Collaborative System for Scheduling the Cloud Services in MVM

*S.Sarathkumar, #C.Vinoth,M.E
*PG Scholar DEPT OF CSE
#Supervisor Name: AP/CSE
King College of Technology, Namakkal

Abstract

Data mining technique has become interesting because of identifying patterns and trends from large collections of data. It is known that the collection and analysis of data that includes personal information of an individual may violate privacy of to whom the information refers. After obtaining this data pattern it is shared among various organizations to improve its services. While sharing the information among the organizations, it may be misused. To overcome this, Access Control Mechanism is used to protect sensitive information from unauthorized user. But still privacy preserving mechanisms is not applied to ensure the privacy of shared information. To ensure the privacy, Top-Down heuristic is applied with kd -tree partitioning. This approach is limited only to static data release not for incremental data. That is it is assumed that a whole dataset is available at the time of data release. To address this issue anonymization approach with generalization and range tree partitioning is applied on incremental data which can enhance security and privacy policies.

Index Terms – *Access Control, Privacy, Mondrian, Imprecision Bound*.

I. INTRODUCTION

Privacy is defined as the ability of an individual or a group to separate information about themselves, and thereby publish it selectively. The privacy is needed whenever the personal information is collected, or the databases are created which contains the personal information, or the data is shared which contains sensitive information.

Nowadays, the data are being collected and used increasingly. Modern business increases the need for sharing, querying and mining data across multiple autonomous enterprises.Most of the collected data are person specific, containing a record for each individual person. The organizations which collect such data often need to publish and share the data for multiple purposes. Sharing is essential to the missions of many organizations. But data sharing requires balancing much privacy, security, and legal interests. The published or shared data usually contains personal sensitive information. The sharing of personal information causes an increasing number of legal, monetary, practical, and privacy issues.

Access Control Mechanism is used to ensure the security of shared data. It uses Role-Based Access Control which controls access to shared or collected to achieve the security. In role-based access control, access decisions are based on the roles that the individual person have as part of an organization. Access rights are grouped by role name, and the use of resources is restricted to individuals authorized to assume the associated role.

Privacy can be protected by using Privacy Protection Mechanism. Anonymization is used to achieve the privacy in Privacy Protection Mechanism. Anonymization is the method of permanently and completely removing personal identifiers from data, such as converting personally identifiable information into aggregated data. When information is disclosed for sharing, the information is should be anonymized. The sensitive information of individuals cannot be identified easily when the data to be shared is fully anonymized.Anonymized data is no longer being associated with an individual in any manner.The goal of anonymization is without violating privacy of any individual, take the advantage of data. So that intruder should be unable to use the published data.

II. LITERATURE REVIEW

A. Access Control

Virtual Private Database (VPD) in oracle is for security purpose. It controls the access to data at the row level and ties the security policy to the table itself. Virtual Private Database limits access to data in certain tables for all users, as required by a corporate policy. For example a company involved in trading might need to limit access to certain tables so they are only accessible during trading hours. It also helps lower the cost of development, by building security once, in the data server, instead of in every application that accesses the data [1].

In Truman model, query modification approach is generalized by using parameterized view framework. The Truman security model provides the personal and restricted view of the complete database for each user. The user queries are modified in the way that the user does not get to see anything more than his/her view of the database.

In Non-Truman model, the query is subjected to validity test, failing which the query is rejected and user notified about this. If the query passes the test, it is to allow execute normally, without modification. If the query can be answered using the information contained in the authorization views available to the user, then the query is said to be valid [2].

Role Based Access Control Standardization which defines sets of basic RBAC elements, relations and functions. The RBAC standard has two main functions: (i) RBAC Reference model which defines a common vocabulary terms and to set the scope of the RBAC features included in the standard. (ii) The RBAC Functional Specification defines requirements over administrative operations for the creation and maintenance of RBAC element sets and relations [3].

B. Anonymization

Anonymization is used to provide privacy to the data. A data release is said to satisfy *k*-anonymity if every tuple released cannot be related to fewer than k respondents. For example, *k*-anonymity with k=2 was provided in the release, each released record could indistinctly belong to at least two individuals. *k*-anonymity can be achieved by using two techniques called generalization and suppression.

Each generalization operation replace values of specific description, typically the quasi identifier attributes, with less specific attributes. For categorical attribute, a specific value can be replaced with a general value. Suppression replaces some values with special value, indicating that replaced values are not disclosed. For suppression different schemes are available. In record suppression, it suppresses entire record. Value suppression, it suppresses every instance of a given value in table and Cell suppression, it suppresses some instances of a given value in Table[4].

Amulti dimensional partitioning approach for *k*-anonymization is used to produce a better quality results. It partitions the domain into ranges rather than generalizing the values. This can be done for attributes which have a totally ordered domain. It chooses a quasi-identifier that is split attribute for partition.

