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: Exploring the Performance of Data Compression Algorithms with Varying Data Sortedness

Background: Data compression is becoming increasingly important in today's data systems with growing rate of data generation. Data compression is particularly useful to reduce the data size of large datasets for easier transfer between devices, or even archival of legacy data.

Objective: This project aims to explore different data compression 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 data compression algorithms available in literature. Huffman encoding, arithmetic encoding, run-length encoding, Lempel-Ziv, X-match, Frequent Value Compression, BDI compression and C-Pack are a few examples of compression algorithms that are interesting to explore [1].

(b) Classify the compression algorithms based on their input data types (numerical/strings/byte encoded). Discuss the internals of the algorithms and contrast their approaches.

(c) Implement an API that can select and execute an algorithm from the above list for a given data stream/collection. Note, that pre-processing may be necessary to convert integer data into the appropriate format as required by a particular algorithm. In case an algorithm supports both strings and data as bytes for encoding, a comparison of performance between both the input types is required.

(d) Provide a comprehensive analysis of the performance of the at least 5 compression algorithms (the student is encouraged to find more) while varying data sortedness of the input stream. Differently sorted data can be generated using the workload generator included with the BoDS Benchmark.

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

References

[1] M. Hosseini, "A Survey of Data Compression Algorithms and their Applications," in Applications of Advanced Algorithms, 2012.
[2] 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