Spring 2023 CS307 Project Part1

> Main Contributors:
>
> Leader and Overall Design: ZHU Yueming
>
> Data Preparation and Documentation: WANG Lishuang, ZHANG Chaozu
>
> Review: MA Yuxin
>
> Extended from the project of Spring 2022

General Requirement:

- It is a group project with **only 2 teammates** who are **in the same lab session**. Each group should finish the project independently and submit only one report written by the teammates.
  - The teammate you select for Project 1 will also be your teammate for Project 2. It is not allowed to change teammates once paired.
- You should submit the report before the deadline. All late submissions after the deadline will receive a score of zero.
- DO NOT copy ANY sentences and figures from the Internet and your classmates. Plagiarism is strictly prohibited in this course.
- The text description should be rigorous, the overall design should be logical organised, the report structure and the layout of diagram should be clear and easy to read, otherwise, you will receive a penalty in the scoring stage.
- The number of pages for your report should be between **4** and **8**. Reports **only or less than 3 pages** and **more than 8 pages** will receive a penalty in the scoring stage.

DBMS can help us manage data in a convenient manner and improve the efficiency of data retrieval. Your work of Project 1 is mainly divided into three parts below:

1. Design an E-R diagram based on the provided data file and data relationships.
2. Design a relational database using PostgreSQL according to the provided data file.
3. Import all data into the database.

Background

Data Description

`posts.json`

> **Post ID**: The id of post, **unique**
>
> **Title**: The title of post, like: `The Benefits of Running in the Morning`
>
> **Category**: The category of post, like: `["Fitness", "Running", "Health"]`

Content: The content of post

Posting Time：The time of posting. *Posting Time is later then the author registration time*

Posting City：The city of posting.*The city is randomly generated and it has no relation to the location of author.*

Author：The Author of post, **unique**

Author Registration Time：The registration time of author.

Author's ID： ID of author. *The ID verification code is valid, but does not verify that the birth time is earlier than the registration time*

Author's Phone：phone of author

Authors Followed By：Accounts who followed by author. *The authors in this field may not appear in* `Author` *field.* `follower account`

Authors Who Favorited the Post：Accounts who favorited the post.*The authors in this field may not appear in* `Author` *field.* `favorite account`

Authors Who Shared the Post：Accounts who shared the post.*The authors in this field may not appear in* `Author` *field.* `sharer account`

Authors Who Liked the Post：Accounts who like the post.*The authors in this field may not appear in* `Author` *field.* `liker account`

`replies.json`

Post ID: The id of post in post.josn, **unique**

Reply Content：the content of reply

Reply Stars：the star of reply

Reply Author：the author of reply.*The authors in this field may not appear in* `Author` *field.*

Secondary Reply Content：the content of sub reply

Secondary Reply Stars：the star count of sub reply

Secondary Reply Author：the author of sub reply. *The authors in this field may not appear in* `Author` *field.*

The Report and your Tasks

Basic Information of Your Group

1. Names, student IDs, and the lab session of the group members
2. You are required to write down the contributions and the percentages of contributions for each group member. **Please clearly state which task(s)/part of the task(s) is/are done by which member in the group.**
    - **If you failed to link a task/part of a task to one of the group members, we will not**

count the score for the task (since we don't know who accomplished this task; maybe it was done by an elf while you were sleeping at night?).

Task 1: E-R Diagram (30% in total)

Make an E-R Diagram of your database design with any diagram software. Hand-drawn results will not be accepted. Please follow the standard of E-R diagrams.

In the report, you are required to provide a snapshot of the E-R diagram. Also, please specify the name of the software/online service you use for drawing the diagram.

Task 2: Relational Database Design (40% in total)

Design the tables and columns based on the background provided above. Generate the E-R diagram via the "Show Visualization" feature. Briefly describe the design of the tables and columns including (but not limited to) the meanings of tables and columns.

In the report, you are required to provide the following content:

1. Attach the snapshot of the E-R diagram generated by DataGrip.
2. Briefly describe the table designs and the meanings of each table and column.

In addition, please submit an SQL file as an attachment that contains the DDLs (`create table` statements) for all the tables you created. **Please make it into a separate file but not copy and paste the statements into the report.**

Notes for the database design:

1. All data items should base on two files `posts.json` and `replies.json` .
2. Your design needs to follow the requirements of the three normal forms
3. Use a primary key and foreign keys to indicate important attributes and relationships about your data
4. Every row in each table should be uniquely identified by its primary key. (You may use a simple or a composite primary key).
5. Every table should be involved in a foreign key. No isolated table is allowed. (每个表要有外键，或者有其他表的外键指向。)
6. Your design should contain no circular foreign-key links. (对于表之间的外键方向，不能有环。例如：A表有外键关联B表，B表有外健关联C表，C表有外健关联A表)
7. Each table should contain at least one mandatory (“Not Null”) column (including the primary key but not the id column).
8. Other than the system-generated self-increment ID column, there should be at least one column with the "unique" constraint. (除了主键自增的id之外，需要有其他unique约束的列)
9. You should use appropriate data types for different fields.
10. Your design should be easy to expand when requirements change.

Task 3: Data Import (30% in total)

In this task, you should write scripts to import the content in those two json files into the database you have designed before. After importing the data, you should also make sure all data is successfully imported.

Task 3.1 Basic Requirements: 10%

1. The script you wrote to import the data file.
2. A description of how you use the script to import data. You should clearly state the steps, necessary prerequisites, and cautions in order to run the script and import data correctly.

Task 3.2 Advanced requirements: 10%

You may also need to finish the following advanced requirement to get the remaining points (10%):

1. Try to **optimize your script**, and find **more than one ways** to import data, and provide a comparative analysis of the computational **efficiencies** between these ways.

For the advanced points, please make sure to describe your test environment, procedures, and actual time costs. It is required to write a paragraph or two to analyze the experiment results.

Task 3.3 Data Accuracy checking: 10%

We will check this part in lab course in April 25th and April 27th.

1. According to the data in two files `posts.json` and `replies.json`, we will **give several questions in lab course** to check whether all data have been correctly imported into your database.
2. If the **author name** appears in following fields but not appears in `Author` field, you need generate `author id` and a resonable `registration time` for the author account, and add it into your database.
    - Authors Followed By
    - Authors Who Favorited the Post
    - Authors Who Shared the Post
    - Authors Who Liked the Post
    - Reply Author
    - Secondary Reply Author

How to Submit Your Report

Submit the report in PDF format with necessary attachments (such as SQL scripts and source code files) on the Sakai website before **23:30 on April 24th, 2023, Beijing Time (UTC+8)**. For attachments, please put them into separate directories based on the task, and compress them into a `.zip` archive.

Disclaimer

The names, characters, businesses, and events in the background of this project are purely fictional. The items in the files are generated by chatGPT. Any resemblance to actual events, entities or persons is entirely coincidental and should not be interpreted as views or implications of the teaching group of CS307.