

Udiddit, a social news aggregator

Introduction

Udiddit, a social news aggregation, web content rating, and discussion website, is currently using a risky and unreliable Postgres database schema to store the forum posts, discussions, and votes made by their users about different topics.

The schema allows posts to be created by registered users on certain topics, and can include a URL or a text content. It also allows registered users to cast an upvote (like) or downvote (dislike) for any forum post that has been created. In addition to this, the schema also allows registered users to add comments on posts.

Here is the DDL used to create the schema:

```
CREATE TABLE bad_posts (  
    id SERIAL PRIMARY KEY,  
    topic VARCHAR(50),  
    username VARCHAR(50),  
    title VARCHAR(150),  
    url VARCHAR(4000) DEFAULT NULL,  
    text_content TEXT DEFAULT NULL,  
    upvotes TEXT,  
    downvotes TEXT  
);  
  
CREATE TABLE bad_comments (  
    id SERIAL PRIMARY KEY,  
    username VARCHAR(50),  
    post_id BIGINT,  
    text_content TEXT  
);
```

Part I: Investigate the existing schema

As a first step, investigate this schema and some of the sample data in the project's SQL workspace. Then, in your own words, outline three (3) specific things that could be improved about this schema. Don't hesitate to outline more if you want to stand out!

1. The data needs to be normalized.

2. bad_posts Table

- title should have a not null constraint
- upvotes should not be text, it should be integer
- downvotes should not be text, it should be integer
- There are no indexes setup for this table, having indexes setup can improve query
- There is no unique constraint on the username, so there can be multiple users with the same username.

3. bad_comments Table

- post_id should be a foreign key
- There are No indices to make searches quicker.

Part II: Create the DDL for your new schema

Having done this initial investigation and assessment, your next goal is to dive deep into the heart of the problem and create a new schema for Udiddit. Your new schema should at least reflect fixes to the shortcomings you pointed to in the previous exercise. To help you create the new schema, a few guidelines are provided to you:

1. Guideline #1: here is a list of features and specifications that Udiddit needs in order to support its website and administrative interface:
 - a. Allow new users to register:
 - i. Each username has to be unique
 - ii. Usernames can be composed of at most 25 characters
 - iii. Usernames can't be empty
 - iv. We won't worry about user passwords for this project
 - b. Allow registered users to create new topics:
 - i. Topic names have to be unique.
 - ii. The topic's name is at most 30 characters
 - iii. The topic's name can't be empty
 - iv. Topics can have an optional description of at most 500 characters.
 - c. Allow registered users to create new posts on existing topics:
 - i. Posts have a required title of at most 100 characters
 - ii. The title of a post can't be empty.
 - iii. Posts should contain either a URL or a text content, **but not both**.
 - iv. If a topic gets deleted, all the posts associated with it should be automatically deleted too.
 - v. If the user who created the post gets deleted, then the post will remain, but it will become dissociated from that user.
 - d. Allow registered users to comment on existing posts:
 - i. A comment's text content can't be empty.
 - ii. Contrary to the current linear comments, the new structure should allow comment threads at arbitrary levels.
 - iii. If a post gets deleted, all comments associated with it should be automatically deleted too.
 - iv. If the user who created the comment gets deleted, then the comment will remain, but it will become dissociated from that user.
 - v. If a comment gets deleted, then all its descendants in the thread structure should be automatically deleted too.
 - e. Make sure that a given user can only vote once on a given post:
 - i. Hint: you can store the (up/down) value of the vote as the values 1 and -1 respectively.
 - ii. If the user who cast a vote gets deleted, then all their votes will remain, but will become dissociated from the user.

- iii. If a post gets deleted, then all the votes for that post should be automatically deleted too.
2. Guideline #2: here is a list of queries that Udiddit needs in order to support its website and administrative interface. Note that you don't need to produce the DQL for those queries: they are only provided to guide the design of your new database schema.
- a. List all users who haven't logged in in the last year.
 - b. List all users who haven't created any post.
 - c. Find a user by their username.
 - d. List all topics that don't have any posts.
 - e. Find a topic by its name.
 - f. List the latest 20 posts for a given topic.
 - g. List the latest 20 posts made by a given user.
 - h. Find all posts that link to a specific URL, for moderation purposes.
 - i. List all the top-level comments (those that don't have a parent comment) for a given post.
 - j. List all the direct children of a parent comment.
 - k. List the latest 20 comments made by a given user.
 - l. Compute the score of a post, defined as the difference between the number of upvotes and the number of downvotes
3. Guideline #3: you'll need to use normalization, various constraints, as well as indexes in your new database schema. You should use named constraints and indexes to make your schema cleaner.
4. Guideline #4: your new database schema will be composed of five (5) tables that should have an auto-incrementing id as their primary key.

