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!

Problems with existing schema and suggestions for improvement:

1. Instead of having a separate “users” table, the present schema has user information such as “username” appearing in both the tables (“bad_posts” and “bad_comments”). The better way could be to create a “users” table with a surrogate key column as “user_id” which will contain a unique id for each username. This column will also be the primary key of the “users” table.
2. The table “bad_posts” is in the denormalized form as its two columns namely “upvotes” and “downvotes” violate the first normal form criteria as each cell in these columns contain multiple values. The better approach would be to create a separate table for storing the voting information.
3. There should be a foreign key constraint on the “post_id” column of the “bad_comments” table referencing the “id” column of the “bad_posts” table
4. The user should be referenced by its “user_id” in the newly created tables whenever required.
5. There is a transitive dependency between “id”, “title” and “topic” columns of the “bad_posts” table which can be removed for having the table in 3rd normal form.
6. Except columns declared as primary key, frequently accessed columns should be indexed for achieving the best query performance

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:

```
-- Following the given guidelines, we decide to create 5 tables in our new schema:
-- 1) "users" table
-- 2) "topics" table
-- 3) "posts" table
-- 4) "comments" table
-- 5) "votes" table

--------------------------------------------------------------------------------
-- Creating table for storing user information
CREATE TABLE "users" (
```

```
      "id" SERIAL PRIMARY KEY,
      "username" VARCHAR(25) UNIQUE NOT NULL,
      "last_login" TIMESTAMP,
      CONSTRAINT "check_usrname_len" CHECK (
      LENGTH( TRIM("username")) > 0
      )
);
-- Adding the index on "username" column for performant query
CREATE INDEX "find_users_by_usrname" ON "users" ("username");
-- Since unique constraint is case sensitive it allows "Shri1" and "shri1"
-- as valid usernames. To have case insensitive usernames, we put following
-- constraint:
CREATE UNIQUE INDEX "check_case_insensitive_username" ON
"users"(LOWER("username"));
--------------------------------------------------------------------------

-- Creating table for storing topics information
CREATE TABLE "topics" (
      "id" SERIAL PRIMARY KEY,
      "topic_name" VARCHAR(30) UNIQUE NOT NULL,
      "description" VARCHAR(500),
      -- put constraint for topic_name length > 0
      CONSTRAINT "check_tpname_len" CHECK (LENGTH( TRIM("topic_name")) > 0)
);
-- create index on topic_name column
CREATE INDEX "find_topic_by_tpname" ON "topics" ("topic_name");
-- in order to create a case insensitive topic_name put constraint
-- this will avoid "topic1" to be different from "Topic1".
CREATE UNIQUE INDEX "check_unique_tpname" ON "topics"(LOWER("topic_name"));

--------------------------------------------------------------------------
-- Create a table posts for storing the post info on existing topics
CREATE TABLE "posts" (
      "id" SERIAL PRIMARY KEY,
      "post_title" VARCHAR(100) NOT NULL,
      "topic_id" INTEGER NOT NULL, -- revised to "not null" as per feedback.
      "user_id" INTEGER, --user-id who posted it
      "url" VARCHAR(4000) DEFAULT NULL,
      "text_content" TEXT DEFAULT NULL,
      "posting_time" TIMESTAMP,
      -- constraints to be imposed
      -- 1) If a topic gets deleted, all the posts associated with it should be
      --    automatically deleted too.
      FOREIGN KEY ("topic_id") REFERENCES "topics"("id") ON DELETE CASCADE,
      -- 2) If the user who created the post gets deleted, then the post will
      --remain, but it will become dissociated from that user.
      FOREIGN KEY ("user_id") REFERENCES "users"("id") ON DELETE SET NULL,
```

```
	-- 3) Posts should contain either a URL or a text content, but not both.
	CONSTRAINT "url_or_content" CHECK (
	("url" IS NOT NULL AND "text_content" IS NULL) OR
	("url" IS NULL AND "text_content" IS NOT NULL)
	),
	-- 4) Title of the post can't be empty
	CONSTRAINT "check_post_title_len" CHECK (
	LENGTH( TRIM("post_title")) > 0
	)
);
-- Creating index for this table
CREATE INDEX "search_posts_by_user" ON "posts"("user_id");
CREATE INDEX "search_posts_by_topic" ON "posts"("topic_id");
CREATE INDEX "url_specific_posts" ON "posts"("url");
--------------------------------------------------------------------------------
-- creating the comments table
CREATE TABLE "comments" (
	"id" SERIAL PRIMARY KEY,
	"user_id" INTEGER,
	"post_id" INTEGER NOT NULL,
	"parent_comment_id" INTEGER DEFAULT NULL,
	"comment_text" TEXT NOT NULL,
	"comment_post_time" TIMESTAMP,

	-- putting constraints
	-- 1) comment should belong to the valid user
	FOREIGN KEY ("user_id") REFERENCES "users"("id") ON DELETE SET NULL,
	-- 2) comment should belong to the valid post id
	FOREIGN KEY ("post_id") REFERENCES "posts"("id") ON DELETE CASCADE,
	-- 3) comment should have a valid parent comment id
	FOREIGN KEY ("parent_comment_id") REFERENCES "comments"("id") ON DELETE
CASCADE,
	-- 4) constraint for "a comments text cant be empty"
	CONSTRAINT "comment_text_not_empty" CHECK (
	LENGTH( TRIM("comment_text")) > 0
	)
);

CREATE INDEX "find_comments_by_user" ON "comments"("user_id");
CREATE INDEX "find_comments_by_post" ON "comments"("post_id");
CREATE INDEX "find_comments_by_parent_id" ON "comments"("parent_comment_id");
--------------------------------------------------------------------------------
-------
-- creating the votes table
CREATE TABLE "votes" (
	"id" SERIAL PRIMARY KEY,
	"user_id" INTEGER,
```

```
        "post_id" INTEGER NOT NULL,
        "vote_value" SMALLINT NOT NULL,

        -- putting constraints
        -- 1) vote should be given by the valid user
        FOREIGN KEY ("user_id") REFERENCES "users"("id") ON DELETE SET NULL,
        -- 2) vote should be considered for the valid post id
        FOREIGN KEY ("post_id") REFERENCES "posts"("id") ON DELETE CASCADE,
        -- 3) constraint for "value of vote must be either 1 or -1"
        CONSTRAINT "check_valid_vote_value" CHECK (
        "vote_value" = 1 OR "vote_value" = -1
        )
        );
CREATE INDEX "find_votes_by_post" ON "votes"("post_id");

-----------------------------------------------------------------------------------
```

