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: Evaluating Sorting Algorithms with Varying Data Sortedness

Background: While many sorting algorithms exist that can establish total order to a data collection/stream, adaptive sorting algorithms are particularly interesting due to their ability to optimize for performance if data is already pre-ordered. One of the most widely known adaptive sorting algorithms is **Insertion Sort**. Insertion sort has a best-case time complexity of O(n) time when the data collection/stream is already fully sorted.

Objective: The goal of this project is to explore different sorting algorithms and their performance when varying the sortedness in the input data stream. The following steps offer a high-level overview of the required effort:

(a) Study various adaptive sorting algorithms available in literature. Particularly, focus on TimSort, (K, L)-adaptive sorting algorithm, Quicksort, MergeSort, Insertion Sort, and std::stable_sort (from C++ STL container). Students are encouraged to add at least another couple of algorithms to this list.
(b) Implement an API that can select and execute an algorithm from the above list for a given data stream/collection.
(c) Provide a comprehensive analysis of the performance of the above algorithms while varying the data sortedness of the input stream. Differently sorted data can be generated using the workload generator included with the BoDS Benchmark.
(d) Explore the possibility of a new hybrid sorting algorithm that combines radix sort and the K,L-sorting algorithm.

Responsible Mentor: *Aneesh Raman (aneeshr@bu.edu)*

References

[1] Ben-Moshe, S., Kanza, Y., Fischer, E., Matsliah, A., Fischer, M., Staelin, C.: Detecting and Exploiting Near-Sortedness for Efficient Relational Query Evaluation.
In: Proceedings of the International Conference on Database Theory (ICDT). pp. 256–267 (2011), http://doi.acm.org/10.1145/1938551.1938584