Once you've taken the time to think about your new schema, write the DDL for it in the space provided here:

-- 1. Users table

```
CREATE TABLE users (  
  "id" BIGSERIAL PRIMARY KEY, "username" VARCHAR(25) UNIQUE,  
  CHECK(LENGTH(TRIM("username"))>0), "last_login" TIMESTAMP  
);  
COMMENT ON COLUMN users.id IS 'large no of users expected';  
COMMENT ON COLUMN users.username IS 'username must be unique, max 25 char,  
not empty';  
CREATE INDEX "search_by_usrname" ON users (username VARCHAR_PATTERN_OPS);
```

-- 2. Topics table

```
CREATE TABLE topics (  
  "id" SERIAL PRIMARY KEY,  
  "topic_name" VARCHAR(30) UNIQUE NOT NULL,  
  "description" VARCHAR(500),  
  CONSTRAINT "non_empty_topic" CHECK(LENGTH(TRIM("topic_name"))>0)  
);
```

```
COMMENT ON COLUMN topics.topic_name IS 'topic must be unique, max 30 char,  
not empty';  
COMMENT ON COLUMN topics.description IS 'description max 500 char, optional';
```

-- 3. Posts table

```
CREATE TABLE posts (  
  "id" SERIAL PRIMARY KEY,  
  "title" VARCHAR(100),  
  "content" TEXT,  
  "url" VARCHAR,  
  "post_timestamp" TIMESTAMP WITHOUT TIME ZONE,  
  "topic_id" INTEGER NOT NULL,  
  "user_id" BIGINT NOT NULL,  
  CONSTRAINT "posts_topic_fk" FOREIGN KEY ("topic_id")  
    REFERENCES topics ("id") ON DELETE CASCADE,  
  CONSTRAINT "posts_usr_fk" FOREIGN KEY ("user_id")  
    REFERENCES users ("id") ON DELETE SET NULL,  
  CONSTRAINT "nospaces_non_empty_title" CHECK (LENGTH(TRIM(title)) > 0),  
  CONSTRAINT "url_content_null_check" CHECK ("url" IS NOT NULL OR "content" IS  
  NOT NULL)  
);  
COMMENT ON COLUMN posts.topic_id IS 'If a topic gets deleted, all the posts  
associated with it should be automatically deleted too';
```

```
COMMENT ON COLUMN posts.user_id IS 'If the user who created the post gets deleted ,
then the post will remain, but it will become dissociated from that user.';
CREATE INDEX "users_no_post_idx" ON posts (user_id,title);
-- CREATE INDEX "posts_url_idx" ON posts (url,)
```

-- 4. Comments table

```
CREATE TABLE post_comments (
    "id" SERIAL PRIMARY KEY,
    "parent_comment_id" INTEGER,
    "comments" TEXT NOT NULL,
    "user_id" BIGINT NOT NULL,
    "post_id" INTEGER NOT NULL,
    CONSTRAINT "user_id_fk" FOREIGN KEY ("user_id")
        REFERENCES users("id") ON DELETE SET NULL,
    CONSTRAINT "post_id_fk" FOREIGN KEY ("post_id")
        REFERENCES posts("id"),
    CONSTRAINT "initial_comment" FOREIGN KEY ("parent_comment_id")
        REFERENCES post_comments("id") ON DELETE CASCADE
);
COMMENT ON CONSTRAINT "initial_comment" ON post_comments IS 'references
post_comments id column with the intent of creating threaded comment
structure';
COMMENT ON COLUMN post_comments.user_id IS 'if user deleted, comment
dissociated from user';
-- CREATE INDEX "all_posts_given_user" ON post_comments (post_id,user_id);
```