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:

```
-- Data Migration from bad_schema to our schema
-----------------------------------------------
-- 1) Migrating the users considering the following guideline
--   "Don't forget that some users only vote or comment, and haven't
created any posts.
--    You'll have to create those users too."

INSERT INTO "users" ("username") (
      SELECT DISTINCT "username" FROM "bad_posts"
      UNION
      SELECT DISTINCT "username" FROM "bad_comments"
      UNION
      SELECT DISTINCT REGEXP_SPLIT_TO_TABLE("upvotes", ',') FROM
"bad_posts"
      UNION
```

```
      SELECT DISTINCT REGEXP_SPLIT_TO_TABLE("downvotes", ',') FROM
"bad_posts"
);

--------------------------------------------------
-- 2) Migrating the topics from bad_schema
INSERT INTO "topics" ("topic_name") (
      SELECT DISTINCT "topic" FROM "bad_posts"
);

--------------------------------------------------
-- Migrating the posts from the bad_schema
INSERT INTO "posts"
("post_title","topic_id","user_id","url","text_content") (
      (
      SELECT SUBSTRING("b"."title",1,100), "t"."id", "u"."id",
      "b"."url","b"."text_content"
      FROM "bad_posts" "b"
      JOIN "topics" "t"
      ON "b"."topic"="t"."topic_name"
      JOIN "users" "u"
      ON "b"."username"="u"."username"
      )
);
--------------------------------------------------
-- Migrating the comments from the bad_schema
INSERT INTO "comments" ("user_id","post_id","comment_text") (
      (
      SELECT  "u"."id", "p"."id", "b"."text_content"
      FROM "bad_comments" "b"
      JOIN "users" "u"
      ON "b"."username"="u"."username"
      JOIN "posts" "p"
      ON "b"."post_id"="p"."id"
      )
);
--------------------------------------------------
--Migrating the votes (upvotes)
INSERT INTO "votes" ("user_id","post_id","vote_value") (
      SELECT "u"."id", "b"."id", 1 AS "up_vote"
      FROM (
      SELECT "id", REGEXP_SPLIT_TO_TABLE("upvotes",',') AS "upvotes"
```

```
      FROM "bad_posts"
      ) "b"
      JOIN "users" "u"
      ON "b"."upvotes"="u"."username"
);
--Migrating the votes (downvotes)
INSERT INTO "votes" ("user_id","post_id","vote_value") (
      SELECT "u"."id", "b"."id", -1 AS "down_vote"
      FROM (
      SELECT "id", REGEXP_SPLIT_TO_TABLE("downvotes",',') AS "downvotes"
      FROM "bad_posts"
      ) "b"
      JOIN "users" "u"
      ON "b"."downvotes"="u"."username"
);
-----------------------------------------------------
```