Technische Universität Berlin

Chair of Database Systems and Information Management

Master's Thesis

Anonymized Access Control for Distributed Event Stores

Henri Tyl Allgöwer
Degree Program: Computer Science
Matriculation Number: 454925

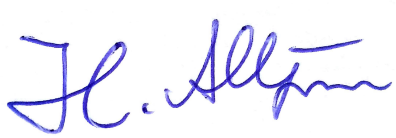
Reviewers
Prof. Dr. Volker Markl
Prof. Dr. Odej Kao

Advisor
Rudi Poepsel Lemaitre

Submission Date
30.11.2023

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

Berlin, October 2, 2023



Henri Tyl Allgöwer

Zusammenfassung

Tipps zum Schreiben dieses Abschnitts finden Sie unter [4]

Abstract

The abstract should be 1-2 paragraphs. It should include:

- a statement about the problem that was addressed in the thesis,
- a specification of the solution approach taken,
- a summary of the key findings.

For additional recommendations see [4].

Acknowledgments

For recommendations on writing your Acknowledgments see [5]. Thank you to the chair at Database Systems and Information Management (DIMA)

Contents

Chapters 3-5 are the core of the thesis, whereas Chapters 1, 2, 6, and 7 provide context. The major contributions should be in Chapters 4 and 5. This structure serves as a guideline and should be customized accordingly. In particular, the generic chapter titles should be replaced with more specific ones, where appropriate (e.g., Chapter 4).

List of Figures

List of Tables

List of Abbreviations

DIMA Database Systems and Information Management

List of Algorithms

1 Introduction

... should include the following:

- motivation (why is this problem interesting? offer examples),
- research challenge (what is the obstacle to be overcome?),
- novelty (was this problem already solved?),
- anticipated impact (how does solving this problem impact our world?).

1.1 Motivation

The increasing popularity of data streaming in corporations highlights the imperative need for incorporating anonymization and data masking techniques in this technology. Particularly noteworthy is the extensive adoption of distributed event stores, which are scalable and fault-tolerant systems designed to capture, store, and process real-time data streams, across various sectors, including Fortune 100 companies, governments, healthcare, and transportation industries [5]. This widespread usage emphasizes the criticality of ensuring data privacy within these distributed event stores, while also highlighting the potential transformative impact of effective anonymization and data masking techniques in this domain.

1.2 Research Challenge

In the contemporary data-driven world, the growing demand for comprehensive data privacy policies is matched by increasingly stringent regulations from governments worldwide [1,2]. However, the underlying infrastructure to adequately support these policies is markedly lacking [3,4]. This disparity poses a unique challenge, particularly when considering the demands of modern database systems to maintain high performance, characterized by low latency and high throughput.

1.3 Novelty

While there exists a body of work focusing on anonymization and data masking for data streaming [6,7,8], there is a noticeable gap of research specifically targeting distributed event stores. Furthermore, although there are enterprise technologies for managing data flowing into such systems [9], there is limited literature on techniques designed for data already within. Most notably, the concept of integrating Role-Based Access Control (RBAC) within this framework, where the role assigned determines the level of anonymity accorded to the data, is a completely novel approach.

1.4 Anticipated Impact

The integration of data privacy policies with modern data stream structures, such as distributed event stores, would be a significant innovation. The introduction of Role-Based Access Control (RBAC) coupled with anonymization in distributed event stores holds the potential to contribute to more advanced, efficient, and secure data handling. By making such tools accessible and cost-effective, companies might be more inclined to prioritize and invest in user data privacy.

1.5 Research Problem

... should include the following:

- a succinct, precise, and unambiguous statement of the research problem or question to be solved,
- goals and subproblems that will be explored, including the scope of the thesis (i.e., what is in and out of scope).

Area (Million sq. miles)	Calling Code
0.29	56
0.3	90
3.8	1
0.5	51
600	9800
Pearson = 1.0	Spearman's = 0.1

Table 1: Correlation in the existence of outlier[2].

1.6 Outline

2 Literature Review

2.1 Distributed Event Stores

2.2 Anonymization

2.3 Role Based Access Control

2.4 Privitar

... should include the following:

- definitions / technical terms,
- theoretical foundations / principles,
- descriptions of algorithms, hardware, software, and/or systems employed.