If the split attribute is numeric, then it chooses a binary split threshold (e.g., Age <=40). If the split attribute is categorical, then the split is defined by specializing a user defined generalization hierarchy. Under *k*-anonymity, a split is allowable if each subset contains at least k tuples [5].

kd-tree is space- partitioning data structure for organizing points in a k-dimensional space. kd-trees are useful for multidimensional search. In kd - tree, the given points or tuples are splitted into two subsets of roughly equal size. One subset contains points smaller than equal to the splitting value; the other subset contains points greater than the splitting value. The splitting value is stored at the root and two subsets are stored recursively in the two subtrees [6].

III. ACCESS CONTROL MECHANISM

An Access Control Mechanism is used to provide access to the various people who are all working in an organization. In Access Control Mechanism two concepts are used to accomplish the security of the data. They are permission and Imprecision Bound.

A. Permission

Permission is same as role based access control. In role based access control, the role is defined for each user in the organization. The role hierarchy is created for assigning roles to the user. The role hierarchy defines an inheritance relationship among roles. It Aggregates permissions and implicitly assigns users to roles. Based on the role, the permission is delineated for each user. In access control policy, permissions are based on selection predicates on the QI attributes.

After the anonymization, the exact tuple values in a relation are replaced by the generalized values. So the relaxed access control enforcement is defined over the generalized data. In relaxed enforcement, the overlap semantics are used to allow access toall partitions that are overlapping the permission. Authorization predicates are violated, by giving access to extra tuples. But it is beneficial for applications where low cost of a false alarm isbearable as compared to the risk associated with a missed event.

B. Imprecision Bound

The imprecision bound is preset by policy administrator, who is the head of the organization. Imprecision Bound defines a threshold, which is the amount of imprecision that can be tolerated. When the anonymization performed, it adds imprecision to the query results. If the rate of false positives is high, then unnecessary false alarms are generated.

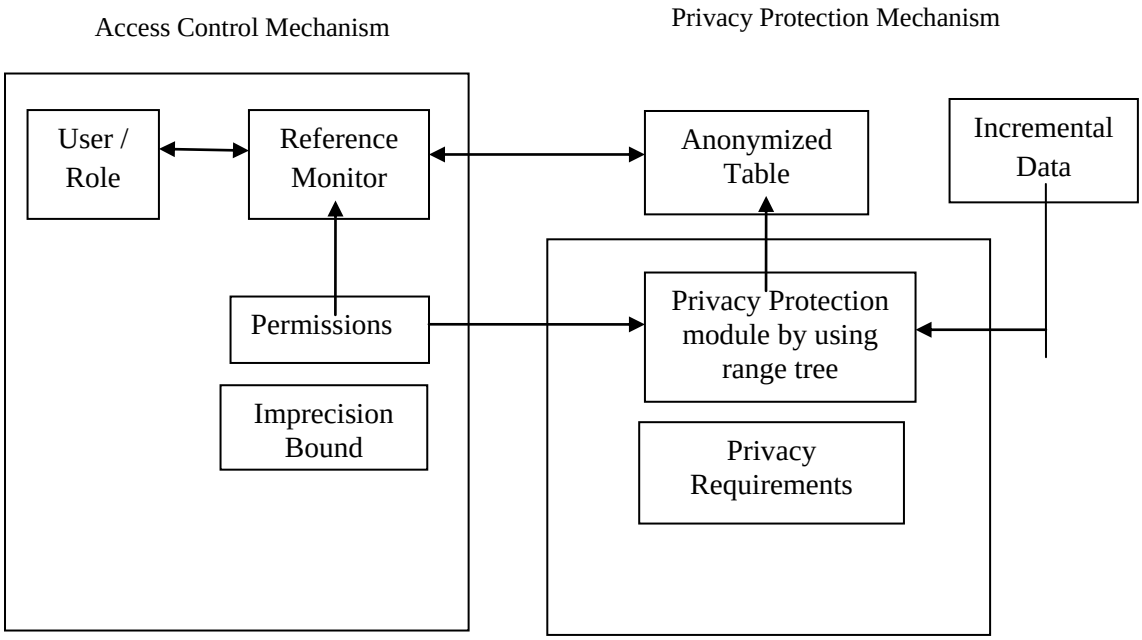

Figure 1 Accuracy – Constrained Privacy Preserving Access Control Mechanism

IV. PRIVACY PROTECTION MECHANISM

When organization shares data, they must do so in a way that fully protects individual privacy. The privacy protection mechanism is used to guarantee the confidentiality or privacy of the data by using anonymization. The anonymization techniques such as generalization and suppression are used to ensure the *k*-anonymity. But generalization and suppression are decrease the quality or utility of data. To overcome this, Top Down Selection Mondrian algorithm is used to achieve the k-anonymity.

A. Top Down Selection Mondrian Algorithm (TDSM)

To minimize the total imprecision for all queries TDSM is used. In TDSM partitions are divided recursively until the new partitions are satisfy the privacy requirement.

