

Slovenská Technická Univerzita
Fakulta Informatiky a Informačných technológií

Pokročilé Databázové Technológie
Zadanie 4

Contents

Zadanie	3
Úloha č.1	4
Povodný Select v postgrese:	4
ERROR, ktorý to vyhodilo po hodine.	5
Zmena plánu.....	5
Úloha č.2	9
Úloha č.3	10
Zhodnotenie.....	12

Zadanie

4. zadanie zamerané na MongoDB:

Odovzdanie do 24.11.2021 23:59 – máte na to 2 týždne – dostanete za to 5 bodov. Otázky 1 a 2 sú za 2 body a posledná je za 1 bod.

1. Navrhnete dátový model (kolekcie a formát dokumentov) v MongoDB pre dataset tweetov využívaný aplikáciou, ktorá bude najmä:

- a. Zobrazovať tweety jednotlivých používateľov vo forme feedu
- b. Po kliknutí na jednotlivé tweety zobrazí retweety daného tweetu

Využite viac vzorov dátového modelovania - embedding aj referencing, podľa vhodnosti. Popíšte hlavné dôvody pre Vami zvolenú dátovú štruktúru. Existuje viacero správnych riešení. Návrh ukážte vo formáte JSON nad nejakým ukážkovým tweetom a štruktúru zôdvocnite.

2. Nainštalujte si lokálne alebo využite online službu pre vytvorenie MongoDB databázy a naimportujte do neho existujúci dataset (len extrémne tweety z prvého zadania), transformovaný na váš dátový model.

3. Napíšte query pre Vašu importovanú databázu pre dva hlavné prípady použitia:

- a. Vypísať posledných 10 tweetov accountu so screen_name = Marndin12, spolu s údajmi o accounte
- b. Vypísať prvých 10 tweetov - text, meno autora, dátum tweetu a hashtagy, ktoré retweetujú tweet s id = 1243427980199641088

Úloha č.1

Najprv som sa zamýšľal ako vôbec spraviť nejaký návrh NoSql databázy. Po prečítaní zadania som sa zamyslel, čo potrebujem najviac a vypadlo z toho, že potrebujem userov, ich tweety a retweety tweetov, čo v najrýchlejšom čase.

Z tohto dôvodu som aj schému navrhol tak, že optimalizujem na čas a nie na priestor. Každý takýto problém mi príjde tradeoff času vs miesta v pamäti.

Na ďalšej strane môžeme vidieť pseudo JSON modelu ako bude jeden account vyzeráť. Jeden account môže mať viacero tweetov, a tweet môže mať viacero retweetov. Ak by sme chceli retweet retweetu, je potrebné nájsť autora retweetu a pozrieť sa tam, čo by trvalo značne dlhšie.

Toto ale zadanie nepožaduje, preto som to navrhol takto. Najviac pamäte kvôli duplicitám tweetov ale zasa nemusím hľadať tweety a retweety cez referencie.

Tak po 5 hodinách som prišiel na nejaké zásadné zistenia. Po viac ako 60 min generovania JSONu z PGAdmina to robilo blbosti s escapovaním JSON vecí, ktoré som riesil asi 2h. Nakoniec som našiel regex, ktorý pomohol. `'(\\V|#|\\t|\\b|\\f|\\n|\\r|\\\\|\\V|)"?'`. Toto ale nie je ideálne lebo alterujem, to čo je reálne v stringu. Len zo záujmu som sa pýtal kamaráta a nejak sme dedukciou zistili, že to musí robiť niektorý z commandov, ktorý sa vykoná nad stringom po `to_json()`. Bolo to aj tak. Tento príkaz bol COPY do súboru. Po spustení selectu v konzole toto už nerobí a vygeneruje validný JSON hneď. Teda nie je treba zložitý Regex.

Povodný Select v postgrese:

```
COPY (SELECT array_to_json(array_agg(results))
```

```
FROM (
```

```
  SELECT id, regexp_replace(screen_name, '(\\V|#|\\t|\\b|\\f|\\n|\\r|\\\\|\\V|)"?', '', 'g') screen_name,
  regexp_replace(name, '(\\V|#|\\t|\\b|\\f|\\n|\\r|\\\\|\\V|)"?', '', 'g') as name,
  regexp_replace(description, '(\\V|#|\\t|\\b|\\f|\\n|\\r|\\\\|\\V|)"?', '', 'g') description, followers_count,
  friends_count, statuses_count,
```

```
  (
```

```
    SELECT array_to_json(array_agg(o))
```

```
  FROM (
```

```
    SELECT id, regexp_replace(content, '(\\V|#|\\t|\\b|\\f|\\n|\\r|\\\\|\\V|)"?', '', 'g') as content,
  location, retweet_count, favorite_count, neg, neu, compound, happened_at, author_id, country_id,
  parent_id,
```

```
    ( SELECT array_to_json(array_agg(b))
```

```
    FROM (
```

```
      SELECT id, regexp_replace(content,
'\\V|#|\\t|\\b|\\f|\\n|\\r|\\\\|\\V|)"?', '', 'g') as content, location, retweet_count, favorite_count, neg,
neu, compound, happened_at, author_id, country_id, parent_id
```

```

FROM tweets rt
WHERE rt.parent_id = t.id
) b
) AS retweets
FROM tweets t
WHERE t.author_id = accounts.id
AND ( t.compound > 0.5 OR t.compound < -0.5 )
) o
) AS tweets
FROM accounts
) results) TO '/home/data/tweets_final.json' WITH (FORMAT text, HEADER FALSE);

