A Novel Approach based on Bucketization for Privacy Preserving Access Control Mechanism for Relational Data

Anuja A. Salunke
PG Student
SRESs College of Engineering
Kopargaon

A. B. Pawar, PhD
Department of Computer Engineering
SRESs College of Engineering
Kopargaon

ABSTRACT
In the todays world of digitalization of data, giving proper access of data to the users is a crucial part. Access control mechanisms are use to protect sensitive information from unauthorized users. Still there is chance of compromising the privacy of the person by authorized user which may lead to identity disclosure. In this paper we proposed a efficient method called bucketization to preserve the privacy of the user. To anonymize and satisfy privacy requirements PPM uses suppression and generalization of relational data like k-anonymity and l-diversity, against identity and attribute disclosure. However, privacy is achieved at the cost of precision of authorized information. The proposed system is implemented and compared with the state of art techniques studied in literature survey.

Keywords
Access control, privacy, k-anonymity, query evaluation

1. INTRODUCTION
Data mining is the process of extracting the useful information from the database. It is possible to efficiently extract or mine knowledge from large amounts of vertically partitioned data within quantifiable security restrictions. In other words the data mining is the process of discovering the interesting knowledge from large amounts of data stored either in databases, data warehouses or other information repositories. Knowledge Discovery in Databases (KDD) is the process of extracting knowledge from large quantities of data. The KDD process assumes that all the data is easily accessible through centralized access mechanisms such as federated databases and virtual warehouses. Moreover, advances in information technology and the ubiquity of networked computers have made personal information much more available. Privacy advocates have been challenging attempts to bring more and more information into integrated collections. Database security is the important requirements of the database. Database security is a very broad area that addresses many issues, like legal and ethical issues regarding the right to access certain information. Some information may be stored to be private and cannot be accessed legally by unauthorized persons. The sensitive data is accessible to authorized users only. The database security provides the security for the sensitive information from the unauthorized access. The database security is based on the access control mechanism and the privacy protection mechanism.

The Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users The sensitive information which is the user is not authorized to access will not accessed by the user. The authorized user can only access the authorized data. In a multiuser database system, the Database Management System (DBMS) must provide techniques to enable certain users or user groups to access selected portions of a database without gaining access to the rest of the database. Its importance comes in a large organization where numerous workers working. There must be some important data which are not published to all the workers. There uses the access controls mechanism for providing the access to the secured data to the particular authorized user only. For example, sore information like employee salaries or performance reviews should be kept secret from most of the database system‟s users. A DBMS typically includes a database security and authorization subsystem that is responsible for ensuring the security of portions of a database against unauthorized access. Privacy is the one of the most mportant concern of human life. It gives more importance to protect the privacy of the personal life. In the case of database, there will be huge amount of data to be kept privately. These data may contain sensitive information about the persons, confidential information about some organizations and so on. These data has to be protected by using some methods. It is the privacy protection mechanism (PPM). The general method is to transform the original data into some anonymous form to prevent from accessing its record owners‟ sensitive information. There are legion methods to render the privacy for the sore data. The anonymization method is one of the important privacy protection mechanisms. The anonymization procedure will transforms the sensitive information to some anonymized form. K-anonymity, l-diversity, etc., are some of the anonymization method. For a given query from unauthorized user, it will render the anonymized data through the privacy preserving methods.

In this paper we proposed bucketization method which deals with the privacy protected access control mechanism. It will provide the security for the sensitive information. For an example, in the case of hospital management system there should be a number of patients. Some of the patients may have the disease which has to be isolated and so on. While publishing the patients‟ data to the state medical board for disease surveillance system, they should anonymize the personal data of the patient. For this purpose it can use the proposed method for the secured access control and privacy protection mechanism.

2. LITERATURE SURVEY
Existing workloadaware anonymization techniques [1], [2] minimize the imprecision aggregate for all queries and therefore the imprecision added to every permission/query in the anonymized small information is not noted. making the privacy requirement more rigorous (e.g., increasing the worth of k or l) ends up in further inexactitude for queries. However, the problem of satisfying accuracy constraints for individual permissions in a policy/workload has not been studied before.

The heuristics projected during this paper for accuracy-constrained privacy-preserving access control also are relevant in the context of workload-aware anonymization. The anonymization for continuous data publishing has been studied in literature [3].

