Pokročilé databázové technológie

Zadanie 6 – MongoDB

Marek Adamovič

Cvičenie: Piatok 11:00 – 12:40

Ing. Ján Balažia, PhD. 2022/2023

Obsah

[1. Dátový model 3](#_Toc122631987)

[2. Import 5](#_Toc122631988)

[3. Dopyty 8](#_Toc122631989)

# 1. Dátový model

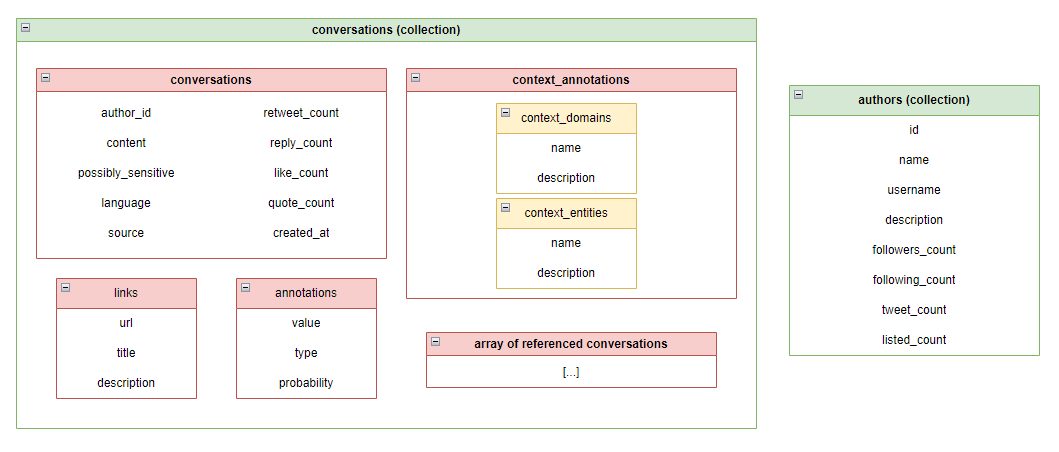
**Otázka:**

Navrhnite dátový model (kolekcie a formát dokumentov) v MongoDB pre dataset tweetov, ktorý bude využívaný mobilnou aplikáciou, ktorá bude:

a. Zobrazovať tweety jednotlivých používateľov vo forme feedov

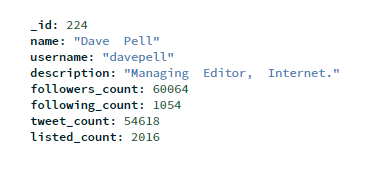
b. Zobrazovať jednotlivé tweety a ich retweets

**Odpoveď:**



Obrázok Dátový model Mongo

Pri tvorbe dátového modelu (obrázok č.1) sme zvolili prístup, v ktorom väčšina dát bude obsiahnutých v jednej collection (embedded prístup), a to v conversations. To hlavne z dôvodu, že tieto dáta sa v našej pomyselnej aplikácií nebudú (často) meniť a bez samotných tweetov by nemali veľký zmysel. To sa však netýka autorov-užívateľov, ktorých ukladáme (referenced prístup) do vlastnej collection authors. Je to najmä kvôli tomu, aby sme pri zmene údajov autora (čo je prípad, ktorý nastáva veľmi často) nemuseli meniť tieto údaje v každom jednom tweete, ktorý tento autor postol. Taktiež majú dáta autorov zmysel aj bez dát o tweetoch, čo je ďalší dôvod, prečo je dobré mať ich vo vlastnej collection. Čo sa týka referencovaných tweetov, tiež sme zvolili embedded prístup, keďže MongoDB neumožňuje referenced prístup v rámci jednej collection. Teda ak by sme chceli využiť referenced prístup, museli by sme mať vlastnú kolekciu pre referencované tweety. To by veľmi nedávalo zmysel, mať dve kolekcie na tweety, kvôli čomu sme zvolili embedded prístup (s tým, že referencované tweety nebudú mať len základné informácie o nich, aby nám napríklad nevznikali nekonečný embedding v rámci jednej collection). Na obrázku č.2,3 si vieme pozrieť príklad štruktúry importovaných dát.

****

Obrázok Príklad štruktúry dát z collection authors

**Obrázok, na ktorom je text

Automaticky generovaný popis** 

Obrázok Príklad štruktúry dát z collection covnersations

# 2. Import

**Otázka:**

Nainštalujte alebo využite online inštanciu MongoDB servera, do ktorého importujte všetky tweets (a s nimi spojené data – anotácie, referencie, odkazy a informácie o kontexte) zo dňa 24.02.2022.

**Odpoveď:**

Na túto úlohu sme využili online inštanciu MongoDB servera. Najskôr sme exportovali dáta v požadovanej štruktúre pomocou query (obrázok č.4 -> authors collection, obrázok č.5 -> conversations collection) z postgres databázy.

Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok Export dát potrebných pre author collection

Obrázok, na ktorom je stôl

Automaticky generovaný popis Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok Export dát potrebných pre conversations collection

Keď sme dokončili export dát z postgresu, pomocou python scriptu (obrázok č.6) sme nahrali dáta do online MongoDB clusteru. Stav databázy po importe si vieme pozrieť na obrázku č.7.

Obrázok, na ktorom je text

Automaticky generovaný popis Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok Python script pre import dát na MongoDB cluster

Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok Stav databázy po importe

# 3. Dopyty

**Otázka:**

Napíšte dotaz, ktorý nad importovanou databázou:

a. Vypíše posledných 10 tweetov pre autora, ktorý má username Newnews\_eu

b. Vypíše posledných 10 retweetov pre tweet, ktorý má id 1496830803736731649.

**Odpoveď:**

Pre úlohu a) sme vytvorili dopyt (obrázok č.8), ktorý najskôr spojí collection conversation s collection authors na základe autorovho id (podobne ako join pri relačných databázach). Následne pomocou unwind nahradí pole autorov (v tomto prípade je autor vždy len jeden, takže je to skôr kozmetická úprava, vďaka ktorej vieme pracovať priamo so záznamom a nie s poľom jedného prvku) prvkom autora z tohto poľa. Pomocou match ponecháme záznamy daného autora (Newnews\_eu) a v závere usporiadame výsledky podľa dátumu created\_at a ponecháme 10 výsledkov.

Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok Query pre vypísanie posledných 10 tweetov autora Newnews\_eu

Pre úlohu b) sme vytvorili dopyt (obrázok č.9), ktorý rozbalí pole s referenciami pomocou unwind a pozrie sa, či v týchto referenciách je retweet konverzácie s požadovaným id. Následne tieto výsledky zoradí podľa času created\_at a vráti prvých 10.

Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok Posledných 10 retweetov tweetu s ID 1496830803736731649