```

[ERROR, ktorý to vyhodilo po hodine.](#)

ERROR: array size exceeds the maximum allowed (1073741823)

SQL state: 54000

Zmena plánu

Po premyslení mi došlo, že to takto nepojde vytiahnuť z postgre. Preto som prehodil model tak, že namiesto celých tweetov tam budu iba ID. Čo ušetrí na pamäti.

Model je približne rovnaký ako na obrázku, s tým že accounts nemá tweety cele ale len ID a retweety má daný tweet ako dokument. A zároveň tweet má v sebe hashtagy.

Account

```

[
{
  _id: ObjectId("619a58dae90eea21cfe33a94"),
  id: 785,
  screen_name: 'kfury',
  followers_count: null,
  friends_count: null,

```

```
statuses_count: null,
tweets: [
  { id_tweet: '1232041399236775937' },
  { id_tweet: '1232042917625135104' },
  { id_tweet: '1233895109168525312' },
  { id_tweet: '1223349793084231680' },
  { id_tweet: '1249841745250009089' },
  { id_tweet: '1255718427353743361' },
  { id_tweet: '1257381971535319040' },
  { id_tweet: '1251024241593470976' },
  { id_tweet: '1260047459432054784' }
]
}
```

Tweet

```
{
  _id: ObjectId("619947968a287b7b37165564"),
  id: '1034289685227614208',
  content: '\n' +
    'PEDOPHILIA IS AN EPIDEMIC IN USA-WORLD: \n' +
    '\n' +
    'So called elites-many pedophile rings connected to Royals-Politicians-Hollywood. Our Justice Dept. @ AG Sessions uses padded gloves = , only small fish-no big fish. \n' +
    '\n' +
    '\n' +
    '\n' +
    '! https://t.co/oScb63oA4y',
  location: null,
  retweet_count: 56,
```

```
favorite_count: 37,  
neg: 0,  
neu: 0.88,  
pos: 0.12,  
compound: 0.5707,  
happened_at: '2018-08-28T04:00:51+00:00',  
author_id: 553991829,  
country_id: null,  
parent_id: null,  
hashtags: [  
  { value: 'TrueMAGA' },  
  { value: 'CrimeMonitor' },  
  { value: 'PedogateIsReal' },  
  { value: 'PedoPope' },  
  { value: 'JUSTICENOW' }  
]  
}
```

Accounts

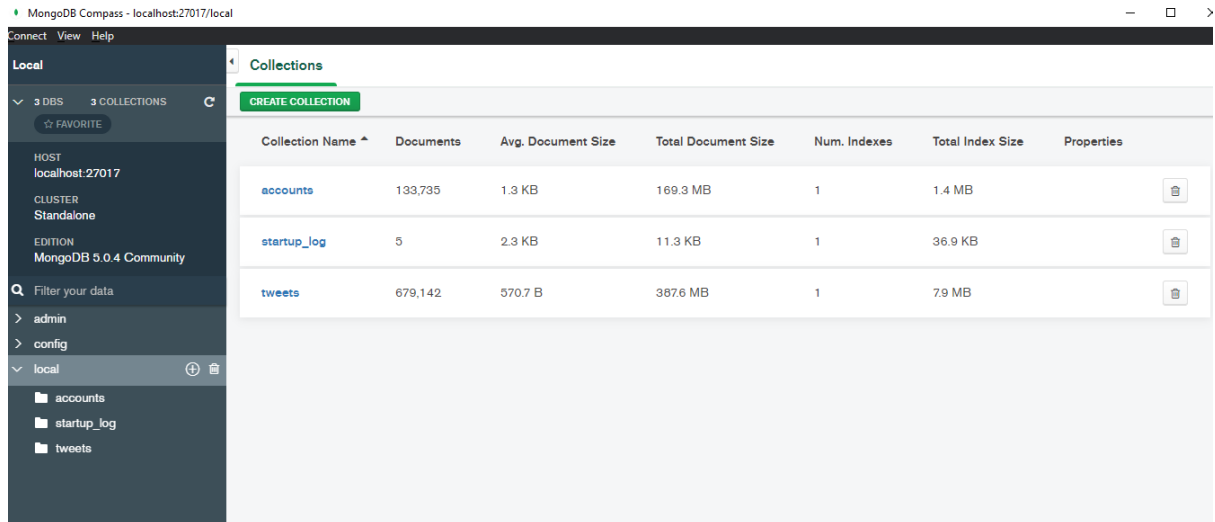
```
{
  id,
  screen_name,
  name,
  description,
  followers_count,
  friends_count,
  statuses_count,
  tweets: [
    {
      id,
      content,
      retweet_count,
      favorite_count,
      neg,
      neu,
      pos,
      compound,
      happened_at,
      author_id,
      country_id,
      parent_id,
      retweets: [
        {
          id,
          content,
          retweet_count,
          favorite_count,
          neg,
          neu,
          pos,
          compound,
          happened_at,
          author_id,
          country_id,
          parent_id
        }
      ]
    },
    {
      id,
      content,
      retweet_count,
      favorite_count,
      neg,
      neu,
      pos,
      compound,
      happened_at,
      author_id,
      country_id,
      parent_id,
      retweets: [
        {
          id,
          content,
          retweet_count,
          favorite_count,
          neg,
          neu,
          pos,
          compound,
          happened_at,
          author_id,
          country_id,
          parent_id
        }
      ]
    }
  ]
}
```



