

Project 1 CAS CS 561

Row-stores vs. Column-stores

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

This programming assignment is also to be completed on Azure. If there is a strong reason you wish to work on it on a local machine, please reach out to the teaching staff for approval.

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.

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 5, 8 and 20. In the submission document, you have to present their performance (query latency).

1. You must run each query 10 times and then calculate the average latency and report the standard deviation in a tabular format. For each trial, run every query once before repeating them, i.e. In trial 1 – run query 5, then query 8, then query 20, and then repeat.
2. Run the following aggregate queries in **both Postgres and MonetDB**:

- a. `select sum(l_quantity) from lineitem;`
- b. `select sum(l_quantity), max(l_discount) from lineitem;`
- c. `select sum(l_quantity), max(l_discount), max(l_tax) from lineitem;`

Analyze how does the performance of both systems vary when executing each of the above queries. Explain why we see a pattern (if any).



3. The reported performance should be accompanied with experimental setup, any tuning done to the system, and information with regards to standard deviation.
4. **Pick one tuning knob from Postgres and explain why tuning that particular knob might help improve performance.**
5. 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.