In the area of discretionary access control models for computer database systems, a crucial early contribution was the event of the System R access management model [4], [5], that powerfully influenced access management models of current commercial relative DBMSs. Some key features of this model enclosed the notion of suburbanised authorization administration, dynamic grant and revoke of authorizations, and therefore the use of views for supporting content-based authorizations. Also, the initial format of well-known commands for grant and revoke of authorizations, that are nowadays a part of the SQL normal, were developed as a part of this model. Later analysis proposals have extended this basic model with a variety of options, such as negative authorization [6], role-based and task-based authorization [7], temporal authorization [8], and context-aware authorization .

Author is [9] projected l-diversity idea that return from this method. k-anonymity is vulnerable to homogeneity attacks once the sensitive price for all the tuples in associate degree equivalence category is that the same. To counter this defect, l-diversity has been projected and needs that every equivalence category of T * contain a minimum of l distinct values of the sensitive attribute. For sensitive numeric attributes, an l-diverse equivalence category will still leak data if the numeric values are on the brink of one another. For such cases, variance diversity has been projected that needs the variance of every equivalence category to be larger than a given variance diversity parameter.

The partitions created by TDSM have dimensions on the median of the parent partition. A compaction procedure has been projected in [10] wherever the created partitions square measure replaced by minimum bounding boxes. This step improves the precision of the anonymized table for any given query workload by reducing the overlapping partitions.

3. PROPOSED SYSTEM

The proposed system consist of 5 modules as follow

1. Access control policy

2. Anonymity

3. Anonymization with impresision Bounds

4. Accuracy-Constrained Privacy-Preserving Access Control

5. Top-Down Heuristic

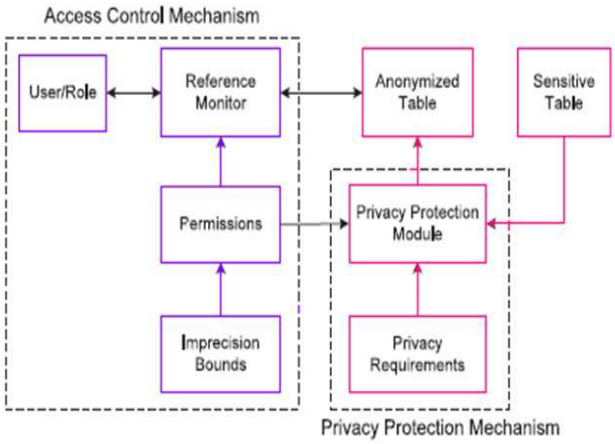
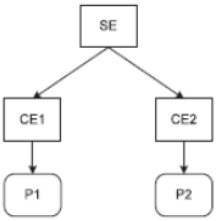


Fig 1: Proposed Architecture

3.1 Access Control Policy

Role	Designation
SE	State Epidemiologist
CE1	County 1 Epidemiologist
CE2	County 2 Epidemiologist



Permission	Authorized Query Predicate
P1	Location = County 1 ∧ Age = 15-65 ∧ Syndrome = Influenza
P2	Location = County 2 ∧ Age = 15-65 ∧ Syndrome = Influenza

Fig 2: Access Control Policy

Syndromic surveillance systems are used at the state and federal levels to detect and monitor threats to public health. The department of health in a state collects the emergency department data (age, gender, location, time of arrival, symptoms, etc.) from county hospitals daily. Generally, each daily update consists of a static instance that is classified into syndrome categories by the department of health. Then, the surveillance data is anonymized and shared with departments of health at each county. An access control policy is given in Fig. 2 that allows the roles to access the tuples under the authorized predicate, e.g., Role CE1 can access tuples under PermissionP1. The epidemiologists at the state and county level suggest community containment measures ,e.g., isolation or quarantine according to the number of persons infected in case of a flu outbreak. According to the population density in a county, an epidemiologist can advise isolation if the number of persons reported with influenza are greater than 1,000 and quarantine if that number is greater than 3,000 in a single day. The anonymization adds imprecision to the query results and the imprecision bound for each query ensures that the results are within the tolerance required. If the imprecision bounds are not satisfied then unnecessary false alarms are generated due to the high rate of false positives.

