Statistical Security in Data Warehouses

[1]Mr. Suraj Sinha, [2]Dr. Sohan Garg
Research Scholar (Mewar University)
[1]surajsinha05@gmail.com

ABSTRACT
The last several years have been characterized by global companies building up massive databases containing computer users' search queries and sites visited; government agencies accruing sensitive data and extrapolating knowledge from uncertain data with little incentive to provide citizens ways of correcting false data; and individuals who can easily combine publicly available data to derive information that – in former times – was not so readily accessible. Security in data warehouses becomes more important as reliable and appropriate security mechanisms are required to achieve the desired level of privacy protection. Data in Data Warehouse (DW) comes from various operational or traditional data sources and these data contains sensitive information of an organization which is used by decision maker to analyze the status and the development of an organization. Such sensitive information needs security so as to prevent these sensitive information from the unauthorized user.

Keywords: Data Warehouse (DW); Data; Security, Unauthorized User, statistical database security; privacy

INTRODUCTION

A statistical database contains information about individuals, but allows only aggregate queries (such as asking for the average age and not an individual's age). Nonetheless, inference can be used to infer some secret information. Data warehouses are built to support data mining. If a data mining tool can be used to derive sensitive information from unclassified information legitimately obtained, there is an inference problem, as discussed by Bertino et al. (2006).

Landwehr (2001) defines how the etymological roots of the term "secure" are found in "se" which means "without," or "apart from," and "cure," i.e. "to care for," or "to be concerned about". While there are many definitions of the primary requirements of security, the classical requirements are summarized by the acronym CIA. CIA is the acronym for confidentiality, integrity, and availability. All other security requirements such as nonrepudiation can be traced back to these three basic properties. Avizienis (2004) defines *confidentiality* as the absence of unauthorized disclosure of information, *integrity* as the absence of improper system alterations and *availability* as readiness for correct service.

- *Dependability* is a broader concept that encompasses all primary aspects of security save confidentiality, and, in addition.
- *Reliability*, which refers to the continuity of correct service.
- *Safety*, defined as the absence of catastrophic consequences for user(s) and environment.
- *Maintainability*, which is the ability to undergo modifications and repairs.

BACKGROUND

Data warehouse is a newest technology in any organization which has various security issues like data integration, data security, data consistency and confidentiality of data. The confidentiality and integrity of the data is very essential to provide security to organization's information. For maintaining the security principles like confidentiality and integrity of data in an organization, encryption technique is used to resist the data from unauthorized users while security obviously encompasses the requirements of the CIA triad this article will focus on the mechanism of access control (AC) as this addresses both confidentiality and—to some extent—integrity. Database security was addressed in the 1960s by introducing *mandatory access control* (MAC), driven mainly by military requirements. Today, *role-based access control* (RBAC) is the commonly used access control model in commercial databases.

RBAC Constraints

Since permissions are organized into tasks by using roles, conflicts of interests are more evident than if dealing with permissions on a per-user basis. In fact, a conflict of interest among permissions on an individual basis is hard if not impossible to determine.

Separation of duties among roles (i.e., defining mutually exclusive roles) provides the administrator with enhanced capabilities to specify and enforce enterprise policies. Since RBAC has static (user-role membership) and dynamic (role activation) aspects, the following two possibilities can be distinguished accordingly.

First, *Static Separation of Duties* (SSD) is based on user-role membership. It enforces constraints on the assignment of users to roles. This means that if a user is authorized as a member of one role, the user is prohibited from being a member of a second role.

Constraints are inherited within a role hierarchy.

Second, *Dynamic Separation of Duties* (DSD) is based on role activation. It is employed when a user is authorized for more roles that must not be activated simultaneously. DSD is necessary to prohibit a user from circumventing a policy requirement by activating another role.

