Zadanie 2 – vyhľadávanie a indexovanie

Odovzdanie do 24.10.2021 23:59 – máte na to 2 týždne – dostanete za to 7,5 boda.

Otázky 2-17 sú dokopy za 5,5 boda (každá rovnako). Zadanie 18 je za 2 body. Teda good luck and have fun.

Zadania prosím neopisujte jednoslovne ale zmysluplnou vetou (nie slohová práca - teda vecne). Vždy priložte screenshot z explain analyse.

1. Vyhľadajte v accounts screen_name s presnou hodnotou ‘realDonaldTrump’ a analyzujte daný select. Akú metódu vám vybral plánovač a prečo - odôvodnite prečo sa rozhodol tak ako sa rozhodol?

2. Koľko workerov pracovalo na danom selecte a na čo slúžia? Zdvihnite počet workerov a povedzte ako to ovplyvňuje čas. Je tam nejaký strop? Ak áno, prečo? Od čoho to závisí?

3. Vytvorte btree index nad screen_name a pozrite ako sa zmenil čas a porovnajte výstup oproti požiadavke bez indexu. Potrebuje plánovač v tejto požiadavke viac workerov? Čo ovplyvnilo zásadnú zmenu času?

4. Vyberte používateľov, ktorý majú followers_count väčší, rovný ako 100 a zároveň menší, rovný 200. Je správanie rovnaké v prvej úlohe? Je správanie rovnaké ako v tretej úlohe? Prečo?

5. Vytvorte index nad 4 úlohou a popíšte prácu s indexom. Čo je to Bitmap Index Scan a prečo je tam Bitmap Heap Scan? Prečo je tam recheck condition?

6. Vyberte používateľov, ktorí majú followers_count väčší, rovný ako 100 a zároveň menší, rovný 1000? V čom je rozdiel, prečo?

7. Vytvorte ďalšie 3 btree indexy na name, friends_count, a description a insertnite si svojho používateľa (to je jedno aké dáta) do accounts. Koľko to trvalo? Dropnite indexy a spravte to ešte raz. Prečo je tu rozdiel?

8. Vytvorte btree index nad tweetami pre retweet_count a pre content. Porovnajte ich dĺžku vytvárania. Prečo je tu taký rozdiel? Čím je ovplyvnená dĺžka vytvárania indexu a prečo?

9. Porovnajte indexy pre retweet_count, content, followers_count, screen_name,... v čom sa líšia a prečo (opíšte výstupné hodnoty pre všetky indexy)?
a. create extension pageinspect;
b. select * from bt_metap('idx_content');
c. select type, live_items, dead_items, avg_item_size, page_size, free_size from bt_page_stats('idx_content',1000);
d. select * from bt_page_items('idx_content',1) limit 1000;

10. Vyhľadajte v tweets.content meno „Gates“ na ľubovoľnom mieste a porovnajte výsledok po tom, ako content naindexujete pomocou btree. V čom je rozdiel a prečo?

11. Vyhľadajte tweet, ktorý začína “The Cabel and Deep State”. Použil sa index?

12. Teraz naindexujte content tak, aby sa použil btree index a zhodnoťte prečo sa pred tým nad “The Cabel and Deep State” nepoužil. Použije sa teraz na „Gates“ na ľubovoľnom mieste? Zdôvodnite použitie alebo nepoužitie indexu?

13. Vytvorte nový btree index, tak aby ste pomocou neho vedeli vyhľadať tweet, ktorý končí reťazcom „idiot #QAnon“ kde nezáleží na tom ako to napíšete. Popíšte čo jednotlivé funkcie robia.

14. Nájdite účty, ktoré majú follower_count menší ako 10 a friends_count väčší ako 1000 a výsledok zoraďte podľa statuses_count. Následne spravte jednoduché indexy a popíšte ktoré má a ktoré nemá zmysel robiť a prečo.

15. Na predošlú query spravte zložený index a porovnajte výsledok s tým, keď je sú indexy separátne. Výsledok zdôvodnite.

16. Upravte query tak, aby bol follower_count menší ako 1000 a friends_count väčší ako 1000. V čom je rozdiel a prečo?

17. Vytvorte vhodný index pre vyhľadávanie písmen bez kontextu nad screen_name v accounts. Porovnajte výsledok pre vyhľadanie presne ‘realDonaldTrump’ voči btree indexu? Ktorý index sa vybral a prečo? Následne vyhľadajte v texte screen_name ‘ldonaldt‘ a porovnajte výsledky. Aký index sa vybral a prečo?

18. Vytvorte query pre slová "John" a "Oliver" pomocou FTS (tsvector a tsquery) v angličtine v stĺpcoch tweets.content, accounts.decription a accounts.name, kde slová sa môžu nachádzať v prvom, druhom ALEBO treťom stĺpci. Teda vyhovujúci záznam je ak aspoň jeden stĺpec má „match“. Výsledky zoraďte podľa retweet_count zostupne. Pre túto query vytvorte vhodné indexy tak, aby sa nepoužil ani raz sekvenčný scan (správna query dobehne rádovo v milisekundách, max sekundách na super starých PC). Zdôvodnite čo je problém s OR podmienkou a prečo AND je v poriadku pri joine.