3.2 Anonymity

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	Diarrhea
4	25	15	Fever
5	22	28	Flu
6	32	35	Fever
7	38	32	Flu
8	35	25	Diarrhea

(a) Sensitive table

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	0-20	10-30	Flu
2	0-20	10-30	Fever
3	20-30	10-30	Diarrhea
4	20-30	10-30	Fever
5	20-30	10-30	Flu
6	30-40	20-40	Fever
7	30-40	20-40	Flu
8	30-40	20-40	Diarrhea

(b) 2-anonymous Table

Fig 3: Anonymity

Anonymity is prone to homogeneity attacks when the sensitive value for all the tuples in an equivalence class is the same. To counter this shortcoming, l-diversity has been proposed and requires that each equivalence Fig. 3 Access control policy. class of T contain at least l distinct values of the sensitive attribute. For sensitive numeric attributes,an l-diverse equivalence class can still leak information if the numeric values are close to each other. For such cases, variance diversity has been proposed that requires the variance of each equivalence class to be greater than a given variance diversity parameter. The table in Fig. 3a does not satisfy k-anonymity because knowing the age and zip code of a person allows associating a disease to that person. The table in Fig. 3b is a 2-anonymous and 2-diverse version of table in Fig. 2a. The ID attribute is removed in the anonymized table and is shown only for identification of tuples. Here, for any combination of selection predicates on the zip code and age attributes, there are at least two tuples in each equivalence class

3.3 Anonymization With Impresision Bounds

we formulate the problem of k-anonymous Partitioning with Imprecision Bounds and present an accuracy-constrained privacy-preserving access control framework. Imprecise data means that some data are known only to the extent that the true values lie within prescribed bounds while other data are known only in terms of ordinal relations. Imprecise data envelopment analysis (IDEA) has been developed to measure the relative efficiency of decision-making units (DMUs) whose input and/or output data are imprecise. In this paper, we show two distinct strategies to arrive at an upper and lower bound of efficiency that the evaluated DMU can have within the given imprecise data. The optimistic strategy pursues the best score among various possible scores of efficiency and the conservative strategy seeks the worst score. In doing so, we do not limit our attention to the treatmentof special forms of imprecise data only, as done in some of the studies associated with IDEA. We target how to deal with imprecise data in a more general form and, under this circumstance, we make it possible to grasp an upper and lower bound of efficiency.

3.4 Accuracy-Constrained Privacy-Preserving Access Control

An accuracy-constrained privacy-preserving access control mechanism. (arrows represent the direction of information flow), is proposed. The privacy protection mechanism ensures that the privacy and accuracy goals are met before the sensitive data is available to the access control mechanism. The permissions in the access control policy are based on selection predicates on the QI attributes. The policy administrator defines the permissions along with the imprecision bound for each permission/query, user-to-role assignments, and role-to permission assignments. The specification of the imprecision bound ensures that the authorized data has the desired level of accuracy. The imprecision bound information is not shared with the users because knowing the imprecision bound can result in violating the Privacy requirement. The privacy protection mechanism is required to meet the privacy requirement along with the imprecision bound for each permission.

3.5 Top-Down Heuristic

In TDSM, the partitions are split along the median. Consider a partition that overlaps a query. If the median also falls inside the query then even after splitting the partition, the imprecision for that query will not change as both the new partitions still overlap the query as illustrated. In this heuristic, we propose to split the partition along the query cut and then choose the dimension along which the imprecision is minimum for all queries. If multiple queries overlap a partition, then the query to be used for the cut needs to be selected. The queries having imprecision greater than zero for the partition are sorted based on the imprecision bound and the query with minimum imprecision bound is selected. The intuition behind this decision is that the queries with smaller bounds have lower tolerance for error and such a partition split ensures the decrease in imprecision for the query with the smallest imprecision bound. If no feasible cut satisfying the privacy requirement is found, then the next query in the sorted list is used to check for partition split. If none of the queries allow partition split, then that partition is split along the median and the resulting partitions are added to the output after compaction.

Top down Heuristic Algorithm

1) Step 1. Initialize the set of candidate partition.

2) Step 2. Sort the queries overlapping the candidate partition with imprecision greater than zero.

3) Step 3. Select the queries with least imprecision bound.

4) Step 4. Checks for the feasible split of the partition along the query interval.