The partitions are divided, by making two decisions. They are,

- The split value is selected along each dimension
- The dimension is selected along which to split

The split value is selected along the median and then the dimension is chosen along which the sum of imprecision isminimum for all queries.In TDSM, thed-dimensional quasi-identifier attribute domain space is divided into non overlapping partitions. After that, the d-dimensional vectorisreplaced by the intervalsof the partition for each tuple to which the tuple belongs. A tuple is belongs toa partition, if it's within that interval [7].

B. Topdown Heuristic 2

Topdown heuristic 2 is used for partitioning the whole tuple space in the given relation. In this, the query bounds are updated as partitions are added to the output. This update is carried out by subtracting the imprecision cost from the imprecision bound. For example, if a partition of size k has imprecision 10 for Query Q1 with imprecision bound 200, then the bound is changed to 190.

If the kd-tree traversal is depth-first, then it will achieve better result. Itensures that a given partition is recursivelysplit till the leaf node is reached. Initially, this approach favours queries with smaller bounds. As more partitions are added to the output, all the queries are treated fairly. If any query violates the imprecision bounds then, that query is put into low priority.

C. Query Cut

If the query is given, the query intervals are used to split the partitions that are defined as query cuts. The query cut splits the partition based on the query interval values.For a query cut using Query, both the startof the query interval and the end of the query interval are considered to split a partition alongthe dimension.

D. Query Imprecision

Query Imprecision is defined as the difference between the number of tuples returned by a query evaluated on an anonymized relation T* and the number of tuples for the same query on the original relation T. [8]

The imprecision for query Qi is denoted by imp_{Qi},

$$imp_{Qi} = |Q_i(T^*)| - |Q_i(T)|$$

$$|Q_i(T^*)| = \sum_{\text{EC overlaps } Q_i} |\text{Equivalence Class}|$$

E. Incremental Data

Real world data sources are often incremental. Data are added in the database at episodic manner. Whenever the incremental data is being anonymized, privacy should be attained. For ensuring the privacy, generalization of the data is increased.A better approach to anonymize the incremental data is anonymizing the whole dataset whenever the data is added to the dataset. The major issue in anonymizing incremental data is the same data may be anonymized and published multiple times, each of the time in different form. This may enable a various type of inferences [9].If any value in the table is inferrable means, then the level of generalization is increased. The increased generalization should satisfy privacy requirement and also it should not reveal any information about individual.

F. Range Tree

The dataset size is kept on increasing, when new data is being added. kd -tree is not suitable for partitioning the large datasets and query processing time is also high compared to range tree. Range tree is used to minimize the query processing time when the number records are very large in the given dataset. kd -tree stores the values based on x co-ordinates but range treestores the values based on x and y co-ordinates [6]. So the query processing time is decreased.

V. RESULTS

By using Accuracy-constrained privacy preserving access control mechanism, the privacy and security of the incremental data might be achieved. Instead of using kd -tree, range tree is used for partitioning which might be minimizing the query processing time.

REFERENCES

[1] K. Browder and M. Davidson (2002) 'The virtual private database in oracle9ir2', Oracle Technical White Paper, Oracle Corporation, vol. 500.

[2] S. Rizvi, A. Mendelzon, S. Sudarshan, and P. Roy (2004) 'Extending query rewriting techniques for fine-grained access control', Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pp. 551–562, ACM.

[3] D. Ferraiolo, R. Sandhu, S. Gavrila, D. Kuhn, and R. Chandramouli (2001) 'Proposed NIST standard for role-based access control', ACM Transactions on Information and System Security, vol. 4, no. 3, pp. 224–274.

[4] P. Samarati (2001) 'Protecting respondents' identities in micro data release', IEEE Transactions on Knowledge and Data Engineering, pp. 1010–1027.

[5] K. LeFevre, D. DeWitt, and R. Ramakrishnan (2006) 'Mondrian multidimensional k-anonymity', in Proceedings of the 22nd International Conference on Data Engineering, pp. 25-25, IEEE.

[6] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars (2008) 'Computational Geometry Algorithms and Applications', Springer-Verlag Berlin Heidelberg, vol. 3, pp. 95-115.

[7] ZahidPervaiz, Walid G. Aref,ArifGhafoor , and NagabhushanaPrabhu (2012) 'Privacy-preserving Access Control',CERIAS Tech Report.

[8] ZahidPervaiz, Walid G. Aref, ArifGhafoor, NagabhushanaPrabhu (2013), 'Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data', IEEE Transactions on Knowledge and Data Engineering, Issue 99.

[9] Ji-Won Byun, Tiancheng Li, Elisa Bertino, Ninghui Li, YonglakSohn (2009) 'Privacy preserving incremental data dissemination', Journal of Computer Security, vol. 17, pp. 43-68.

[10] A. Frank and A. Asuncion(2010) 'UCI Machine Learning Repository' http://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD).