JOHANNES KEPLER
UNIVERSITY LINZ

Master's Thesis

Floating-Point Compression of Time Series

Student: Konstantin Urbanides, BSc
Advisor: Dipl.-Ing. Dr. Markus Weninger, BSc
Company advisor (Dynatrace): Günter Schwarzbauer
Start date: September 2024

Dipl.-Ing. Dr. Markus Weninger, Bsc
Institute for System Software
T +43-732-2468-4361
markus.weninger@jku.at

Problem Statement

The aim of this thesis is to evaluate the existing techniques for floating-point compression and to correlate them with the behavioral differences that arise with time series obtained from real-world data. A significant number of papers in this field utilize synthetic datasets that do not reflect real world use cases. Accordingly, this thesis employs data gathered in production systems for the analysis of compression techniques. A limitation of many papers in this field is that they primarily focus on the compression rate, while neglecting the fact that the usage of such algorithms in real world applications depends on a multitude of factors, including performance and cost reduction with regard to resource usage. The primary trade-off when utilizing such compression algorithms is between performance and compression rate. Additionally, the characteristics of the data being compressed exert a considerable influence on the efficacy of the compression process. In practical applications, metrics exhibit a range of behaviors with regard to value variance. Consequently, metrics can be grouped into distinct categories, such as cyclic or event-driven metrics. This diversity of metrics and their associated categories explains why some compression techniques are more effective than others for specific metric categories.

Scope

The compression algorithms examined in this thesis are specifically designed for time series and produce a bitstream containing the compressed input. As such, time series can be characterized as a sequence of pairs, each containing a timestamp and floating-point values. While time series are made up of timestamps and floating-point values, only the compression of floating-point values is relevant in this work. Furthermore, the behavior of the different compression techniques with respect to certain categories of metrics will be examined.
From the text above, the following summary can be derived:
- Only consider metrics that contain floating-point values
- A time series can be defined as a sequence of pairs, each comprising a timestamp and floating-point values
- Evaluation of the compression algorithms on real-world data
- Monitoring data gathered from systems in operation containing a multitude of metric categories

Goals

The principal objective of this thesis is to evaluate the performance of compression techniques on real-world data by implementing a selection of algorithms identified through a preliminary research phase. Illustrative examples of these algorithms include GorillaDB, CHIMP, and ELF. The performance of these algorithms is significantly influenced by the nature of the data employed. Consequently, the comparison of compression techniques will be based on real data from operational environments. This thesis diverges from existing studies due to its focus on real-world data, providing insights into actual performance rather than employing algorithms on

synthetic data. The monitoring data from production systems is distinguished by the fact that it encompasses data generated by thousands of services, each of which may contain tens or hundreds of metrics. Given the disparate responses of the algorithms to the various types of metrics, it is imperative to conduct a comprehensive examination of these differences. Based on the findings, recommendations can be formulated for the practical implementation of theses findings, with the aim of enhancing the efficacy of existing techniques.

Nice to Have

- Combination of several compression techniques.

One such technique could be the application of general compression algorithms, such as Zstandard, as a post-compression step. This approach can utilize bit patterns that emerge subsequent to the application of a compression algorithm, thereby facilitating a further reduction of the compressed data size.

- Optimize/adapt selected compression algorithm

Modalities:

The progress of the project should be discussed at least every three to four weeks with the advisor. A time schedule and a milestone plan must be set up within the first two weeks and discussed with the advisor(s). It should be continuously refined and monitored to make sure that the thesis will be completed in time. The final version of the thesis is expected to be finished before 31.08.2025.