Administrating RBAC
Definition of roles and constraints, assigning permissions to roles, and granting membership to roles are the most common administrative tasks in RBAC. When a new employee enters the company, the administrator simply adds this person to one or more existing roles according to the users tasks and needs. Similarly, users can be removed from a role when they leave the company or added to new roles when their functions change.
It is commonly agreed that one of RBAC's biggest advantages is its easy administration. Nonetheless, managing a large number of roles can still be a difficult task. However, Sandhu and Coyne (1996) present an intriguing concept that shows how RBAC might be used to manage itself. An administrative role hierarchy is introduced, which is mapped to a subset of the role hierarchy it manages.

Coexistence with MAC / DAC
Mandatory access control is based on distinct levels of security to which subjects and objects are assigned. Discretionary access control (DAC) controls access to an object on the basis of an individual user's permissions and/or prohibitions. RBAC, however, is an independent component of these access controls and can coexist with MAC and DAC.
RBAC can be used to enforce MAC and DAC policies as shown in (2000). The authors point out the possibilities and configurations necessary to use RBAC in the sense of MAC or DAC. For a detailed discussion on defining and organizing roles please refer to Nyanchama and Osborn (1994), who introduce a formal role graph to facilitate role administration. Ferraiolo and Kuhn (1992), for example, published fundamental concepts on granting and revoking membership to the set of specified named roles.

SCIENTIFIC CONCEPTS
Classic access control is still the mechanism of choice to protect not only databases but also data warehouses. The difference between a database and a data warehouse is that database is designed and optimized to process individual tuples and the data warehouse is optimized to respond to queries that analyze aggregated data. OLTP (On-Line Transaction Processing) systems are secured by controlling access to individual tuples but for data warehouses the issue of data protection is more complex. For typical access control there are several shortcomings. First and foremost, users can do anything with the data once they have access to it; Second, even if access to fine grained detail data is not permitted, querying different similar datasets can reveal fine details; this is also known as inference
attacks. The first issue can be addressed—in theory—by usage control as described by Park and Sandhu (2004), the second by several methods of statistical database security. Both topics are very active fields of research.

Usage Control
The main problem with data collection is that people might allow companies to use data for specific reasons (such as recommending related products) but do not consent to other uses of the same data. Usage control byPark and Sandhu (2004) is a concept that makes it possible to enforce pre- and postconditions when using data. It is similar to a traditional reference monitor, only that the restrictions are enforced during the entire access, as proposed by Thuraisingham (2005): The privacy control would "limit and watch access to the DBMS (that access the data in the database)."

Statistical Database Security
Well-established protection concepts for statistical database security, such as: restriction based techniques, query set size control, expanded query set size control-audit based (assumed information base), perturbation-based techniques, data swapping (distribution unchanged), random-sample queries, fixed perturbation (modify data), and query-based perturbation. For an in-depth description Castano et al. (1994) and Willenborg and De Waal (1996) are excellent sources.
Query set size control Enforcing a minimum set size for returned information does not offer adequate protection. Denning (1982) described trackers that are sequences of queries all within the size limits allowed by the database; when combined with AND statements and negations, information on individuals can be inferred. While simple trackers require some background information, Denning (1979) as well as Denning and Schlorer (1983) show how general trackers can be used without in-depth background knowledge.

In ***audit-based expanded query set size control*** aka. Nabil and Worthmann's (1989) 'query set overlap control' the system decided whether to grant access to an "assumed information base," which is the history of all the requests issued by the user. The assumed information base contains all possible inferences that can be generated with the results of all previously issued queries; before

answering a new query the system has to decide whether the query could be combined with the assumed information base to infer confidential information.

Perturbation-based techniques (cf. Table 1) are characterized by modifying the data so that the privacy of individuals can still be guaranteed even if more detailed data is returned than in restriction-based techniques. Data can be modified in the original data or in the results returned.

Data swapping	Data is exchanged between different records; individual information is thus protected while calculated statistics are not impacted.
Random sample queries	A set of answers to a specific query are created dynamically by selecting a random subset instead of all data item. This approach works well only for large datasets.
Fixed perturbation	Data is modified—not swapped—as soon as it is loaded into the data warehouse.
Query-based perturbation	Data is modified for each query dynamically. The advantage is that the accuracy can be varied individually depending on the user's trustworthiness.

