УДК 004.6

OPTIMIZING DATA ORGANIZATION: A JOURNEY INTO SORTED STRING TABLES

Vashchenko M.
Academic Supervisor – Candidate of Technical Sciences,
Associate Professor Kravets N.
Kharkiv National University of Radio Electronics, Department of MEEPP,
Kharkiv, Ukraine
e-mail: vashchenkonik@nure.ua

This article delves into the adoption of Sorted String Tables (SSTables) as a solution for efficient and organized data storage in modern information systems. It addresses the challenges posed by disorderly data arrangement and presents SSTables as a remedy, with a focus on key aspects such as key uniqueness and the compaction process. The advantages of SSTables, including streamlined data merging and two-level indexing, are underscored alongside strategies for risk minimization. Moreover, the article highlights the widespread adoption of SSTables in various data stores and distributed systems, advocating for further research to optimize performance and enhance data recovery mechanisms.

In contemporary information systems, there is a pressing need for efficient and structured data storage. Disorderly arrangement of keys and their associated values in journal tables presents a significant challenge. This article discusses the shift towards a more organized data storage approach using SSTables, along with methods to ensure data orderliness on disk.

Ensuring key uniqueness in each data segment is vital for SSTables' effective operation, aiming to prevent data duplication and maintain information consistency. The compaction process plays a pivotal role in achieving this goal by merging and restructuring data segments, eliminating duplication, and ensuring key uniqueness. This involves analyzing, ordering, and merging keys from different segments to meet uniqueness requirements, thereby preserving data order and integrity in SSTables. Additionally, data management mechanisms like replication strategies, data version control, and conflict resolution further contribute to ensuring key uniqueness, reinforcing the structuredness and integrity of data in SSTables.

The advantages of transitioning to SSTables encompass a wide array of benefits that significantly enhance the efficiency and performance of data storage systems. One notable advantage is the capability to merge data, even when the volume of keys exceeds available RAM. This capability opens up new possibilities for handling large volumes of data, previously hindered by memory limitations. Transitioning to SSTables, based on the merge sort algorithm, offers an efficient mechanism for combining data from different segments, enabling optimal utilization of available resources. Moreover, data merging facilitates

effective management of changing data volumes and enables dynamic scaling of system performance according to application requirements. Furthermore, the data merging mechanism ensures proper handling of matching keys, retaining only the latest values for identical keys, thereby reducing information duplication and optimizing data storage utilization. Consequently, transitioning to SSTables not only ensures efficient management of data volumes but also guarantees the integrity and consistency of information in storage.

Data block compression is essential for efficient utilization of storage resources and bandwidth in data storage systems. Maintaining data orderliness in SSTables necessitates the application of a two-level indexing strategy. This strategy involves sorting data both on disk and in memory to enhance query performance and ensure rapid access to data. Data indexing occurs on disk at the first level, using specialized data structures such as B-trees or LSM-trees, facilitating quick data retrieval and efficient disk space utilization. The second level of indexing, conducted in memory using data structures such as hash tables or search trees like AVL or red-black trees, enables fast access to data in memory, minimizing system response time to queries. Thus, the two-level indexing strategy, both in memory and on disk, ensures efficient data management in SSTables, facilitating swift and convenient access to information.

The transition to SSTables not only streamlines data storage but also facilitates the formation of optimal queries by ensuring data orderliness and efficiency. The systematic arrangement of data segments in SSTables enables streamlined query execution, reducing computational overhead and latency. Additionally, the inherent compaction process aids in maintaining query integrity by eliminating redundant data and ensuring key uniqueness, thereby enhancing the accuracy and reliability of query result [1].

The vulnerability of data storage systems based on memtable, particularly in the event of application failures, poses a risk of data loss. To mitigate this risk, a combined mechanism for saving data to disk and memory is proposed. Memtable is utilized for quick data recording to RAM before their long-term storage on disk in SSTables. Periodic data recording to disk ensures reliable data recovery following application failures, minimizing the risk of data loss and ensuring data integrity in storage.

References:

1. Shubin, I., & Kozyriev, A. (2023). Method for solving quantifier linear equations for formation of optimal queries to databases. Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2023). CEUR Workshop Proceedings, 3396, 449–459. https://ceur-ws.org/Vol-3396/paper36.pdf.