Mahek Aggarwal, Vivian Li

DS 4300

January 22nd, 2024

Twitter in a RDB

In this assignment, we utilized MySQL 8.0.36 as our relational database to store the tweets. We created a connection to the MySQL database through Python using the MySQL Connector and configured sensitive information using the getpass library so users are able to input their personal details. In addition, the pandas/time/random/mysql.connector libraries were used. The hardware configurations included an Apple MacBook Pro with a M3 processor chip with 16GB of RAM and 8 cores.

Our results:

Number of tweets processed in one second:	7041.33
Home timelines retrieved per second:	96.3

We used multiple functions to interact with the database to collect information. Initially, we had approximately 2000 tweets in one second with approximately 300 home timelines. After re-running the codes several times, we witnessed that the number of tweets and home timelines retrieved increased gradually. We are unsure of the cause of this; however, we have some hypotheses. Our main hypothesis is that depending on the random user selected, the home timelines retrieval time may differ based on the followings of that user. In addition, the size of the tweets file may cause delays in processing. Some possible improvements for the future include implementing secondary indexing so that data is retrieved more efficiently from the database.

Both team members contributed equally to this project as we read and understood the assignment together and then collaboratively set up the database in MySQL. We each worked on one of the functions and assisted each other when we ran into any bugs.