SECURITY AND PRIVACY OF RELATIONAL DATA USING ACCESS CONTROL AND PRIVACY PROTECTION MECHANISM

Limy Sebastian and Panchami V.
Computer Science and Engineering, Toc H Institute of Science and Technology, Ernakulam, India
E-Mail: limysebastian@gmail.com

ABSTRACT

Authentication and authorization are two pivotal security mechanisms generally implemented to enforce security both on data and resource levels in computer applications, especially over the internet. Once a user is authenticated, application starts a session for the user. But authentication does not mean license for anything for user. Several application resources and data sources are further secured by a Role based Access Control mechanisms. The sensitive and confidential information in many organizations is shared by the means of authorized access. But an authorized user can still compromise the privacy of persons leading to an unwanted identity disclosure. An anonymization technique is normally employed to suppress identifiable attributes yet capable of sharing information. Anyway some accuracy of data is compromised for the sake of privacy protection. In this paper a framework is developed which puts an additional aspect of accuracy constraint for multiple roles. Once the access control and data anonymization techniques are integrated they work together as a service for any application as a configurable privacy preserving role based access control framework.

Keywords: authentication, authorization, access control, anonymization, accuracy.

1. INTRODUCTION

Companies and organizations collect and use large amount of sensitive user data whose release must be carefully controlled. The phenomenal advance in information technology over the past few decades has literally transformed our lives. Particularly, the explosive growth of the Internet and e-commerce has enabled people to carry out daily activities online, for example, online shopping, e-banking and even consulting a doctor over the Internet. Such prevalent online activities imply that a vast amount of personal data is electronically produced and collected continuously. Such collected data represent an important asset today as they can be used for various purposes ranging from scientific research to demographic trend analysis or marketing purposes or in syndrome surveillance etc. The collection and use of personal data is accepted as a common business practice today. But this trend raises a significant concern for information privacy. Individuals are afraid that their personal information might fall into wrongful hands and be abused against their will.

Information privacy implies much more than confidentiality of personal information. The problem of information privacy is not about how to conceal critical information, but how to ensure that such information is disclosed only under appropriate circumstances. Thus, a complete solution to information security must meet the following three requirements: secrecy or confidentiality, which refers to the protection of data against unauthorized disclosure, integrity, which refers to the prevention of unauthorized and improper data modification, and availability, which refers to the prevention and recovery from hardware and software errors and from malicious data access denials making the database system unavailable. These three requirements arise in practically all application environments.

Today, the protection of data that is entrusted to enterprise information systems is more challenging than ever. There is an increased focus by industry and research towards improving security of cyber infrastructures. The data becomes more valuable with the increasing quality of data. The potential to be gained from unauthorized access and the potential damage that can be done if the data is corrupted increases as the data becomes more valuable.

Any access to sensitive information, in the traditional environment was through employees. Employees are not always reliable but they are known and their access to sensitive information is mainly limited by the function they perform. The environment changed drastically, when the activities are moved to the Internet. The off shoring of data management functions result in companies knowing little or nothing about the users accessing their systems. There is also a rise in the trend of ubiquitous computing. Due to this, data must be available to users anywhere and at anytime. Because of all these increased risks, adequate protection of information systems is needed [1]. All the above motivations are strong reasons for the adoption of solutions that are more comprehensive and articulated than the ones available today.

The rest of the work is organized as follows. In Section II, relevant background is discussed. The various privacy requirements and how to implement the privacy requirements are also explained. Section III covers the proposed system which combines access control and privacy protection mechanisms. In Section IV implementation of the proposed work is explained and Section V concludes the paper.

2. BACKGROUND

In this section access control and data anonymization are discussed. The various privacy requirements are also explained.

A. Access control

Access Control Mechanisms are used to ensure that only authorized information is available to users.

Access control constraints what a user can do directly, as well as what programs executing on behalf of the users are allowed to do. In this way access control tries to prevent activities that could lead to a breach of security [2]. The three different access control policies that commonly occur in computer systems: Discretionary Policies, Mandatory policies and Role Based policies.

Discretionary policies govern the access of users to the information on the basis of the user's identity and authorizations (or rules) that specify, for each user (or group of users) and each object in the system, the access modes (e.g., read, write or execute) the user is allowed on the object. The request that the user makes to access an object is checked against a set of authorizations that are specified. If there exist an authorization stating that the user can access the object in a particular specific mode, the access is granted. Otherwise the access is denied [2].

