Programming Assignment #2.2 CAS CS 460

Row-stores vs. Column-stores

Due: 12/07 11:59 pm on gradescope.

This programming assignment is for groups of two. If there is a strong reason you wish to work on it alone, please reach out to the teaching staff and explain why.

Task 2

2.1 Introduction

A common question in production is which system to use for a specific use-case. A good data engineer can provide such answers through experience, benchmarking, and intuition. The goal of this task is to start building these skills, starting with benchmarking.

Traditional DBMS architectures today follow two main approaches: a row-*major* and *column-major* approach. In this task you will study these approaches. We will use PostgreSQL as a row-major system and MonetDB as a column-major system. The goal of the project is to compare the performance of these two systems for a set of analytical queries taken from an industry-grade database systems benchmark.

2.2 Set-up

Create a new database in PostgreSQL named 'task2database_psql'. Do the same for MonetDB and create a database named 'task2database_monet'. You may follow instructions from the Azure Support Manual to do so. Then you will need to load data in those two databases. To do that, download this folder from here. Import this file from your local machine to the azure virtual machine like we did in **Task 1**. Remember, to change the folder name of the newly downloaded file in the older command before importing.

Follow these steps after importing:

1. Create a file that will setup the monetdb environment variables:
    a. Create a file called **.monetdb** using this command:
    ```
    touch .monetdb
    ```
    b. Using a text editor of your choice (**nano or vim**), enter the following contents into this file:
    ```
    user=monetdb
    password=monetdb
    ```
    c. Save the file and exit the text editor.
2. The, let us go inside the pa2-task2 directory. If you have imported the folder correctly, you can use this command to do so:
    ```
    cd ./pa2-task2/
    ```

3. Then, navigate to the **/dbgen** folder using this command:
   ```
   cd ./dbgen/
   ```
4. In this folder run the "**`make clean`**" and then "**`make`**" commands. If you are not familiar with what make is or what it does, you can learn by navigating to the resources section of this part.
5. After that, you should run "dbgen" executable by just typing this command:
   ```
   ./dbgen
   ```
6. Then, run the **setup.sh** shell script which will generate the contents of the **/data** and **/queries** folders, and will also load the data in your running database systems.
   a. First we will navigate out of the dbgen folder using:
      ```
      cd ../
      ```
   b. We will now run **setup.sh** using:
      ```
      ./setup.sh
      ```

The required data will be setup in both the monetdb and postgres databases automatically through the setup.sh script. It will take a few minutes to complete setup.

2.3 Exercise

You will have to prepare a document where the two systems will be compared for 3 different queries from the TPC-H benchmark. You should execute query 1, 4 and 6. In the submission document, you have to present their performance (query latency).

- You must run each query 5 times and then calculate the average latency and report the standard deviation in a tabular format.
- The reported performance should be accompanied with experimental setup, any tuning done to the system, and information with regards to standard deviation.
- **Pick one tuning knob from Postgres and explain why tuning that particular knob might help improve performance.** You may view tuning knobs using the "**`show all;`**" command within the postgres terminal (after using **`psql postgres`** from the terminal of your VM).
- Finally, the report should discuss which of the two systems is preferable for what type of queries based on the observed results and the intuition developed throughout the experimentation with the systems.

2.4 Resources

MonetDB: basic tutorial, server connection, for windows users
"Make" documentation
TPC-H Benchmark Spec File (go over Chapters 0, 1, and 2)

Submission instructions:

Please submit a single PDF file in gradescope. Handwritten report is NOT allowed!
Make sure that you submit as a group! In this file include your full names and BU IDs. There should be only one submission per group.