Zadanie 1 – import tweetov do PostgreSQL

Odovzdanie do 7.10.2021 23:59 – dostanete za to 7 bodov.

Prvé zadanie je zamerané na overenie zručností efektívnej práce s veľkými dátami. Vašou úlohou je podľa priloženej schémy importovať dáta z `authors.jsonl.gz` a `conversations.jsonl.gz`. Pre pochopenie formátu si naštudujte https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet a https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/user. Programovať môžete v hocijakom jazyku a môžete používať aj ORM. Váš zdroják sa odovzdáva do AISu, no rovnako **MUSÍ** byť zavesený na GitHub – v dokumente na začiatku uveďte vždy linku na projekt. Ak nebude GitHub alebo odkaz v dokumente, nebudú body. Okrem zdrojového kódu odovzdávate aj dokument kde opisujete prečo ste zvolili daný prístup a aký to malo vplyv na spracovávanie záznamov.

Hodnotenie

- importovanie správneho počtu záznamov 3 body
- efektívnosť čítania dát zo súboru 1 bod
- efektívnosť zapisovania dát do databázy 1 bod
- estetika kódu 1 bod
- úroveň protokolu a z neho zjavnosť výveru vašeho riešenia 1 bod

Obsah protokolu

1. Opis algoritmu a postupu pri importe
2. Použité technológie - zdôvodnite výber (argument, že je to jediná technológia, s ktorou viem pracovať nie je platný)
3. Napísané a vysvetlené každé SQL, ktoré program vykoná a zhodnotená jeho efektívnosť (aj keď používate ORM)
4. Dĺžka trvania importu a časový opis priebehu
    1. potrebné evidovať pre každých spracovaných 10k záznamov aktuálny čas (vo formáte ISO-8601) s aktuálnou dĺžkou trvania celého importu a tiež trvanie spracovania aktuálneho bloku vo formáte `mm:ss` v priloženom CSV v UTF-8 a semi-colomn separated. Ukážka (aktuálny čas; celkový čas; čas pre aktuálny blok): `2022-09-16T19:35Z;81:22;10:22`
5. Počet a veľkosť záznamov v každej tabuľke ako screenshot

Upozornenia

1. Pozor na UTF-8 NULL
2. Vstup môže obsahovať duplikáty konverzácií (môžete naraziť na `conversations`, ktoré majú rovnaké ID)
3. Nositeľ jedinečnosti je
    1. `id` v `conversations`
    2. `tag` v `hashtags`
4. Číselníky sú vždy unikátna k-tica, rozlišujeme malé a veľké znaky (príklad: `PDT`, `pdT` a `pdt` sú 3 rozdielne záznamy)
5. Maximálna dĺžka URL je 2048 znakov, ak narazíte na viac, preskočte daný záznam (záznam v `conversations` ale bude vytvorený, nevznikne nový záznam v `links`)
6. Import môže trvať dlho (veľmi neoptimálne riešenie bežalo 2 dni na MacBook M1 MAX - toto sa dá rádovo znížiť správnym prístupom na hodiny) - odporúčame nezačať v piatok ráno
7. Optimalizujte si svoj PostgreSQL server (pred importom)
8. Dodržiavajte naming z diagramu
9. Výsledná databáza má okolo 49GB

