Reading Notes

##TVStore Yanzhe An and Wang (2022)

Abstract

A pressing demand emerges for storing extreme-scale time series data, which are widely generated by industry and research at an increasing speed. Automatically constraining data storage can lower expenses and improve performance, as well as saving storage maintenance efforts at the resourceconstrained conditions. However, two challenges exist: 1) how to preserve data as much and as long as possible within the storage bound; and, 2) how to respect the importance of data that generally changes with data age.

To address the above challenges, we propose time-varying compression that respects data values by compressing data to functions with time as input. Based on time-varying compression, we prove the fundamental design choices regarding when compression must be initiated to guarantee bounded storage. We implement a storage-bounded time series store TVStore based on an open-source time series database. Extensive evaluation results validate the storageboundedness of TVStore and its time-varying pattern of compression on both synthetic and real-world data, as well as demonstrating its efficiency in writes and queries.

Introduction

- "Time series databases are becoming the most popular type of databases in recent years." Newest data seems to suggest otherwise? gmbh (2023)
- "... the fast increasing volume of time series data has placed an unprecedented requirement on computing resources, especially storage space [6, 79]." pg 83
- "as the significance of time series data is highly correlated with the age of the data [22,37,89], it is desirable to have a storage management strategy that takes data ages into account [3, 7]." pg 83
- " Significant prior work has addressed the storage-control problem by compression, which can be lossless or lossy. Lossless compression [10, 33, 57, 71, 73] preserves the complete data, but its achievable upper bound on compression ratio [93] might not be satisfactory for applications." pg 83
- "Hence, time series databases commonly control storage consumption by directly discarding data older than a given time [43] or exceeding a storage threshold [67]. But discarding historical data causes a loss [94]." pg 83
- "Another common approach is to exploit lossy compression [15,41,65], which preserves partial data and trades off precision for space. But existent approaches to lossless and lossy compression are only best-effort about the final size of compressed data size [13, 24, 99]." pg 83
- *Problem statement* "We consider the problem of automatically bounding the storage of a time series store by compression. To enable this, our key insight is that time series data can be compressed losslessly or lossily according to its importance, which is in turn related to its age, as users commonly accept information loss on less important old data [12, 14, 23, 38, 40]." pg 83
- "The goal is to reduce computing resource consumption and improve performance." pg 84

Background and Motivation

- "…mounting demands have emerged for keeping time series data for future analysis [94]. But time series data are generated at a growing speed that is outpacing the increase of computing capabilities [17, 79]. Many application scenarios cannot afford enough computing resources such as storage and network bandwidth to accommodate the processing needs for time series data." pg 84

- "Sensors of a connected car can generate about 30 terabytes (TB) of data per day [62, 77]…. Since a 30TB disk can cost around $1200, a month's worth of data can fill up a 960TB disk, causing a cost of $30,000." pg 84

- "In the oil and gas industry, a typical offshore oil platform generates more than 1TB of data [19] daily. But common data transmission via satellite connection allows only a speed from 64 Kbps to 2Mbps for these offshore oil platforms. If all data are transmitted back for processing, it would take more than 12 days to move 1 day's worth of data to the processing backend [8]." pg 84

- "… cosmological simulations generate petabytes of data per simulation run [34] and climate simulations generate tens of terabytes per second [29]." pg 84

- "Data reduction is necessary to enable data processing and analytics within a reasonable amount of resource and time [92]." pg 84

- "The importance of time series data changes along with time, as reflected by applications' favoring recent data over old data [5, 18, 31], or favoring some events at certain moments over others [49,83]." pg 85

- "As a result, we have seen a plethora of research on data series analysis and prediction considering the timechanging pattern [9,22,36,37,89]." pg 85

- "As time series data can be identified by timestamps, we use a time-dependent function to denote the changing importance of data. Hence, the compression ratios can also be deduced from the function." pg 85

- "We must deduce the proper moments for compression initiation when 1) it is not too late that the storage space is exceeded during compression; and, 2) it is not too early that unnecessary compression is applied to some recent data or that an improperly high compression ratio is used." pg 85

- "Challenges: As a result, two main challenges exist in automatically bounding the time series storage by timevarying compression: 1) how time-varying compression can be executed on an ever-increasing volume of data and with error bounds computed, when the compression ratios keep changing… and, 2) how to automatically decide the conditions for running time-varying compression such that storage space is always bounded but not too much…" pg 85

- time-varying compression (TVC)

- "To guarantee that data are compressed to the compression ratio sequence r_1,r_2,…,r_k, TVC groups r_i data chunks into the ith segment. Each segment is then compressed to an output chunk. Hence, the ith chunk of the compression output has a compression ratio r_i, complying to the definition of r(t)." pg 86

gmbh, solidIT consulting & software development. 2023. "DBMS Popularity Broken down by Database Model." https://db-engines.com/en/ranking_categories.

Yanzhe An, Yuqing Zhu, Yue Su, and Jianmin Wang. 2022. "TVStore: Automatically Bounding Time Series Storage via Time-Varying Compression." In *Proceedings of the 20th USENIX Conference on File and Storage Technologies*, 83–99. USENIX Conference on File and STorage Technologies. Santa Clara, CA, USA: USENIX Association.