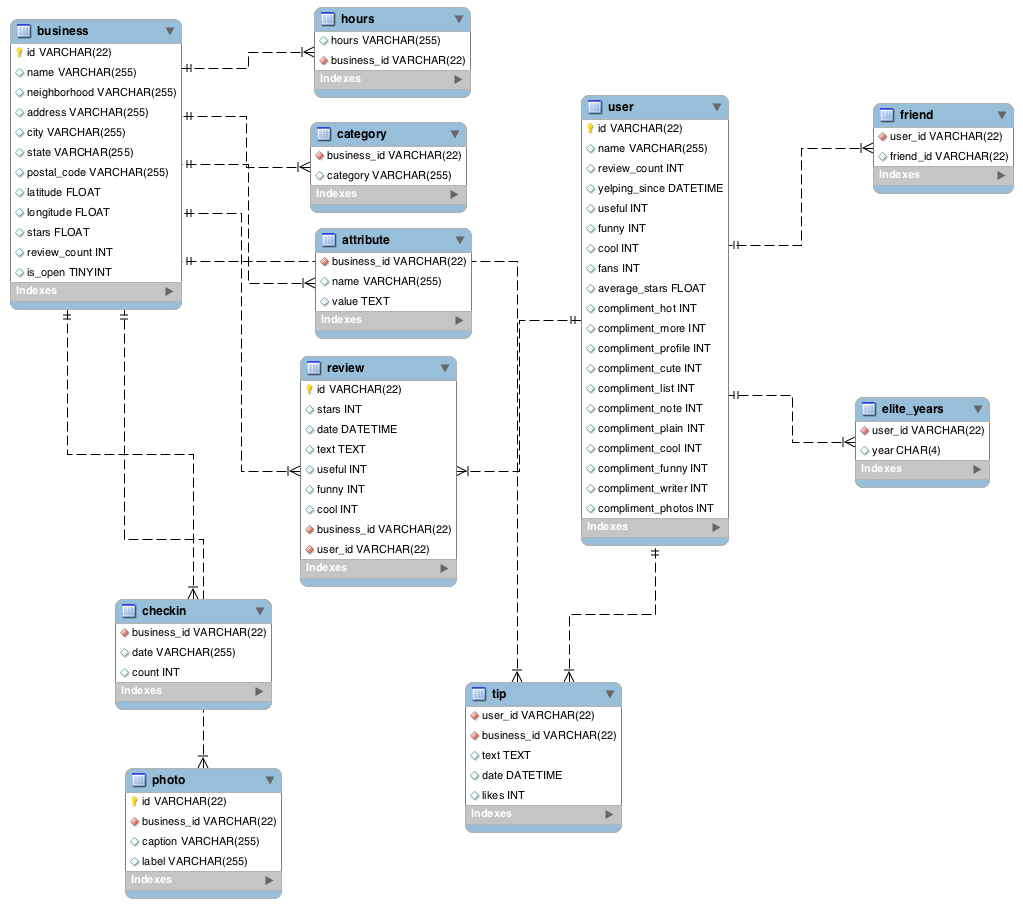
# Uvod

Pri predmetu Tehnologije upravljanja podatkov sva v okviru seminarske naloge želela analizirati podatke s strani Yelp z uporabo oblačnih storitev, ki jih ponuja Microsoftov oblak Azure. Vendar sva kmalu naletela na nepremostljiv problem, ker je v okviru študentske licence, ki si jo lahko študent FRIja pridobi preko programa IMAGINE, v SQL podatkovno bazo mogoče brezplačno naložiti le 32MB podatkov. V najinem primeru, ko je teh podatkov slabih 7GB, to pač ni mogoče. Zato sva bila prisiljena poiskati alternativno rešitev in jo našla v uporabi brezplačne instance na grškem akademskem oblaku Okenaos, ki deluje v sklopu omrežja GEANT.

