Assignment 1: Database Creation & Loading

CSE 511: Data Processing at Scale

Points: 100

Introduction & Background

In this assignment you will learn how to deal with loading large datasets into a database from scratch and then build applications on the top of this database as we move forward with the course. For this assignment we will be using a research based Reddit database.

Numerous scientific research has centered on Reddit in particular and is also thought of as the self-described "first page of the Internet." Technical impediments to the data acquisition still exist, despite Reddit being more open to data collection than other social media networks like Facebook and Twitter. As a result, collecting and carefully analyzing the billions of comments, millions of subreddits, and hundreds of millions of users on Reddit is time consuming. The Reddit Pushshift dataset is a result of this time consuming work and has been gathered and made available to researchers by Pushshift, a social media data gathering, analysis, and archiving tool. The Reddit dataset from Pushshift contains historical data going all the way back to Reddit's founding and is updated in real-time. This dataset has been extensively used for educational and research purposes.

Dataset Description

All of the submissions and comments posted on Reddit between June 2005 and April 2019 are accessible through Pushshift. The original dataset comprises **2,611,778,198 submissions 5,601,331,385 comments, and 2,888,885 subreddits.**

Important Note:

1. In the scope of this assignment and future assignments we will only use a portion of the Pushshift dataset which we have carefully curated for the students. The original dataset was stripped down, decreasing the number of columns in the dataset such that it is easier to manage. You can find the stripped down files which you **MUST** use to complete the assignment here**.** Also the sample entries of the multiple dataset files only highlight the available columns after the dataset is stripped and cleaned. This is NOT the original dataset. **Students MUST use only the link provided to the specific files of the dataset for the assignment and NOT the original dataset.**

2. The sample entries provided with each file is only for understanding of the columns in the dataset. Due to space limitation, some of the textual data may be cut down.

The stripped down Pushshift Reddit dataset is made up of multiple files and for ease of access we have created .csv for the same.

1. **submissions:**
    a. Below is a sample entry form “submission”, showing all the available columns and their sample values.
    b. Link to the submissions file. Total number of entries: **1,263,936**


```
"downs": 1,
"url": http://thinkprogress.org/economy/2011/06/29/some_url",
"id": "iemqy",
"edited": "False",
"num_reports": "",
"created_utc": 1309564792,
"name": "t3_iemqy",
"title": "A somewhat long title with some political discussion",
"author": "[deleted]",
"permalink": "/r/politics/comments/iemqy/another_string/",
"num_comments": 1,
"likes": "",
"subreddit_id": "t5_2cneq",
"ups": 3
```


2. **comments:**
    a. Below is a sample entry form “comments”, showing all the available columns and their sample values.
    b. Link to the comments file. Total number of entries: **10,557,466**


```
"distinguished": "",
"downs": 0,
"created_utc": 1309478400,
"controversiality": 0,
"edited": "False",
"gilded": 0,
"author_flair_css_class": "mordekaiser",
"id": "c22x4aq",
"author": "username",
"retrieved_on": 1427302516,
"score_hidden": "False",
"subreddit_id": "t5_2rfxx",
"score": 1,
"name": "t1_c22x4aq",
```

```
"author_flair_text": "[username] (NA)",
"link_id": "t3_id1nc",
"archived": "True",
"ups": 1,
"parent_id": "t3_id1nc",
"subreddit": "leagueoflegends",
"body": "Good lord.  Yes."
```


3. **authors:**
    a. Below is a sample entry form “authors”, showing all the available columns and their sample values.
    b. Link to the author file Total number of entries: **6,158,212**


```
"id": "t2_1rr1",
"retrieved_on": 1532086586,
"name": "duncan",
"created_utc": 1298437200,
"link_karma": 1029,
"comment_karma": 9943,
"profile_img": "https://www.redditstatic.com/avatars/avatar_default_02_25B79F.png",
"profile_color": "",
"profile_over_18": "False"
```


4. **subreddits**:
    a. Below is a sample entry form “subreddit”, showing all the available columns and their sample values.
    b. Link to the subreddit file Total number of entries: **914,067**


```
"banner_background_image": "",
"created_utc": 1137700161,
"description": "A very very long description of what the subreddit is about. SHown maybe only to the members?",
"display_name": "John Doe",
"header_img": "https://b.thumbs.redditmedia.com/h5RmvyztneDL1.png",
"hide_ads": "False",
"id": "vf2",
"over_18": "True",
"public_description": "Still a description but this one shown to people not part of the subreddit",
"retrieved_utc": 1591839904,
"name": "t5_vf2",
"subreddit_type": "public",
"subscribers": 1880887,
"title": "Another Random Subreddit",
"whitelist_status": "all"
```

Please find the corresponding diagram showing the relations across the tables for the stipped down Pushshift dataset.