-- 5. Post Up/Downvotes table

```
CREATE TABLE post_votes (  
    "post_id" BIGINT NOT NULL,  
    "user_id" BIGINT NOT NULL,  
    "vote" SMALLINT DEFAULT 0,  
    CONSTRAINT "post_id_delete_votes_idx" FOREIGN KEY ("post_id")  
REFERENCES posts("id") ON DELETE CASCADE,  
    CONSTRAINT "user_id_idx" FOREIGN KEY ("user_id") REFERENCES users("id")  
ON DELETE SET NULL,  
    CONSTRAINT "post_user_composite_pk" PRIMARY KEY ("post_id", "user_id"),  
    CONSTRAINT "up_down_vote_chk" CHECK ("vote" = 1 OR "vote" = -1)  
);
```

Part III: Migrate the provided data

Now that your new schema is created, it's time to migrate the data from the provided schema in the project's SQL Workspace to your own schema. This will allow you to review some DML and DQL concepts, as you'll be using INSERT...SELECT queries to do so. Here are a few guidelines to help you in this process:

1. Topic descriptions can all be empty
2. Since the bad_comments table doesn't have the threading feature, you can migrate all comments as top-level comments, i.e. without a parent

3. You can use the Postgres string function **regexp_split_to_table** to unwind the comma-separated votes values into separate rows
4. Don't forget that some users only vote or comment, and haven't created any posts. You'll have to create those users too.
5. The order of your migrations matter! For example, since posts depend on users and topics, you'll have to migrate the latter first.
6. Tip: You can start by running only SELECTs to fine-tune your queries, and use a LIMIT to avoid large data sets. Once you know you have the correct query, you can then run your full INSERT...SELECT query.
7. **NOTE:** The data in your SQL Workspace contains thousands of posts and comments. The DML queries may take at least 10-15 seconds to run.

Write the DML to migrate the current data in bad_posts and bad_comments to your new database schema:

```
-- For users table
-- Populate users table with unique usernames from bad_posts
```

```
INSERT INTO users ("username")
SELECT DISTINCT username
FROM bad_posts
UNION
SELECT DISTINCT username
FROM bad_comments;
```

```
-- Populate topic_name with topics from bad_posts
```

```
INSERT INTO topics ("topic_name") SELECT DISTINCT LEFT(topic,30) FROM bad_posts;
```


-- Populate topic_id and user_id of posts table after innerjoin with bad_posts

```
INSERT INTO posts ("topic_id", "user_id", "url", "content", "title")
SELECT bp_sub.topid, users.id, bp_sub.url, bp_sub.tc, LEFT(bp_sub.title, 100)
FROM (
SELECT topics.id topid, bp.username uname, bp.url, bp.text_content tc, bp.title
FROM topics
JOIN bad_posts bp
ON bp.topic = topics.topic_name
) bp_sub
JOIN users
ON users.username = bp_sub.uname JOIN topics t
ON t.id = bp_sub.topid;
```

-
- Populate post_comments table with comments from bad_comments, be sure to inner join with username and post_id and select for text_content

```
INSERT INTO post_comments ("user_id", "post_id", "comments")
SELECT usr_sub.id, bc.post_id, bc.text_content
FROM (
SELECT id, username
FROM users
) usr_sub
JOIN bad_comments bc
ON bc.username = usr_sub.username;
```

-- Populate votes for all posts by a user by other users
-- inner join based on the username and then assign a value +1/- 1 based on whether its taken from upvotes or downvotes column
-- For all upvotes, select ids that match and assign 1 to upvote column

```
INSERT INTO post_votes ("post_id", "user_id", "vote")
SELECT bp_u.id, u.id, 1 upvote
FROM (
SELECT id, REGEXP_SPLIT_TO_TABLE(upvotes, ',') usernames
FROM bad_posts
) bp_u
JOIN users u
ON u.username = bp_u.usernames;
```

-- For all downvotes, select ids that match and assign -1 to downvote column

```
INSERT INTO post_votes ("post_id","user_id","vote") SELECT bp_d.id, u.id,  
-1 AS downvote  
FROM (  
SELECT id, REGEXP_SPLIT_TO_TABLE(downvotes,',') usernames  
FROM bad_posts  
)bp_d  
JOIN users u  
ON u.username=bp_d.usernames;
```

```
DROP TABLE bad_posts;  
DROP TABLE bad_comments;
```