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 7, 2023

Henri Tyl Allgöwer

Zusammenfassung

Tipps zum Schreiben dieses Abschnitts finden Sie unter [6]

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 [6].

Acknowledgments

For recommendations on writing your Acknowledgments see [7]. Thank you to the chair at Database Systems and Information Management (DIMA)

Contents

List of Abbreviations . . . . . **x**

List of Algorithms . . . . . **xi**

1 Introduction **1**
1.1 Motivation 1
1.2 Research Challenge 1
1.3 Novelty 2
1.4 Anticipated Impact 2
1.5 Research Problem 2
1.6 Outline 3

2 Literature Review **4**
2.1 Distributed Event Stores 4
2.2 Anonymization 4
2.3 Role Based Access Control 4
2.4 Privitar 4

3 Theoretical Framework **5**
3.1 Managing Different Anonymization Granularity 5
3.1.1 Use Case Example 6
3.2 Masking Functions 9
3.3 System Requirements 9
3.3.1 Data Stream Integration 9
3.3.2 Administration 9
3.3.3 Performance 9
3.3.4 Adaptability 9
3.3.5 Scalability 9
3.3.6 Reliability 9

4 System Design **10**
4.1 Anonymized kafka 10
4.2 Role Based Access Control 10
4.3 Data Pipeline 10

5 Implementation **11**
5.1 Anonymized Kafka 11
5.1.1 Stream Manager 11
5.1.2 Configuration parsing 11
5.1.3 Validation 11
5.1.4 Stream Config Builder 11
5.1.5 Kafka Streams 11
5.1.6 Anonymizers 11
5.2 Test Suite 11
5.2.1 Data Generator 11
5.2.2 Kafka Connector 11
5.2.3 Consumers 11
5.3 Docker 11

6 Testing and Evaluation **13**
6.1 Experimental Setup 13
5.X Design and an Interpretation of the Results (For each Experiment Class X) 13

7 Conclusion **14**
7.1 Future Work 14

Appendix A. Further Details on the Solution Approach **16**

Appendix B. Extended Version of the Experimental Results **17**

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

1 Strong scaling for Visit Count[2]. 4

List of Tables

1 Correlation in the existence of outlier[3]. 2

2 Example datum of a diabetes patient 6
3 Data available for the nurse staff. Note that pid, address, insurance number and additional medical information. 7
4 Data available for the administration. Note that only the insurance information, medication and diagnosis are not suppressed. 8
5 Data available for external research. Note the untouched medical, the suppressed personally identifying and generalized quasi identifiable attributes. 9

List of Abbreviations

DIMA Database Systems and Information Management

CCPA California Consumer Privacy Act

GDPR General Data Protection Regulation

ICD International Statistical Classification of Diseases and Related Health Problems

List of Algorithms

1 Splitting a Session[5]. 12

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[3].

Is there also indentation here?
Hard to tell without two lines
Oh wow there is

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.

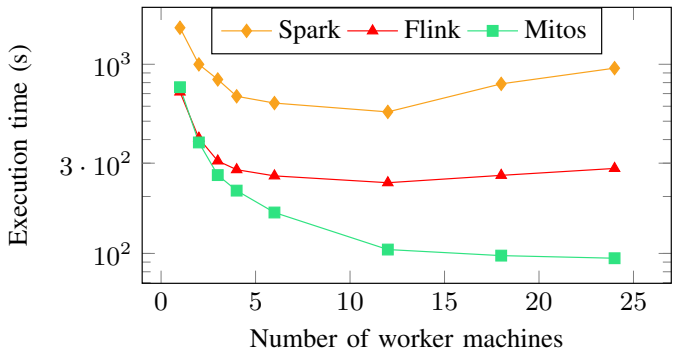

Figure 1: Strong scaling for Visit Count[2].

