EECS 341 Final Report

Jiawei Wu (jxw585), John Donnelly (jed126), Billy Barbaro (wxb107)

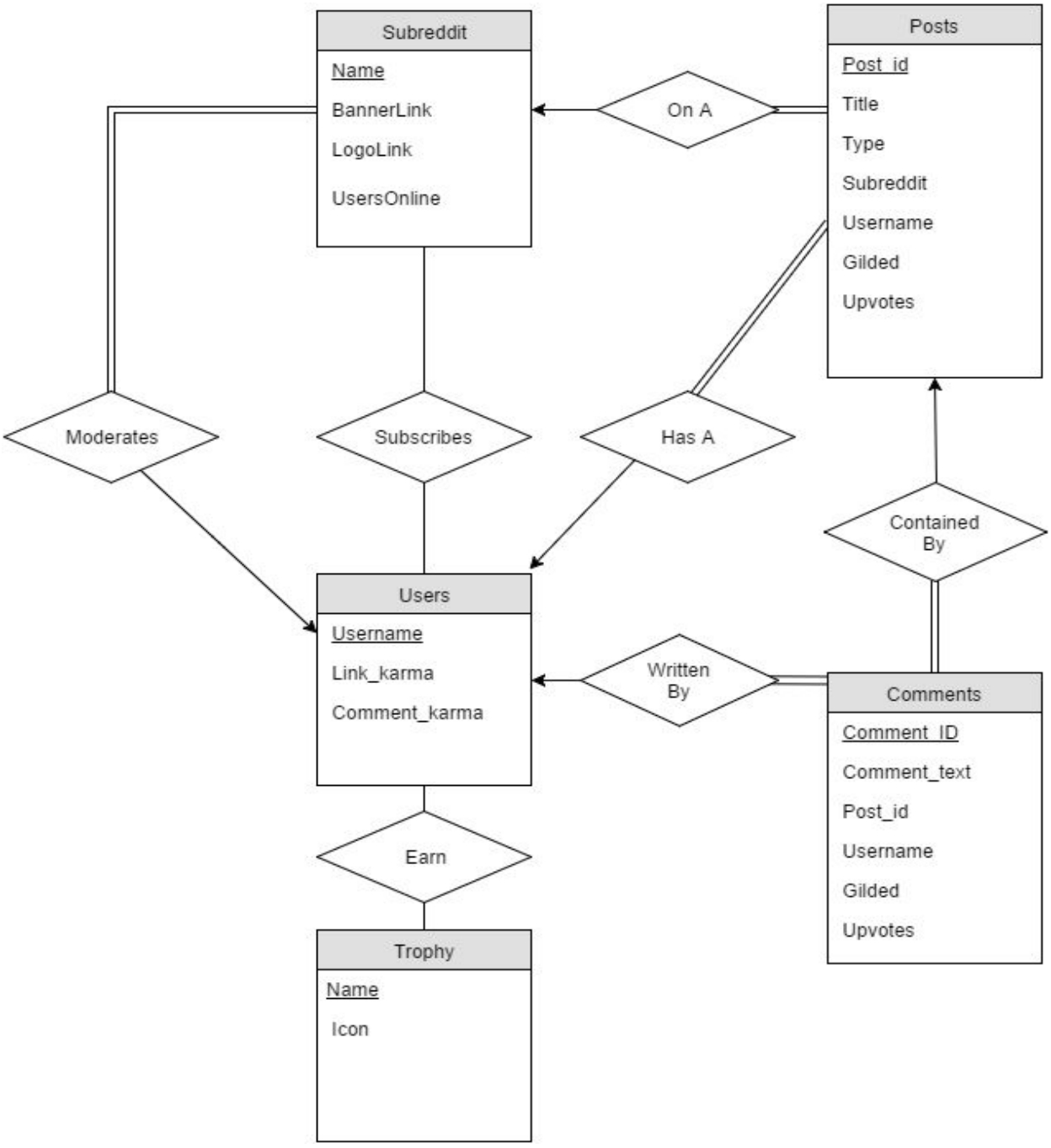
Application Background:

Almost every website that supports user posting, commenting, voting, etc. leverages some form of database application to persist and present this data to users. In our project, we aim to mimic the databases used by the popular website reddit. Reddit is essentially a giant forum, covering nearly every conceivable topic. reddit is made up of "*subreddits*" which are communities of *users* who *subscribe* to the subreddits to discuss common interests using *posts* and *commenting* on those posts. Users vote on posts to decide which posts will be on the "front page" easily visible to everyone. Each of these subreddits have *moderators* who choose the rules for the subreddit and enforce them. Additionally, users can receive *trophies* for various activities and *gildings* for exceptional comments and posts.