5) Step 5. If a feasible cut is found, then the resulting partitions are added to the candidate partition.

6) Step 6. If feasible cut is not found, then the candidate partition is checked for the median cut.

The heuristic algorithm will helps to provide the secured access control mechanism. The imprecision bound is set by the administrator. The imprecision bound is not known to the user. So it provides the secured access control method.

4. RESULTS

The proposed system that combines the idea of secured access control mechanism and privacy protection mechanism for the relational data. The heuristic algorithm along with bucketization used here improves the efficiency of the access control mechanism. The combination of the anonymization and fragmentation used here it has improved the privacy of the sensitive information in the relational data.

The below result shows that how the data is sliced for the user and only needed fields are shown and confidential data is kept hidden.

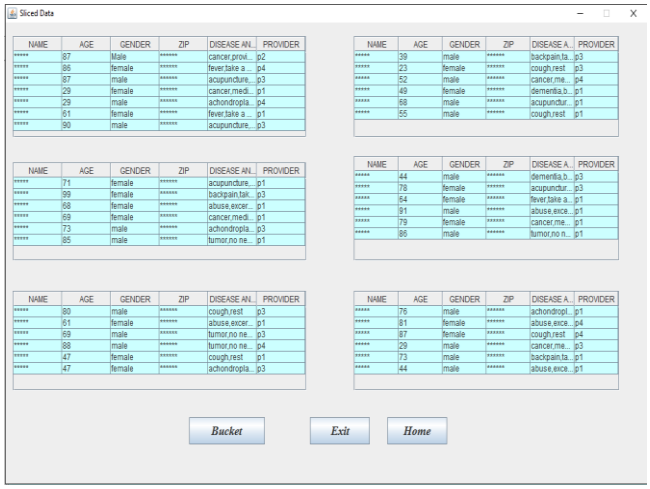


Fig 4: Result

5. CONCLUSION

In secured relational data storage, it needs good access control mechanism and privacy preserving access control mechanism. In this paper a privacy-preserving access control framework for relational data has been proposed. The proposed framework is a combination of access control and privacy protection mechanisms. The access control mechanism allows only the authorized query predicates on sensitive data. The privacy preserving module anonymized and fragmented the data to meet privacy requirements and imprecision constraints on predicates set by the access control mechanism. For the anonymization process proposed a k-anonymity method and for the fragmentation introduces the clustering analysis method. It formulates this interaction as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB). It gives hardness results for the k-PIB problem. This paper presents a heuristics method for partitioning the data to satisfy the privacy constraints and the imprecision bounds using bucketization. In the current work, static access control and relational data model has been assumed. For future work, it plan to extend the proposed privacy-preserving access control to cell level access control and can use the l-diversity instead of k-anonymity method.

6. REFERENCES

[1] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization Techniques for Large-Scale Datasets," ACM Trans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.

[2] T. Iwuchukwu and J. Naughton, "K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 746-757, 2007.

[3] B. Fung, K. Wang, R. Chen, and P. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, article 14, 2010

[4] P.G. Griffiths and B. Wade, "An Authorization Mechanism for a Relational Database," ACM Trans. Database Systems, vol. 1, no. 3, pp. 242-255, 1976.

[5] R. Fagin, "On an Authorization Mechanism," ACM Trans. Database Systems, vol. 3, no. 3, pp. 310-319, 1978.

[6] E. Bertino, S. Jajodia, and P. Samarati, "An Extended Authorization Model," IEEE Trans. Knowledge and Data Eng.,vol. 9, no. 1, pp. 85-101, 1997

[7] R. Sandhu, E.J. Coyne, H.L. Feinstein, and C.E. Youman, "Role-Based Access Control Models,"Computer,vol. 29, no. 2, pp. 38-47, 1996

[8] E. Bertino, C. Bettini, E. Ferrari, and P. Samarati, "An Access Control Model Supporting Periodicity Constraints and Temporal Reasoning,"ACM Trans. Database Systems,vol. 23, no. 3, pp. 231-285, 1998.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubrama- niam, L-Diversity: Privacy Beyond k-anonymity, ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

[10] T. Iwuchukwu and J. Naughton, K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization, Proc. 33rd Intl Conf. Very Large Data Bases, pp. 746-757, 2007.