International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN: 2147-6799 www.ijisae.org Original Research Paper

A Generic Data Privacy Approach for Relational Databases and Data Warehouses

Rahul Kumar Sharma[1*], Vivek Kapoor[2]

Submitted: 22/04/2023 **Revised**: 27/06/2023 **Accepted**: 12/07/2023

Abstract: Currently data privacy is a big challenge for relational databases and data warehouses. Data privacy concerns are extremely critical in today's data driven world due to data privacy regulations like Global Data Protection Legislation (GDPR). Data privacy in relational databases and data warehouses is typically implemented using row and column level security. Currently most of the databases and data warehouses do not support row and column level security and very few of the databases and data warehouses support it in a very customized or specific way. This paper discusses an approach that can be used to achieve data privacy for relational databases and data warehouses in a generic and integrated way. A generic and integrated data privacy layer is proposed in this paper. This generic and integrated data privacy layer will provide row and column level security using rules and policies that can be defined at user, group or role level and will work with relational databases and data warehouses.

Keywords*: relational, privacy, generic*

I. Introduction

Data privacy has always been a challenge for relational databases and data warehouse. Row and column level security features are not available in many relational databases and data warehouses. Database management system providers and Data warehouse providers have been trying to provide customized solutions for solving data privacy challenges

Even if a database management system or data warehouse supports some limited capability for supporting row and column level security for data privacy, the implementation is customized and specific. There is no generic data privacy mechanism available that can work with all the existing relational databases and data warehouses.

Data privacy regulations like Global Data Protection Legislation (GDPR) govern the way in which we can use, process, and store personal data (information about an identifiable, living person). GDPR also makes it mandatory to have data privacy mechanisms in place for securing users' personal data.

The main purpose of our work is to provide a generic approach that can be used to achieve data privacy for relational databases and data warehouses.

This paper is organized as follows. Section 2 contains the literature review. Section 3 explores the problem domain and mentions privacy challenges in relational databases and data warehouses. Section 4 explains the proposed solution and proposed system architecture. Section 5 discusses the experimental analysis of the implemented methodology. Section 6, summarizes the conclusions and mentions about future work

[1*]*PhD Scholar, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India*

[2]*Asst. Professor, IET, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India*

II. Related Work

K. Shirudkar and D. Motwani [1] have discussed about Big data security challenges and have mentioned that it has become a major issue and concern due increasing volume and velocity of data being ingested in big data systems. Big data security has its own set of challenges due to the huge volume of data; current security mechanisms are slow and sometimes not effective for big data domain.

Duygu Sinanc Terzi et al. [2] have mentioned that it is very difficult to store big data and analyze it with traditional applications and it has challenging privacy and security problems. They have also provided categorization of various big data security and privacy studies. Extra requirements are needed for security and privacy in data gathering, storing, analyzing, and transferring for Big data.

V. N. Inukollu et al [3] have presented the security challenges related with Big Data systems for both traditional data centers and cloud-based environments. Cloud computing plays a critical role in protecting data, applications and related infrastructure using technologies, policies, and big data tools. Cloud computing and big data applications are likely to represent the most promising new frontiers in computing.

Yuan Tian [4] has mentioned current data security measures for bit data including – Authentication, Authorization, Data protection and auditing at preliminary level. Top 10 challenges for Big data security is also mentioned along with a proposed intelligent security model for achieving best big data security.

Anjana Gosain & Amar Arora [5] have compared the various security aspects of traditional data warehouse systems including – Encrypted data, Audit control, extendibility, platform independence model security, transformation, creation of platform specific model, Query/View/Transformations support and integration of multi-platform data.

Eduardo Ferna´ndez-Medina et al [6] have discussed about need for specifying security measures from the early stages of the data warehouse design and enforcing them due to the sensitive data contained in data warehouses. Traditional access control models for relational databases based on columns and rows of tables are not appropriate for DWs. Security and audit rules defined for data warehouses need to be specified based on the multidimensional modeling used to design data warehouses. They also presented an Access Control and Audit (ACA) model designed to represent major confidentiality and audit aspects in the conceptual modeling of data warehouses

Rahul Kumar Sharma & Dr. Vivek Kapoor [7] have proposed a data authorization framework to implement row and column level security in Hive. An integrated approach is provided to enhance the security of Hive/Hadoop for achieving row and column level security.