Data Description:

- Every subreddit is uniquely identified by its name. It may also have a link to the images used for its logo and banner in addition to a count of the number of users currently online.
- Every user is uniquely identified by their username.
- Users have both link karma and comment karma which are always integers. The link karma denotes the number of upvotes a user's post has received. The comment karma denotes the upvote count on their comments.
- Users do not need to subscribe to any subreddits, post, or comment. They are free to just have an account and view the site's content.
- Every subreddit must have one and only one moderator
- A user can be the moderator to multiple subreddits.
- Users can subscribe to as many subreddits as they want and a subreddit can have unlimited subscribers.
- A user can receive any number of trophies which have a name and an icon associated with them for accomplishments (like 1,000 upvotes on a post, or being a member for a year).
- Every post has a title, type, number of upvotes, and it can be stickied or gilded (both boolean values).
- Every post belongs to a subreddit, but subreddits can have multiple posts.
- Every post is posted by a single user.
- Comments on a post consist of the comment text, the number of upvotes on the comment, and whether or not the comment has been gilded.
- Each comment belongs to only 1 post (although a post can have many comments).
- Each comment is posted by a single user (although a user can post multiple comments).

ER Diagram:


Subreddit
Name
BannerLink
LogoLink
UsersOnline
On A
Posts
Post_id
Title
Type
Subreddit
Username
Gilded
Upvotes
Moderates
Subscribes
Has A
Contained By
Users
Username
Link_karma
Comment_karma
Written By
Comments
Comment_ID
Comment_text
Post_id
Username
Gilded
Upvotes
Earn
Trophy
Name
Icon

Functional Dependencies:

For Trophy:
Name -> Icon

For Comments:
Comment_ID -> comment_text, Username, Gilded, Upvotes, Post_id

For Users:
Username -> Link_Karma, Comment_Karma

For Posts:
Post_id -> Title, Type, Subreddit, Username, Gilded, Upvotes

For Subreddit:
Name -> BannerLink, LogoLink, UsersOnline

Schema:

Subreddit(Name, BannerLink, LogoLink, UserOnline)
Users(Username, Link_Karma, Comment_Karma)
Trophy(name, icon)
Posts(Post_id, title, type, subreddit, username, gilded, upvotes)
Comments(Comment_ID, Comment_text, post_id, Username, Gilded, Upvotes)
Earn(Username, TrophyName)
Subscribes(Username, Subreddit)

Example Queries:

- Select all users subscribed to the science subreddit
  - SELECT username FROM subscribes WHERE name ='science';
- How many posts from the pics subreddit are from Imgur?
  - SELECT COUNT(*) FROM posts WHERE subreddit = "pics" AND (type='imgur.com' OR type='i.imgur.com');
- Average numbers of users online across all subreddits.
  - SELECT AVG(usersonline) FROM subreddit;
- Select the subreddit with most combined upvotes for its posts
  - SELECT sr FROM
    ((SELECT subreddit AS sr, (SUM(upvotes)) AS cnt FROM (Subreddit NATURAL JOIN Post) GROUP BY subreddit) AS newtable)
    WHERE cnt =
    (SELECT MAX(cnt) FROM((SELECT subreddit AS sr, (SUM(upvotes)) AS cnt FROM (Subreddit NATURAL JOIN Post) GROUP BY subreddit) AS newtable));

- Select all user subscribed to the videos subreddit with more than 500 link karma:
  - SELECT s.username FROM subscribes s NATURAL JOIN users u WHERE u.link_karma >= 500;
- Select the user with the most trophies.
  - SELECT E1.username FROM earn E1 GROUP BY E1.username HAVING Count(*) >= ALL(SELECT COUNT(*) FROM earn E2 GROUP BY E2.username);

Implementation:

In order to make use of real data for our project, we leveraged the REST API reddit provides to users (more information available at https://www.reddit.com/dev/api). Using a python wrapper on the API (https://github.com/praw-dev/praw), a python script was developed (included shortparser2.py in addition to parallelization.py which helps with the trophy data) to pull down the data necessary to populate our tables. The gathered data was then parsed and formatted to fit into the schema we had set up, then saved out as a CSV. To save us the trouble of having to go through and write each of the INSERT statements for the data, the online tool at https://sqlizer.io was used, which takes in a CSV and outputs SQL insert statements corresponding to each row. While this work could have been done in python using an ORM such as SQLAlchemy, for the sake of simplicity and time, we decided against it.

Our database implementation was done using MySQL. An AWS (Amazon Web Services) server was set up to run an instance of our database. This remote instance allowed the entire team to view and work on the same data. The attached script Create_DB1.2.sql was run to create the schema. Then the contents of AllInserts.sql (generated using the process described above) were run to populate the database.

Team Roles:

Wu
- Wrote scripts to take the data off of reddit and convert it to usable data
- Helped with presentation and report
- Various other minor roles

Donnelly
- Designed ER diagram
- Wrote database schema
- Helped with presentation and report
- Various other minor roles

Barbaro
- Wrote SQL to create tables
- Helped with presentation and report
- Various other minor roles