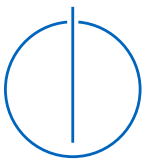
SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Implementing an Efficient Shuffle Operator for Streaming Database Systems

Jonas Ladner



SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Implementing an Efficient Shuffle Operator for Streaming Database Systems

Implementierung eines effizienten Shuffle-Operators für Streaming-Datenbanksysteme

Author:	Jonas Ladner
Examiner:	Prof. Dr. Thomas Neumann
Supervisor:	Maximilian Rieger M.Sc.
Submission Date:	17.02.2025

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 17.02.2025 Jonas Ladner

Acknowledgments

Abstract

Modern streaming database systems rely on efficient data partitioning to achieve scalability and high performance across processing nodes. Partitioned data shuffling is a crucial operation, as it is used to prepare and distribute data for further processing on distributed systems.

In this thesis, various partitioning implementations are compared and evaluated by simulating real-world usage of the shuffle operator. The implementations have to process incoming tuple batches and partition them into output buckets, which are based on slotted pages and can be passed to subsequent operators. The implementations are evaluated based on their performance, scalability and memory consumption.

The results demonstrate that using a lock- and Software Managed Buffers (SMB)-based approach yields the best overall performance, offering both efficiency and ease of implementation. Notably, the proposed locking mechanism minimizes the duration of holding a lock, ensuring minimal contention and contributing to the approach's superior performance.

Contents

1 Introduction

1.1 Motivation

Growing demand for real-time data analysis and increasing data volume create significant challenges [1–3]. Stream processing systems address these issues by distributing tasks across worker nodes [1, 4]. Data shuffling prepares and distributes tuples among worker nodes [5]. As a core streaming component, the shuffle operator must achieve high throughput and low latency to support efficient downstream processing. This thesis addresses these challenges by focusing on efficient implementations of the shuffle operator.

1.2 Streaming processing engines

Streaming processing engines are designed to process data as soon as it arrives rather than relying on traditional pre-computed information and index structures [1]. Their core operations include partitioning and distributing incoming traffic across worker nodes. The distribution process is frequently based on a partitioning function. These partitions can then improve the performance of further operators by ensuring the data locality of interdependent tuples [3]. Data locality within the worker node is crucial for maintaining performance and scalability in large-scale deployments.

1.3 Shuffle operator

The shuffle operator provides a partitioned distribution of tuples, enabling downstream operators to leverage the data locality of partitioned data blocks. For instance, the throughput of the join operator can significantly be improved when tuples assigned to the same hash bucket are shuffled to the same worker node [3]. Implementing a shuffle operator involves addressing challenges like memory consumption, latency, and scalability. This thesis proposes and evaluates different implementations of the shuffle operator, focusing on their efficiency and performance.

1.4 Slotted pages

Slotted pages are a common way to store variable-size tuples within fixed size memory blocks [6]. These fixed size memory blocks can then be either stored on disk or easily be sent to worker nodes.

Typically, the pages consist of three sections: metadata, slots and a variable-size data section. The metadata area contains information like an identifier for the page, what fields the tuples have and the amount of tuples on this page. This fixed-size metadata section is then followed by the slot section. A single slot contains the fixed-size properties, the variable-size length and its start offset in the variable-size data section. In contrast to the previous two sections, the variable-size data section grows from the end of the page towards the slot section of the page.

1.5 Contribution

The key contribution of this thesis is the creation and evaluation of various implementations of the shuffle operator. In order to simulate the real-world usage of the shuffle operator, the following three-step shuffle-simulation is proposed:

1. Tuple generation: The tuples are generated in a batched manner using a pseudo-random generator. Each implementation requests the ad-hoc generation of tuples, which contain 32-bit key field and an optional variable-size data field.

2. Data shuffle: The requested, random-generated tuples are then processed using the different implementations and stored within partition buckets. Each partition bucket consists of slotted pages, where the tuple of this partition are stored.

3. Storing tuples on slotted pages: The implementations range from thread-local to shared slotted-page write-out strategies. The implementations using a shared write-out policy, can then be further categorized into locking and lock-free approaches.

While the implementations are optimized for the process above, the underlying algorithms can be transferred to any streaming system, that uses data communication platform based on slotted pages.

2 Related work

2.1 Radix Partitioning

2.2 Software Managed Buffers

3 Implementations

3.1 On-Demand Partitioning

3.1.1 Overview

3.2 Software Managed Buffers-based Partitioning

3.2.1 Overview

3.2.2 Lock-based Implementations

3.2.3 Lock-free Implementations

3.3 Histogram-based Partitioning

3.3.1 Overview

3.3.2 Radix Partitioning

3.3.3 Ad-hoc Radix ("Hybrid") Partitioning

3.4 Thread-Local Pages and Merge-based Partitioning

3.4.1 Overview

3.5 Collaborative Morsel Processing

3.5.1 Overview

3.5.2 Collaborative Morsel Processing using Software Managed Buffers

3.5.3 Collaborative Morsel Processing using Processing Units

3.6 Complexity Analysis

3.6.1 Time Complexity

Tuple Access Count

3.6.2 Space Complexity

Memory Consumption

4 Evaluation

4.1 Experimental Setup

4.1.1 Hardware

4.1.2 Software

4.2 Tuple Generation

4.3 Tuple Write Benchmark

4.4 Shuffle Benchmark

4.4.1 Memory Consumption

4.4.2 Performance

4.5 Comparison with Stream Processing Systems

5 Conclusion

5.1 Conclusion

5.2 Future Work

Abbreviations

SMB Software Managed Buffers

List of Figures

List of Tables

Bibliography

[1] F. Gürcan and M. Berigel. "Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges." In: *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. Oct. 2018, pp. 1–6. DOI: 10.1109/ISMSIT.2018.8567061.

[2] E. Zamanian, C. Binnig, and A. Salama. "Locality-aware Partitioning in Parallel Database Systems." In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, pp. 17–30. ISBN: 9781450327589. DOI: 10.1145/2723372.2723718.

[3] S. Chu, M. Balazinska, and D. Suciu. "From Theory to Practice: Efficient Join Query Evaluation in a Parallel Database System." In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, pp. 63–78. ISBN: 9781450327589. DOI: 10.1145/2723372.2750545.

[4] G. Andrade, D. Griebler, R. Santos, and L. G. Fernandes. "A parallel programming assessment for stream processing applications on multi-core systems." In: *Computer Standards & Interfaces* 84 (2023), p. 103691. ISSN: 0920-5489. DOI: https://doi.org/10.1016/j.csi.2022.103691.

[5] F. Liu, L. Yin, and S. Blanas. "Design and Evaluation of an RDMA-aware Data Shuffling Operator for Parallel Database Systems." In: *ACM Trans. Database Syst.* 44.4 (Dec. 2019). ISSN: 0362-5915. DOI: 10.1145/3360900.

[6] A. Ailamaki, D. J. DeWitt, and M. D. Hill. "Data page layouts for relational databases on deep memory hierarchies." In: *The VLDB Journal* 11.3 (Nov. 2002), pp. 198–215. ISSN: 1066-8888. DOI: 10.1007/s00778-002-0074-9.