tweets.json

Úloha č.2

Nainstaloval som si Mongo shard u seba. A importoval obrovské JSONy z postgre... Trvalo mi asi 8 hodín, kým všetko sadlo tak ako ma a podarilo sa mi dostať sem. Stále to failovalo kvôli nejakým znakom v name atd...



The screenshot shows the MongoDB Compass interface for a local database. The left sidebar displays the database structure, including the 'local' database and its collections: 'accounts', 'startup_log', and 'tweets'. The main panel shows a table of collections with the following data:

Collection Name	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size	Properties
accounts	133,735	1.3 KB	169.3 MB	1	1.4 MB	
startup_log	5	2.3 KB	11.3 KB	1	36.9 KB	
tweets	679,142	570.7 B	387.6 MB	1	7.9 MB	

Úloha č.3

Vypísať posledných 10 tweetov accountu so screen_name = Marndin12, spolu s údajmi o accounte

Moja funkcia

```
function gif_10_tweets(Name) {  
    var myCursor = db.accounts.find({"screen_name": Name});  
    var myDocument = myCursor.hasNext() ? myCursor.next() : null;  
    if (myDocument) {  
        var tweet_ids = myDocument.tweets;  
        var tweets_arr = [];  
        for (let i = 0; i < tweet_ids.length; i++) {  
            tweets_arr.push(tweet_ids[i].id_tweet);  
        }  
        return db.tweets.aggregate([  
            { $match: { "id": { $in: tweets_arr } }},  
            { $sort: { happened_at : -1 }},  
            { $limit: 10}  
        ]);  
    }  
}
```

gif_10_tweets("Marndin12")



10_tweetov.txt

Vypísať prvých 10 tweetov - text, meno autora, dátum tweetu a hashtagy, ktoré retweetujú tweet s id = 1246874043682299904

Kedže som už na tomto zadani strávil neuuumerne veľa času(aktualne asi 10 hodín) nejdem už ďalej dorábať, pretože som zistil že som si zabudol z postgresu vybrať aj retweety pre tweety. Lahko by som ale upravil select aby to vracal, importoval. Upravil som tento kusok kódu a vypadlo mi z toho všetko čo treba. Zasa ale ďalší čas, ktorý bohužiaľ už nemám.

```
function gif_10_retweets(ID) {  
  var myCursor = db.accounts.find({"screen_name": ID});  
  var myDocument = myCursor.hasNext() ? myCursor.next() : null;  
  if (myDocument) {  
    var tweet_ids = myDocument.retweets;  
    var tweets_arr = [];  
    for (let i = 0; i < tweet_ids.length; i++) {  
      tweets_arr.push(tweet_ids[i].id_tweet);  
    }  
    return db.tweets.aggregate([  
      { $match: { "id": { $in: tweets_arr } }},  
      { $sort: { happened_at: -1 }},  
      { $limit: 10 }  
    ]);  
  }  
}
```

Zhodnotenie

Nekonečné zadanie za 5 bodov, ktoré ma síce bavilo ale za počet bodov, povedzme si úprimne, nestojí. Zo samotného zadania drvivú väčšinu času bralo exportovanie postgresu do JSON, pretože nemajú spravený rozumný copy do súboru ale robí to hovadiny. Mongo ako také je zaujímavé a možno sa mu budem venovať ďalej. **Ďakujem za pochopenie** a dúfam, že moju snahu oceníte.