# Priprava podatkovne baze

## Podatki

Podatki s katerimi sva se v tej nalogi ukvarjala so s spletnega mesta YELP, ki objavlja uporabniške ocene lokalnih podjetij in restavracij. Omogoča možnost komentiranja in ocenjevanja. Shema podatkovne baze je na Sliki 1. Podatki so dostopni preko spleta, skupna velikost le-teh pa skupaj nanese slabih 7GB podatkov. Shema in podatki so v formatu, ki ga podpira podatkovna baza MySQL. Podatki zbrani v bazi so ocene uporabnikov o lokalnih podjetjih z opisi, delovnimi urami, kategorizacijo, lokacijo… Možno je tudi dodajanje slik za posamezno podjetje.



Slika 1: shema podatkovne baze Yelp

## Namestitev podatkovne baze

Zaradi lažje organizacije in velike količine podatkov, sva podatkovno bazo gostovala v oblaku, na grškem omrežju Okeanos, ki znotraj akademskega kroga GEANT, ponuja brezplačno gostovanje v oblaku (na voljo nam je 80GB diskovnega spomina, 4GB delovnega spomina, do 4 procesorska jedra). Na virtualni stroj sva namestila Windows Server 2012, ki je prav tako na voljo preko grškega oblaka. Na njem teče instanca Microsoft SQL Serverja (MS SQL) z nameščenimi Anaylsis Services, Integration Services in Reporting Services. Strežnik uporablja razvijalsko licenco, zato so performanse nekoliko omejene. Zaradi podatkov v obliki primerni za MySQL je bilo potrebno namestiti tudi MySQL strežnik.

## Priprava podatkov in migracija

Podatki, dostopni preko spleta, so v dialektu SQL, ki ga podpira MySQL, uvoz neposredno v MS SQL pa ni neposredno podprt. Za premostitev težave sva na strežnik namestila tudi MySQL, podatke najprej uvozila v ta sistem za upravljanje podatkovne baze, potem pa s pomočjo orodja, ki ga ponuja Microsoft, namenjenemu migraciji podatkov iz različnih sistemov v MS SQL, podatke (dokaj) uspešno prenesla v ciljni sistem. Pri tem postopku migracija ni bila opravljena v celoti, v tabeli spodaj je predstavljena uspešnost migracije posamezne relacije iz MySQL v MS SQL.

Tabela 1: migracija podatkov iz MySQL v MS SQL

|  |  |  |
| --- | --- | --- |
| Originalna tabela | Ciljna tabela | Uspešnost % |
| yelp\_db.attribute | yelp\_db.attribute | 72,73 |
| yelp\_db.bussiness | yelp\_db.bussiness | 100 |
| yelp\_db.category | yelp\_db.category | 100 |
| yelp\_db.checkin | yelp\_db.checkin | 29,42 |
| yelp\_db.elite\_years | yelp\_db.elite\_years | 100 |
| yelp\_db.friend | yelp\_db.friend | NA |
| yelp\_db.hours | yelp\_db.hours | 100 |
| yelp\_db.photo | yelp\_db.photo | 100 |
| yelp\_db.review | yelp\_db.review | 5,21 |
| yelp\_db.tip | yelp\_db.tip | 42,52 |
| yelp\_db.user | yelp\_db.user | 30,95 |

Zaradi dokaj nizkega uspeha prenosov je drugi poskus prenosa podatkov v MS SQL strežnik potekal »ročno«, z uporabo JSON datotek, ki so prav tako na voljo na spletu. Tu je bila uspešnost prenosa 100%, le potrebno je bilo vložiti nekaj več truda, saj so podatki predstavljeni v deloma drugačni obliki, oziroma so nekatere tabele med sabo združene.

## Kreiranje projekta

Za povezavo in ustvarjanje kocke se v okolju Windows uporablja programski paket SSDT (SQL Server Data Tools), ki skupaj z razvojnim okoljem Visual Studio (VS) omogoča grafično načrtovanje kocke, dimenzij ter pogledov. Potrebno je vzpostaviti povezavo z bazo, nato pa kreirati kocko, dimenzije. V tem primeru sva ustvarila novo kocko, ki se jo potem z orodjem VS tudi ustvari na strežniku. Do nje in do pripadajočih dimenzij imamo potem tudi dostop preko Reporting Services, Excela…

# Analiza podatkov

## Aktivnost uporabnikov

## Ocene poslovnih subjektov

V tem sklopu me je zanimalo, kakšne so povprečne ocene subjektov glede na kategorije poslovanja, odvisnosti od lokacije (okrožja), mesta