3 Theoretical Framework

3.1 Managing Different Anonymization Granularity

When planning to integrate anonymization techniques into existing systems, there are many things to consider. First one must understand the data flowing through the system. Does it include personally identifiable information? Is there further sensitive data? What part of it is necessary for system maintenance? Maybe there is additional data collected for statistics. Bearing this in mind the next thought would what needs to be anonymized. For this privacy agreements with the user must be taken into account. There may be additional government regulations in place like the California Consumer Privacy Act (CCPA) in the United States of America or the General Data Protection Regulation (GDPR) in the European Union. The next course of action is deciding on specific masking functions. As was discussed in 2.2 there is a plethora to chose from. It is important to note that all forms of anonymization lead to a loss of information. While choosing Blurring or Suppression, two methods that replace attributes with a placeholder, for all critical data fields derived in the previous assessment will ensure that all privacy concerns are addressed, it will also diminish all intelligence gained from collecting this data in the first place. It is questionable if not collecting this type of data in the first place would then be the better solution as it would safe storage and computing power. An alternative approach would be to invest heavily in the IT security ensuring that no intruder with malicious intent can gain access to sensitive data. Keeping in mind, however, that social engineering attacks are nowadays the most common and effective strategy (SAUCE) giving a guarantee of safety can be impossible. It also stands to reason that the more employees a company has the risk for social engineering attacks increases. Both of these radical approaches do not seem to adequately solve the problem. Fortunately, there is a way to navigate between these two extremes. An attentive observer of the company's data operations will likely notice that data can and should be restricted similarily to permissions: Only the data needed to furfill the user's duty should be accessible to the user. In addition, there are anonymization techniques, which do not lead to total information loss like the two mentioned before. Generalization for example can be employed to significantly reduce the re-identification of individuals, while simultaneously retaining some information. With data restriction and

different anonymization techniques in mind there is a middle ground to be found which maximizes security and minimizes information loss. Consider the following example:

3.1.1 Use Case Example

Hospitals depend on the collection, management and analysis of data to administer the best and most accurate care of their patients. In a modern hospital all data would be stored in a centralized hospital database. Here all data for individual patients are brought together. Table 2 shows an exemplary entry of a patient in the diabetes and endocrinology ward of a german hospital. Note that the table has been shortened to enhance its clarity and readability. (A MORE DETAILED VERSION CAN BE FOUND IN THE APPENDIX)

pid	name	addr.	gender	age	ins. co.	ins. no.	diag.	gluc.	hba1c	med.
1	F. Ott	Berlin	M	28	TK	K15489	E10	22.1	8.74	Insulin

Table 2: Example datum of a diabetes patient

The header of table 2 shows eleven attributes. First the person id (*pid*), which is the primary key for each patient as it uniquely identifies the patient in the database. Then *name*, address (*addr.*), *gender* and *age* are included as additional personal information. This would typically also include the full address not just the city name, contact information, height as well as weight to adapt the dosage of the medication. The subsequent two attributes include information about the patient's insurance information. Insurance companies in Germany are uniquely identified with a nine digit institutional identifier. In this example the insurance company (*ins. co.*) has the value 101575519, which matches the identifier of the Techniker Krankenkasse (TK). Each client is then assigned a number unique to that insurance company called the insurance number (*ins. no.*). It always starts with a letter followed by digits. Finally, the datum references the medical information. It starts with the diagnosis (*diag.*) classified according to the International Statistical Classification of Diseases and Related Health Problems (ICD). E10 being the label for Type 1 Diabetes Mellitus. The most important medical measurement for the treatment of this disease is the current amount of glucose (*gluc.*) in the blood. This determines the quantity of medication (*med.*) to be administered to the patient. For Type 1 Diabetes this is Insulin. Lastly, the table includes an attribute called *hba1c*. This is the body's own three-month average of blood

glucose. By means of which diabetes is diagnosed. In this case it is also symbolic for all additional diagnostic findings. Glucose and HbA1c are intentionally distinguished as separate attributes in this dataset, despite both being blood-derived metrics, due to their distinct measurement methodologies and relevance in immediate treatment contexts. Glucose can be ascertained with a single drop of blood, providing critical information for the immediate treatment. Conversely, HbA1c is derived from a complete blood count and does not require instant action.

