CAS CS 561: Data Systems Architectures
Data-intensive Systems and Computing Lab
Department of Computer Science
College of Arts and Sciences, Boston University
http://bu-disc.github.io/CS561/

CS561 Spring 2023 - Research Project

Title: *Measuring the Robustness of Modern Key-Value Stores*

Background: Modern key-value databases are commonly tuned using fixed deterministic workloads. While this may work well at first glance, if either the workload used for tuning and the observed workload differ by a wide margin, or we experience workload drift over time, there is a chance that an optimal tuning may experience a large drop in performance.

Recent work has shown that when accounting for workload uncertainty in the tuning process, we can improve performance consistency under changing workloads [1,2]. One of the first steps in this line of work is developing a technique to numerically describe workloads that are closer and farther away. Depending on the type of system different descriptors of workload distance are used, however, both techniques utilize this distance to test the robustness of databases. They do so by first tuning the DB for a single fixed workload, then evaluating this configuration on workloads that are further and further away. Therefore, this project will first focus on defining a distance, then using this to measure the robustness of a particular database.

Objective: In this project, we will develop a way to investigate and measure the robustness of databases. We will be focusing on taking a look at a relatively new key-value store called **SplinterDB** developed by VMWare that is based on a B$-\varepsilon$ tree [3].

(a) Read the two reference papers to get an idea of robustness and its definition.
(b) Define both a workload and a distance measure to differentiate between workloads that are close and far.
(c) Build a test bench suite to run arbitrary workloads on SplinterDB.
(d) Incorporate this distance measure into our test bench suite.
(e) Hand-tune SplinterDB for a particular workload. As SplinterDB is relatively new there are not many configurations to look at afterward, we can work together to make sense of what configuration is optimal.
(f) Evaluate this configuration on workloads that are further away to measure the robustness of a B$-\varepsilon$ tree.

Responsible Mentor: *Andy Huynh (ndhuynh@bu.edu)*

References:
[1] Huynh, A., Chaudhari, H. A., Terzi, E., & Athanassoulis, M. Endure: A Robust Tuning Paradigm for LSM Trees Under Workload Uncertainty. *Proceedings of the VLDB Endowment, 15(8),* VLDB, 2022.
[2] Barzan Mozafari, Eugene Zhen Ye Goh, and Dong Young Yoon. CliffGuard: A Principled Framework for Finding Robust Database Designs. *ACM SIGMOD International Conference on Management of Data*, 2015.
[3] https://github.com/vmware/splinterdb