K. Michael and K. W. Miller [8] have discussed about key challenges of big data is to preserve individual privacy. Our digital footprints, that we leave behind, could denote unique aspects about ourselves when analyzed. Those unique aspects would otherwise go unnoticed, akin to digital DNA.

Fabian Prasser et al [9] mentioned that careful consideration is needed in privacy protection mechanisms for pooled data or when data is re-used for secondary purposes. They have also mentioned data anonymization is an important protection mechanism and suggested expert-level anonymization methodologies that can be integrated with ETL workflows

Janos Mezaros [10] has mentioned one of the most important changes in data privacy regulations - General Data Protection Regulation (GDPR) [18] where the aim is to protect data privacy in an online environment. GDPR requires stronger consent to protect the data subjects and introduced new rights, such as the right to be forgotten and data portability.

Curt Cortner et al [11] have described methods that provide multilevel and mandatory access control for a database management system. The access control techniques provide access control at the row level in a relational database table.

III.Problem Domain

Data privacy mechanisms are very crucial for relational databases and data warehouses. There is a dire need to have generic data privacy mechanisms that can work with relational databases and with data warehouses.

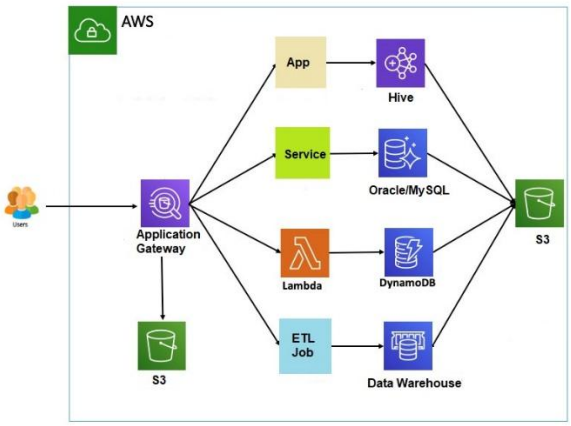
Data privacy can be achieved by row and column level security. The row level security includes allowing or denying access to specific subset of data in data set based on users' role or group. Column level security can be achieved by mechanisms like full/partial/custom data masking, nullifying, hashing and encryption.

The current mechanism of supporting row and column level security in most of the relational databases is through views. The relational databases and data warehouses that don't support views are not able to provide row and column level security. Even for the relational databases and warehouses that support views, supporting row and column level security through views is a cumbersome and not scalable approach as separate views would be required for each user/role for which row and column level security is required.

IV.Proposed Solution

We thus put forth a proposal to have a data privacy framework that be used with the relational databases including Oracle, PostgreSQL, MySQL and with the big data warehouses like Hive. The framework will support the row and column security mechanisms for data privacy mentioned above and it will be generic so that it can work with both relational databases and data warehouses

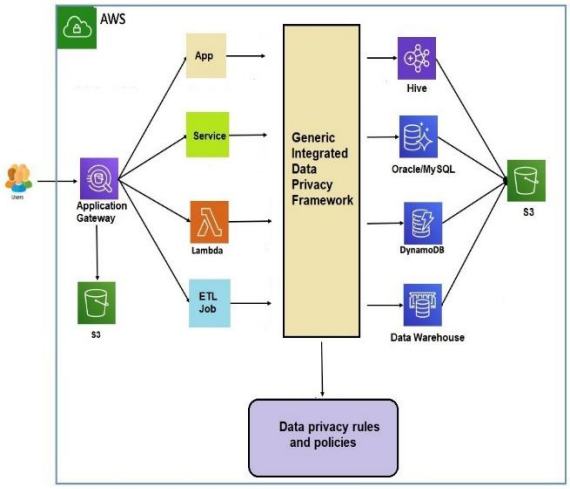
The current architecture of enterprise application is as shown below -



In this architecture there are multiple components at service layer like applications, services, lambda, or step functions and ETL jobs in a data center or over cloud that can interact with variety of relational databases like Oracle/MySQL, Dynamo DB or data warehouses

including big data warehouses like Hive at the data access layer.