Problem Statement

- Considering the **four** tables **submissions, comments, author and subreddit** your task is to create tables and load the corresponding data to the table.
- You need to figure out the primary keys, foreign keys, constraints or other necessary settings by yourself. The key information in the requirement is not complete and attributes can be primary keys and foreign keys at the same time.
- All table names and attribute names **must** be in lowercase letters and exactly the same as mentioned.
- Your code should **NOT** contain the commands to **create a database, change database or set encoding!**

Grading

- The assignment will be graded using automated scripts.
- 100 points of the assignment is divided into two categories
    - **80%** for Database creation, normal insertion (i.e. without any optimization) and correctly assigning constraints.
    - **20%** for the optimized data insertion.
    - We have optimized the code using pg_bulkload, but you are free to also explore any other optimization techniques if interested. Please find some references below on how the data insertion can be optimized.
        - dbi Blog (dbi-services.com)
        - http://ossc-db.github.io/pg_bulkload/pg_bulkload.html
        - https://www.postgresql.org/docs/current/populate.html
        - sql - How to speed up insertion performance in PostgreSQL - Stack Overflow
        - 4.pg_bulkload Data Loading Use and Example - www.cqdba.cn - Blog Park (cnblogs.com)

          Note: The threshold to get full 20% grade points in the optimization test case is **~300 seconds** to insert all the entries. For example, Our optimized code performs the entire problem statement in ~**220 seconds**.

- You **MUST** provide **assignment1.sh** which will be responsible for the following:
    - We have optimized the code using pg_bulkload, but you are free to also explore any other optimization techniques if interested. In case you choose to use any additional program, your .sh script **MUST** install any third party and necessary libraries which you are using. The grading will be automated hence we won't install any other dependencies explicitly to grade, you MUST mention it as a part of your .sh script.
    - The assignment1.sh **MUST** call any required .sql file(s) which should **NOT** contain the commands to create a database, change database or set encoding. Please test your code, you will receive 0 grade points if your code crashes the grading environment!
    - The automated grading scripts will **only** run the .sh file from your submissions to test your code.
- You can test your code with a subset of the data but the grading scripts will use the entire mentioned dataset.

Submission Requirements & Guidelines

Assignment 1 is due on **01/25/2023 11:59 PM**. Submit the assignment following the below guidelines

1. This is an **individual** assignment
2. Naming nomenclature:
    a. You **MUST** name your .sh file as **assignment1.sh**
    b. Name the zip file following the naming convention. **"Assignment-1.zip"** to submit on canvas including **both .sh and .sql files**.
    c. README.md incase you want to mention anything to the grading team.

Submission Policies

1. Late submissions will *absolutely not* be graded (unless you have verifiable proof of emergency). It is much better to submit partial work on time and get partial credit for your work than to submit late for no credit.
2. Every student needs to *work independently* on this exercise. We encourage high-level discussions among students to help each other understand the concepts and principles. However, a code-level discussion is prohibited and plagiarism will directly lead to failure of this course. We will use anti-plagiarism tools to detect violations of this policy.

Common Questions

1. Does the time constraint mentioned in the assignment (300 seconds) include the time required to install the libraries or just the time required for the optimized insertion (20%)?
   The 300 seconds include table insertion, data insertion and constraint creation only.

2. Are commands for installing PostgreSQL/pg_bulkload to be written in the .sh file?
   The grading machines will have the following tools already set up:
   (1) psql
   (2) pg_bulkload

   Additionally, the scripts will be run after being logged into the postgres user. You do not need to write the switch user command. However, you will have to write the command to correctly access the correct database from the correct database user

3. What path to have for the csv files?
   The same folder as your .sh script: ./filename.csv

4. What are the details of the database?
   PostgreSQL v14 Configurations:
   - username: postgres
   - password: postgres
   - Database Name: postgres
   - Database IP: 127.0.0.1:5432

   You do not need to use all the details, only the ones that are necessary

~ The crux of the assignment is to gain the ability to ingest a (somewhat) large amount of data and to understand how that is performed! Make sure that you focus your time on that aspect ~