In a hospital setting, numerous actors engage with the aforementioned dataset. The most straightforward and prominent is the doctor. She will need all data to fulfill her duties. The doctor's letter contains all personal information. The medical data is needed for diagnosis and treatment. She will also need to keep the insurance information in mind as the covered treatment options are oftentimes different for each company. Additionally, she will need to write the patient's insurance information on the rescriptions. Only the pid could be omitted, but is debatable if the overhead is worth it, considering the pid can be easily inferred with all the given information. Therefore, no anonymization to the doctor's data makes the most sense.
Supporting the doctor is the nurse staff. One of their main tasks is to monitor patients and administer medication. To accomplish this they require the diagnosis, medication and in this case the glucose data. As the HbA1c value is not relevant for the immediate treatment it can be safely ommitted. Again insurance information is necessary as nurses typically do have the liberty of administer medication according to their own judgement. This is especially important when considering how understaffed hospitals in Germany are most of the time. On the other hand the patient's personal insurance number does not play into this. As nurses also interact directly with the patients they need some basic personal information like name and gender. Pid and address, however, are not required. Therefore, the data for the nurse staff can be anonymized as shown in Table 3 without limiting the nurses or losing valuable information.

pid	name	addr.	gender	age	ins. co.	ins. no.	diag.	gluc.	hba1c	med.
*	F. Ott	*	M	28	TK	*	E10	22.1	*	Insulin

Table 3: Data available for the nurse staff. Note that pid, address, insurance number and additional medical information.

In tandem with the stay and medical treatment of the patient, the administartion of the hospital will want to collect the money from the patient's insurance.

The insurance company together with the patient's personal insurance number will suffice as identification. Administered medication will be imperative as this dictates the amount of money the hospital will get in addition to the fees for the stay. For this the diagnosis will typically have to be added as a suitable reason. No further information is required. Limiting the amount of data here is crucial as here the data is exported to a third party. Which means that additional regulations will take effect. Minimizing the data leaving the hospital minimizes security risks. With these strict rules in place the data can be adjusted as seen in Table 4.
Note at this point that an attacker, who has gained access to both the data of the nurse staff and that of the administration, would struggle to correlate the entries. The shared available data fields insurance company, diagnosis and medication are likely generic enough to not point to a singular but to many patients.

pid	name	addr.	gender	age	ins. co.	ins. no.	diag.	gluc.	hba1c	med.
*	*	*	*	*	TK	K15489	E10	*	*	Insulin

Table 4: Data available for the administration. Note that only the insurance information, medication and diagnosis are not suppressed.

Diabetes, which afflicts over ten percent of the global population and demonstrates a rising prevalence, stands as one of the most common chronic diseases worldwide [1, 4]. Given its mostly non-lethal progression and lifetime dependency on medication, it has given rise to a substatial market. As cause, optimal treatment and cure remain subject to research, data of especially newer diabetes patients is in hot demand. To provide this data to research institutes in accordance with the regulations in place the hospital must ensure that no concrete patient can be reidentified. Here, advanced anonymization techniques such as K-Anonymization come into place. Each attribute of the data entry can be assigned to one of three categories: personally identifiable, quasi indentifying and remaining attributes. To achieve k anonymity each entry must supress the personally identifiable attributes, while keeping the remaining attributes untouched. Most importantly the quasi identifiable attributes of each data entry must be the same for at least k - 1 other entries of a data set. This is typically achieving with generalization of these attributes until k entries are found. In this use case the personally identifiable attributes are *pid*, *name*, *address* and *insurance number*. The quasi indentifying attributes are *gender*, *age* and *insurance company*. The medical data comprive the remaining attributes. A K anonymous version of this data extry is depicted in Table 5.

FAZIT?

pid	name	addr.	gender	age	ins. co.	ins. no.	diag.	gluc.	hba1c	med.
*	*	*	{M, F, X}	[20 - 30]	ins. co.	*	E10	22.1	8.74	Insulin

Table 5: Data available for external research. Note the untouched medical, the suppressed personally identifying and generalized quasi identifiable attributes.

3.2 Masking Functions

3.3 System Requirements

3.3.1 Data Stream Integration

3.3.2 Administration

3.3.3 Performance

3.3.4 Adaptability

3.3.5 Scalability

3.3.6 Reliability

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[5].

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] Federation, I.D.: Diabetes facts & figures (2023), https://idf.org/about-diabetes/diabetes-facts-figures/, accessed: 2023-10-01

[2] 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

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

[4] Organization, W.H.: Diabetes fact sheet (2023), ttps://www.who.int/news-room/fact-sheets/detail/diabetes, accessed: 2023-10-01

[5] 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

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

[7] 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