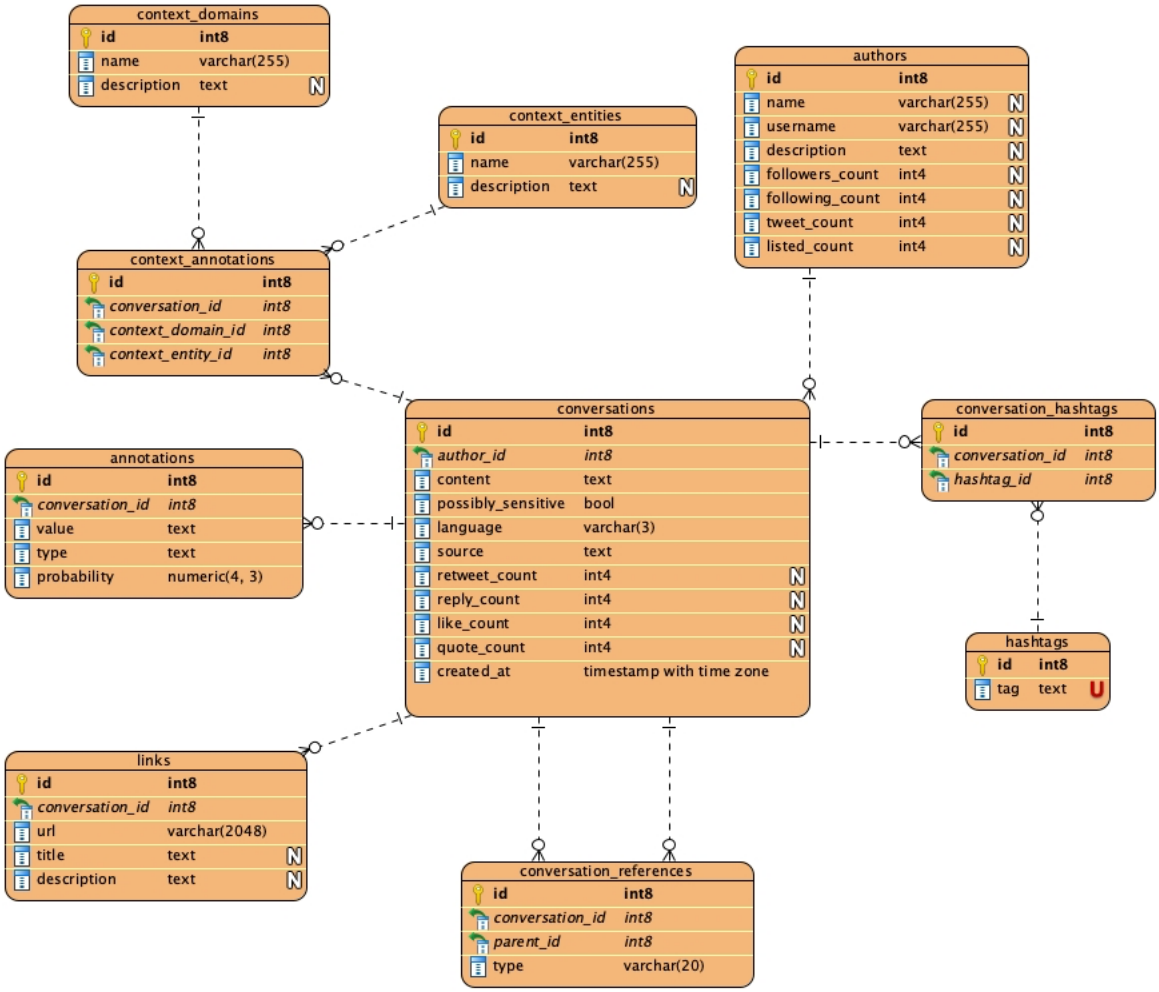
Vstupný dataset

Vstupné dáta sú publikované na Google Drive a pozostávajú z dvoch súborov, ktoré sú vo formáte JSON Lines (každý riadok obsahuje práve jeden JSON objekt podľa Twitter dokumentácie). Vstup je ešte komprimovaný pomocou gzip (surové súbory sú naozaj veľké - `conversations.jsonl` po dekompresii má cca 63GB). Pri implementácii je vhodné využívať všetky milé vlastnosti gzip formátu (alebo si budeš musieť na pár týždňov odinštalovať svoju obľúbenú hru).

Pristupujte cez váš stuba.sk Google účet (požiadavky na prístup, cez súkromný email budú zamietnuté a navyše z nich budeme trochu mrzutí).

https://drive.google.com/drive/folders/1sHMYxkSK9WHCUa0rL0IaP7KaLhsek5qS?usp=sharing

Model



Referencie

https://www.researchgate.net/publication/332373250_For_Whom_the_Bot_Tolls_A_Neural_Networks_Approach_to_Measuring_Political_Orientation_of_Twitter_Bots_in_Russia

https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet

https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/user