Merai Dandouch

Graph Database Design Intern - Summer 2022

The BU Neuromics Lab is working towards building a database that stores a myriad of data types such as genomic modalities, clinical data and subject information from the PTSD Brain Bank repository. The PTSD Brain Bank is led by Veterans Affair National Center for PTSD which stores brain tissue samples from people who had PTSD and experienced traumatizing events. It's also integrated with VA's Boston-based brain banks that focuses on Alzheimer's disease (Friedman et al., 2017). Their goal is to answer the question of how stress impacts brain functioning.

To help researchers answer this question, a database will be used to process and retrieve clinical and high-throughput sequencing data. A major problem lies in the fact that PTSDBB data is multidimensional that perhaps a conventional relational database may not be adept at connecting and storing this data. I worked with the Neuromics Lab to help build a database infrastructure to store genomic and clinical information using a Neo4j database platform.

Neo4j is a NoSQL, native graph database that is an ACID transactional utilizing index free adjacency and the cypher query language. Neo4j is useful in storing data that contains numerous layers and dimensions. Anytime there is a change in the database, it will not be fully committed until all the transactions are successfully completed. This means Neo4j allows for more efficiency regarding querying and traversal of data by using the 'pointer hopping' method. It stores the address of every record node in RAM and lookups adjacent nodes and hops to it using a pointer. Neo4j does not need to rely on indexes, although it can. While this can be computationally effective, "pointer-hopping" consumes a lot of RAM and disk space.

A recurring error I encountered when trying to load the data in was a memory error. For example, I was given a practice GTEx RNA-Seq dataset (56382 genes x 8555 samples) in csv format around 1GB in size. It was impossible to load the data in one transaction. To overcome this challenge, I partitioned the dataset into 8555 separate files where each file represented a sample and its corresponding gene symbol and raw read count. I iterated through each file as a separate transaction and loaded the file into the database. After each file was processed, the transaction result was then removed from memory. This method proved to be successful and consumed less ram but took up more disk space overall. For example, partitioning the dataset from 1 file into many separate files, increased the file size from 1GB to 14 GB. Lastly, for every node and relationship entity stored in the database, the database disk size increased by 15 - 41 bytes due to the data being stored as a linked list. After successfully loading the data into database, the database took up 36 GB in disk.

The clinical data contained participant information including therapeutics, bmi, sex, age, environment, smoking history, childhood etc. This data will be stored in a separate database than the one containing genomic information. I discovered that there was a paywall behind the multiple database functionality in Neo4j and due to time constraints, the clinical data remains to be stored in the second database. Nevertheless, the clinical data loading script was successful in storing this information into a Neo4j database and did not result in a memory error.

I worked on a side project which involved adding a front side web framework to the database. This included a HTML form that takes in user input to query the PTSDBB database and a back-end FLASK API application which communicates with the server and retrieval process.

Through this experience, I was able to learn a new database and back-end API language and solve the many Neo4j errors I encountered. More importantly, I gained autonomy, independence, confidence, and acceptance with failing and learning. Although there is still a lot to learn about database architecture and functionality, I have a newly founded appreciation and passion for databases.

Works Cited

Friedman, M.J., Huber, B.R., Brady, C.B., Ursano, R.J., Benedek, D.M., Kowall, N.W. and McKee, A.C., 2017. VA's National PTSD Brain Bank: a national resource for research. *Current Psychiatry Reports*, *19*(10), pp.1-8.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 08/2022 and/or dbGaP accession number phs000424.v6.p1 on 08 /2022.