Mandatory policies govern access on the basis of classification of subjects and objects in the system. Each user and each object in the system is assigned a security level. The security level associated with an object reflects the sensitivity of the information contained in the object. Sensitivity of the information means the potential damage that could result from unauthorized disclosure of the information. The security level associated with a user is called clearance. It reflects the user's trustworthiness not to disclose sensitive information to users not authorized to view it. Each security level is said to dominate itself and all others below in this hierarchy. Access to an object by a subject is granted only if some relationship (depending on the type of access) is satisfied between the security levels associated with the two. In particular, the following two principles are required to hold: Read down (a subject's clearance must dominate the security level of the object being read and Write up (a subject's clearance must be dominated by the security level of the object being written).

Role-based policies regulate users' access to information on the basis of the user executes on the system. Role-based policies require the identification of roles in the system. A role can be defines as a set of actions and responsibilities associated with a particular working activity. Instead of specifying all the accesses, each user is allowed to execute access on objects that are specified based on roles. The user who is assigned a role is allowed to execute all the accesses for which the role is authorized. The same role can be played by several users, perhaps simultaneously [3].

B. Data anonymization

In the past sharing and dissemination of information was mostly in tabular and statistical form, known as macrodata. But many situations now require that the specific stored data themselves, called microdata, be released. To protect the privacy of the individuals to which the data refer and also to prevent attribute disclosure, released data are generally modified by removing certain identifiers such as name, phone numbers etc. The techniques that are used are known as data anonymization techniques. Data anonymization is used to preserve privacy of data. Data anonymity is particularly important in public databases such as health records or census data collected by government agencies, in situations where an organization wishes to allow third parties to access its customer data. In such cases, it cannot be guaranteed that the third parties respect the privacy policy of the data. Thus it is the responsibility of the organization to assure customers' privacy by removing all information that can link data items with individuals.

Data anonymization can be defined as the process of encrypting or removing personally identifiable information (i.e. information that can be used on its own or with other information to identify an individual) from data sets, so that the individuals to whom the data describe remain anonymous. Anonymity is an important concept for privacy. Anonymization algorithms use suppression and generalization of records to satisfy privacy requirements with minimal distortion of data. Suppression means to remove data from the table so that they are not released. That is, while performing suppression of attributes, certain attribute values are replaced by a '*' or any other special character. Generalization is achieved by replacing certain attribute values with broader categories [4].

Given a relational database $R = \{A_1, A_2, \ldots, A_n\}$, where A_i is an attribute. R* is the anonymized version of the relation R. We assume that R is an incremental database table into which data can be added by the administrator. The attributes in the database can be of the following types:

- **Identifier:** Attributes that can uniquely identify an individual. These attributes are suppressed in the anonymized relation. E.g., name and social security.
- **Quasi-identifier (QI):** Attributes that can identify an individual based on other information available to an adversary. QI attributes are generalized. E.g., gender, zip code, birth date.
- **Sensitive attribute:** Some attributes cause a privacy breach when associated with a unique individual. Such attributes are called sensitive attributes. E.g., disease or salary.

A key difficulty of data anonymization comes from the fact that data utility (i.e., data quality) and data privacy are conflicting goals. Data privacy can be enhanced by hiding more data values, but it inevitably decreases data utility. Revealing more data values increases data utility, but it may decrease data privacy.

C. Privacy requirements

The anonymized data that is released is still prone to certain attacks. Some attributes from the released anonymized data can be combined with external publicly available data to re-identify the individuals. Such attacks are known as linking attacks. To perform such linking attacks, the attacker needs two pieces of prior knowledge: the victim's record in the released data and the quasi-identifier of the victim [5]. k- anonymity has been proposed by Samarati et al. [6] to reduce the risk of this type of attack. In k-anonymity, the data privacy is guaranteed by ensuring that any record in the released data is indistinguishable from at least (k-1) other records with respect to a set of attributes called quasi-identifier. In other words it can be said that, a data release is said to satisfy k-anonymity if every tuple released cannot be related to fewer than k other tuples, where k is a positive integer set by the data holder [7]. These k tuples form an equivalence class. That is, an equivalence class is a set of tuples having the same QI attribute values. Generalization and suppression are two different approaches for obtaining, from a given table, a table that satisfies k-anonymity. These two approaches produce the best results when applied jointly. These two approaches provide protection, but some information is lost on the data being released. Generalization leads to loss of accuracy, as the information released is less precise. Suppression leads to loss of completeness as some information is being removed.