We propose to have a generic integrated data privacy layer in between that is placed between the service layer and data access layer -



This generic integrated data privacy layer will use the data privacy rules and policies stored in an isolated data store and will ensure data privacy at row and column level.

The generic data privacy layer can be implemented as a database driver that can intercept the queries from service layer and can look into a data store that stores data privacy rules and policies and apply the row and column level filters based on access given to the user or user group or user role and return the filtered data to the service layer.

Following steps will be followed in the call from service layer to data access layer to get data from relational databases or data warehouses –

1. Service layer sends a query to the data access layer for retrieving some data as requested by user or application
2. The query is sent to the database driver. The generic data privacy layer is already integrated with the database driver
3. The database driver checks the user, user group and user role and fetches the data privacy rules and policies applicable for this user from a separate data store that stores the data privacy rules and policies
4. The database driver filters the rows and columns in the dataset returned by the relational database or data warehouse.

a. It will remove any rows for which the current user does not have access.

b. For columns multiple options like full masking, partial masking, nullifying, hashing, encryption etc. will be provided so that the column value is not visible for any columns to which current user does not have access.

It is also proposed that the query execution time will be used as a performance metric for the generic integrated data privacy layer and the degradation in query execution time should not exceed more than 10%.

V. Experimental Analysis

To be done

VI. Conclusion and Future Work

The current work has emphasized the need for a generic and integrated approach for achieving data privacy for relational databases and data warehouses. This approach checks for row and column level security based on user, role, or group.

The suggested approach to implement the generic integrated data privacy layer is to implement it as a database driver for various relational databases and data warehouses.

The next steps are to implement a reference implementation of database driver for relational database (MySQL) and for a data warehouse (Hive) and provide a working implementation of this approach. Functional testing will be done to ensure that row and column security based on user, role and group are working.

Also, performance testing would be done to ensure that the degradation in query execution time would not be more than 10%.

References

[1] D. Dayong, "Overview of Big Data and Hive," in Apache Hive Essentials, Packt Publishing, 2015. K. Shirudkar and D. Motwani, "Big Data Security," International Journal of Advance Research in Computer Science and Software Engineering, vol. 5, no. 3, pp. 1102-1105, 2015.

[2] D. S. Terzi, R. Terzi, and S. Sagiroglu, "A Survey on Security and Privacy Issues in Big Data," in The 10th International Conference for Internet Technology and Secured Transactions, 2015.

[3] V. N. Inukollu, S. Arsi and S. R. Ravuri, "Security Issues associated with Big Data in cloud computing," International Journal of Network Security & Its Applications (IJNSA), vol. 6, no. 3, pp. 51-55, 2014.

[4] Y. Tian, "Towards the Development of Best Data Security for Big Data," Communications and Networks, vol. 9, pp. 291-301, 2017.

[5] A. Gosain and A. Arora, "Security Issues in Data Warehouse: A Systematic Review," in International Conference on Intelligent Computing,

Communication & Convergence, Bhubneshwar, 2015.
[6] E. Fernandez-Medina, J. Trujillo, R. Villaroel and M. Piattini, "Access control and audit model for the multidimensional modeling," Decision Support System, Elsevier, vol. 42, pp. 1270-1289, 2005.
[7] R. K. Sharma and V. Kapoor, "Implementing Row and Column Level Security in Hive," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 6, no. 9, pp. 1329-1332, 2017.
[8] K. Michael and K. W. Miller, "Big Data: New Opportunities and New Challenges," IEEE Computer Society, pp. 22-24, 2013.
[9] F. Prasser, H. Spengler, R. Bild, J. Eicher and K. A. Kuhn, "Privacy-enhancing ETL-processes for biomedical data," International Journal of Medical Informatics, vol. 126, pp. 72-81, 2019.
[10] J. Meszaros, "The conflict between privacy and scientific research in the GDPR," in Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC), 2018.
[11] Curt Cotner and Roger Lee Miller, “Row level security in Database Management System”, US patent no. 9,870,483 B2