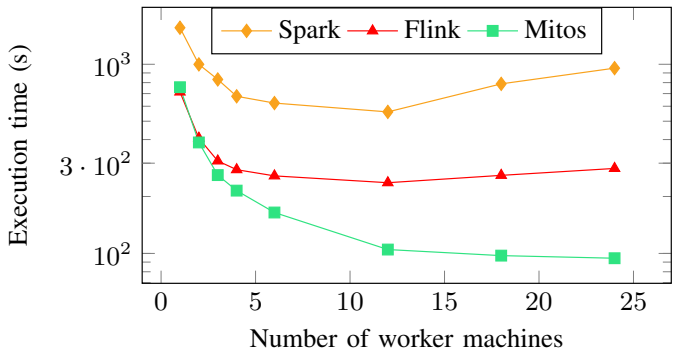


Figure 1: Strong scaling for Visit Count[1].

3 Theoretical Framework

3.1 Apache ZooKeeper

3.2 Apache Kafka

3.2.1 Architecture

3.2.2 Administration

3.2.3 Connector

3.2.4 Streams

3.3 Masking Functions

3.4 Anonymization Levels

Use case RBAC

4 System Design

4.1 Anonymized kafka

UML Sequence diagram High level Component Diagram Anonymizer Registry Configuration file The role of the Data Officer

4.2 Role Based Access Control

4.3 Data Pipeline

5 Implementation

5.1 Anonymized Kafka

Class Diagram

5.1.1 Stream Manager

5.1.2 Configuration parsing

5.1.3 Validation

5.1.4 Stream Config Builder

5.1.5 Kafka Streams

5.1.6 Anonymizers

5.2 Test Suite

5.2.1 Data Generator

5.2.2 Kafka Connector

5.2.3 Consumers

5.3 Docker

Putting it all together Docker compose Network Volumes Dependencies Individual Dockerfiles

... should include the following:

- research methodology (e.g., prototype and experiments, case study, literature survey, theoretical analysis),
- derivations and descriptions of algorithms, hardware, software, and/or systems developed.

Algorithm 1: Splitting a Session[3].

Parameters:
e : Tuple to be inserted.
te (e) : Event-time of e.

S ← slice that covers te (e);
if *S starts at te (e)* **then**
  //Slice before S must be fixed.
  change the type of the slice before S to combined;
  add e to S;
else
  // S does not start at te (e).
  change tend (S) to te (e) (excluding te (e) from S);
  change type of S to flexible;
  add slice in [te (e), former tend (S)] with former type of S.
  add e to the new slice.
end

6 Testing and Evaluation

TODO

6.1 Experimental Setup

... should include the following:

- define experimental data and workload(s),
- discussion about the selection and interpretation of the evaluation metrics,
- discussion about the computing environment, including hardware, software, tools.

5.X Design and an Interpretation of the Results (For each Experiment Class X)

... should include the following:

- which experiments will be conducted and why?
- for each experiment, what are objectives, baselines, and expected results?
- description and an interpretation of the experimental results,
- explanation for any anomalies or any unexpected behavior.

7 Conclusion

7.1 Future Work

... should include the following:

- problem restated and a brief summary of the methodology,
- student contributions (e.g., survey, open-source software, journal publication),
- a brief summary of the findings and results,
- limitations and generalizability of the findings and results.
- lessons learned,
- recommendations for future research.

Bibliography

[1] Gévay, G.E., Rabl, T., Breß, S., Madai-Tahy, L., Quiané-Ruiz, J.A., Markl, V.: Efficient control flow in dataflow systems: When ease-of-use meets high performance (2021), https://www.researchgate.net/publication/349768477_Efficient_Control_Flow_in_Dataflow_Systems_When_Ease-of-Use_Meets_High_Performance, to be published

[2] Mahdi Esmailoghli, Jorge-Arnulfo Quiané-Ruiz, Z.A.: Cocoa: Correlation coefficient-aware data augmentation (2021)

[3] Traub, J., Grulich, P.M., Cuéllar, A.R., Bress, S., Katsifodimos, A., Rabl, T., Markl, V.: Scotty: General and efficient open-source window aggregation for stream processing systems (2020), https://www.redaktion.tu-berlin.de/fileadmin/fg131/Publikation/Papers/Traub_TODS-21-Scotty_preprint.pdf

[4] Wallwork, A.: English for writing research papers, chap. Abstracts, pp. 177–245. Springer International Publishing Switzerland (2011)

[5] Wallwork, A.: English for writing research papers, p. 306. Springer International Publishing Switzerland (2011)

Appendix A. Further Details on the Solution Approach

Appendix B. Extended Version of the Experimental Results