Certain attacks are still possible on the k-anonymous dataset. An attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. In other words, k-anonymity can create groups that leak information due to lack of diversity in the sensitive attribute. Suck attacks are known as homogeneity attacks. Another attack that is possible is the background knowledge attack. Attackers often have background knowledge. An adversary can use background knowledge to discover sensitive information and k-anonymity does not guarantee privacy against suck attackers. Thus a stronger definition of privacy is needed. To avoid suck attacks, the notion of ℓ-diversity is used. A q-block is said to be ℓ-diverse if it contains at least ℓ different values for the sensitive attribute [8]. With this additional constraint, the homogeneity attack is no longer applicable because each block has a set of ℓ distinct values of sensitive attributes. The background knowledge attack also becomes more complicated as the value of ℓ increases as the attacker needs more knowledge to isolate a unique value associable to a predefined entry.

D. Implementing the privacy requirements

An algorithm to implement the privacy requirements have been proposed by LeFevre et al. [11]. This algorithm using a multidimensional partitioning approach to partition the whole tuple space into a number of equivalence classes with identical quasi identifier attribute values. That is, consider a d-dimensional QI domain space. The idea of this algorithm is to divide this domain space into non overlapping rectangular regions. This partitioning is then used to define a global recoding function ($\alpha: D_{Q1} \times \ldots \times D_{Qd} \rightarrow D^d$) that maps each tuple in the domain to the region in which it is contained. α is then applied to the input relation R to produce R*. A partitioning, after applying α to the QI attributes, is said to be allowable with respect to an input relation R, if the relation R* satisfies the anonymity requirements.

The proposed algorithm is based on greedy recursive partitioning. The input of the recursive procedure is d-dimensional rectangular domain and a set of tuples, R. The algorithm first chooses a quasi identifier split attribute. The algorithm also chooses a binary split threshold (e.g., Age ≤ 50; Age > 50) if the split attribute is numeric. In case of categorical attributes, the split is defined by specializing a user defined generalization hierarchy.

The split attribute (and threshold) define a division of the input domain into m non-overlapping regions that cover the input domain. The split also defines a corresponding partitioning of the input data (R) into disjoint subsets, R_1, …, R_m. The split is said to be allowable if each R_i satisfies the given anonymity requirements. The procedure is executed recursively on each resulting partition (R_i) until there no longer exists an allowable split. For example, under k-anonymity, a split is allowable if each R_i contains at least k tuples.

3. PROPOSED WORK

Traditionally, research in the database community in the area of data security can be broadly classified into two- access control research and data privacy research. The idea of access control is to authorize a user to access only a subset of the data. This authorization is enforced by explicitly rewriting queries to limit access to the authorized subset. The main limitation of traditional access control mechanism in supporting data privacy is that it is "black and white" [10]. That is, the access control mechanism offers only two choices: release no aggregate information thereby preserving privacy at the expense of utility, or release accurate aggregates thus risking privacy breaches for utility. Thus, a hybrid system is needed that combines a set of authorization predicates restricting access per user to a subset of data and privacy preserving mechanism.

In the proposed system, a relational table, containing sensitive information, is taken. This table is anonymized. The database contains incremental data, with the administrator having the permission to add data into the table. The table has to be anonymized each time data is added into the database. Role based access control is being used here. The concept of role-based access control (RBAC) began with multi-user and multi-application on-line systems. The central notion of RBAC is that permissions are associated with roles and users are assigned to appropriate roles. This greatly simplifies management of permissions. Roles are created for the various job functions in an organization and users are assigned roles based on their responsibilities and qualifications. Users can be easily reassigned from one role to another [3]. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k-anonymity or ℓ-diversity [6].

Another constraint that needs to be satisfied by the privacy protection mechanism is the imprecision bound for each selection predicate. Query imprecision is defined as the difference between the number of tuples returned by a query evaluated on an anonymized relation R* and the number of tuples for the same query on the original relation R [9]. The imprecision bound for each permission define a threshold on the amount of imprecision that can be tolerated. If imprecision bound is not satisfied, then unnecessary false alarms are generated due to high rate of false positives. The imprecision bound is preset by the administrator and this information is not shared with the users because knowing the imprecision bound can result in violating the privacy requirement. The imprecision bound will be different for the different roles that exist within the organization. So, in a nutshell, it can be said that the privacy preserving module anonymizes the data to meet the privacy requirement along with the imprecision bound for each permission.

4. IMPLEMENTATION

The proposed system ensures security and privacy of relational databases. The overall system can be divided into 5 modules, they are:

A. User management module

This module explains about user registration. A user has to create his own account to use the software. At the time of creation of the account, the user will be asked to enter certain details that would be stored into a database. The user must also specify a username and password. When he/she tries to login, the user has to provide the username and password. The verification process is done to provide access to the system.

B. Administrative module

The administrator is the user with special privileges. The database used in the proposed work is an incremental one, with the administrator having the permission to add details into the database. The administrator manages the various users of the system. The administrator also assigns the various roles to the different users of the system, for the purpose of implementing the role based access control. The privacy requirement for each role is also set by the administrator. The administrator also sets the imprecision bound for each role.

C. Query builder module

The user requests to the database are made using a user friendly graphic user interface. These requests are transformed into SQL queries by the query builder module. It is these SQL queries that get the required data from the database that a particular user having a specific role is allowed to access. When a user tries to access a data item to which he is not authorized, an error message will be generated.

D. Authorization module

The authorization used is based on role based access control. Whenever a user executes a query, the system checks whether the user is authorized to do so, and then displays the result accordingly. The user will be allowed to view only those tuples to which he has access to, based on the roles assigned to him by the administrator. That is, when a user assigned a role executes a query, the tuples satisfying the conjunction of the query predicate and the permission is returned. Imprecision bound and different privacy constraints are also set for the different roles.

E. Data anonymization module

The risk of identification can be reduced or even eliminated by removing certain attributes and coarsening the values of other attributes. This is what is done in the data anonymization module. Anonymization can be achieved using either generalization or suppression techniques. Anonymity requirements like k-anonymity and ℓ-diversity have to be satisfied by the anonymized data. The value of k determines the number of tuples having the same QI attribute values, thereby reducing the possibility of linking attacks. The value of ℓ determines the least number of well represented values for the sensitive attribute in an equivalence class, thereby reducing homogeneity and background knowledge attacks. An imprecision bound should also be satisfied for each role which ensures that the authorized data has the desired levels of accuracy.

5. CONCLUSIONS

Databases are at the core of successful businesses. Due to the large amount of personal data being held by companies today, preserving privacy has become a crucial requirement for operating a business or any other organization. The principal concern that arises is the security of the data that organizations and individuals distribute. An organization’s data is among its most important assets, so business managers and technical leads must choose and implement appropriate devices and products to protect the very assets they wish to share [1]. Access control mechanisms ensure that only authorized information is available to users. But the problem is that sensitive information the even after the removal of identifying attributes is prone to linking attacks by the authorized users. This is because the data may contain other data such as birth date, zip code and sex, which uniquely or almost uniquely correspond to specific respondents and make them stand out from others. By linking these identifying characteristics to publicly available databases, the data recipients can determine to which respondent each piece of released data belongs, or restrict their uncertainty to a specific subset of individuals. The large amount of information easily accessible today and the increased computational power available to the attackers make such linking attacks a serious problem. The system designed prevents the authorized users from compromising the privacy of an individual leading to identity disclosure. The privacy protection mechanism that is being used is called data anonymization. The access control mechanism allows only authorized query predicates on sensitive data. The privacy preserving module uses suppression and generalization techniques to anonymize the data to meet privacy requirements and the imprecision bound. The system is accuracy constrained because some amount of imprecision is added when the database is updated. Imprecision constraints are introduced on the predicates set by the access control mechanism. The database is incremental in nature with the administrator having the authority to add data into the database. So anonymization has to be applied each time the database is updated. The combined use of access control mechanism and data anonymization ensures both privacy and security of the sensitive information.

REFERENCES

E. Bertino and R. Sandhu, “Database Security-Concepts, Approaches, and Challenges”, IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, January-March 2005.

Ravi S. Sandhu, Pierangela Samarati, “Access Control: Principles and Practise”, IEEE Communications Magazine,in Advances in Cryptology EUROCRYPT 1999, pp. 223-238.

Dipmala Salunke, Anilkumar Udadhyay, Amol Sarwade, Vaibhav Marde and Sachin Kandekar, “A survey paper on Role Based Access Control”, in International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, issue 3, March 2013.

J. Yashwanth Kumar and G. Kalpana,“Data Management By Privacy Preserving in the Anonymous Database using Suppression and Generalization protocols”, in International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, issue 9, November 2012.

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.

V. Ciriani, S. De Capitani di Vimercati, S. Foresti and P. Samarati, “k-Anonymity”, Advances in Information Security, 2007.

Latanya Sweeney, “k-Anonymity: A Model for Protecting Privacy”, in International Journal on Uncertainity, Fuzziness and Knowlegde-based Systems, pp. 557-570, 2002.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “L-Diversity: Privacy Beyond k-anonymity”, ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

Zahid Pervaiz, Walid G. Aref, Arif Ghafoor and Nagabhushana Prabhu, "Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 4, pp. 795-807, April 2014.

S. Chaudhuri, R. Kaushik, and R. Ramamurthy, “Database Access Control and Privacy: Is There a Common Ground?” Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR), pp. 96-103, 2011.

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.