Table 1 : Perturbation-based techniques

CONCLUSION:

This work summarizes the security solutions, which is used to secure the data warehouse. Number of researchers has given numerous security solutions at each level (business level, conceptual level, logical level and physical level) of data warehouse design. Some researchers has implemented security at access level after data warehouse design to prevent the critical or important data from unauthorized access. After implementing these security solutions in data warehouse still there is a need of such a measure which can eradicates all the security issues in data warehouse which can prevent the sensitive or critical data of an organization.

Future research in data warehouse security will address several issues. First, with the increasing size of DWHs containing very personal information, privacy-preserving techniques will become more important. This area of research has also received more attention because nation-wide data gathering programs for national security are established. Second, while this theoretical research is certainly important, there are many more aspects to security that need to be considered. A nationwide DWH needs to be secured as an entire system including the mechanisms of data delivery, data querying, and usage. Security in DWH rests on three tiers: (1) technical infrastructure such as firewalls, encryption, (2) security in data gathering, privacy preserving techniques, and (3) secure applications including authentication, access control, authorization and auditing

REFERENCES

[1] Nabil, R. A., & John, C. Worthmann (1989). Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4), 515– 556.

[2] Charu, C. (2005). Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st VLDB Conference*.

[3] Algirdas Avizienis, J.-C., Laprie, Randell, B., & Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions of Dependable and Secure Computing*, 1(1), 11–33.

[4] Bell, D., & La Padula, L. (1975). *Secure computer system: Unified exposition and multics interpretation*. Esd-tr-75-306, Technical Report mtr-2997, Bedford, MA: The MITRE Corporation.

[5] Bertino, E., Khan, L. R., Sandhu, R., & Thuraisingham, B. (2006, May). Secure knowledge management: Confidentiality, trust, and privacy. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 36(3), 429– 438.

[6] Castano, S., Martella, G., Samarati, P., & Fugini, M. (1994). *Database Security*. Addison-Wesely, ACM Press.

[7] Mitre Corporation (1997, April). *Integrity considerations for secure computer systems.* Technical Report esdtr-76-372, esd,/afsc, mtr 3153, Bedford, MA: Mitre Corporation.

[8] Denning, (1982). *Cryptography and Data Security*. Addison Wesley.

[9] Denning, D. E., & Denning, P. J. (1979). Data security. *ACM Comput. Surv.*, 11(3), 227–249.

[10] Denning, D. E., & Schlorer, J. (1983, July). Inference controls for statistical databases. *IEEE Computer*.

[11] Fernández-Medina, E., Trujillo, J., Villarroel, R., & Piattini, M. (2006). Access control and audit model for the multidimensional modeling of data warehouses. *Decis. Support Syst.*, 42(3), 1270–1289.

[12] Ferraiolo, D.F., & Kuhn, R. (1992, October). Role-based access control (rbac).

In *Proc. 15th NIST-NSA National Computer Security Conference*, Baltimore, MD.
[13] Guimaraes, M. (2006). New challenges in teaching database security. In *InfoSecCD '06: roceedings of the 3rd annual conference on Information security curriculum development*, 64–67, New York, NY, USA, ACM.
[14] Landwehr, C.E. (2001). Computer security. *Int. Journal of Information Security*, 1(1), 3–13.
[15] Loukides, G., & Shao, J. (2007). Capturing data usefulness and privacy protection in k-anonymisation. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, 370–374, New York, NY, USA: ACM.
[16] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). L-diversity: Privacy beyond k -anonymity. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, 24, Washington, DC, USA: IEEE Computer Society.
[17] Nyanchama, M., & Osborn, S. (1994). Ifip wg 11.3 working conf. on database security. database security viii: Status and prospects. In *Proc. 15th Annual Computer Security Applications Conference*, North-Holland.
[18] Osborn, S., Sandhu, R.S. & Munawer, Q. (2000). Configuring role-based access control to enforce mandatory and discretionary access control policies. *ACM Transaction on Information and System Security*, 3(2), 85–206.
[19] Park, J., & Sandhu, R. (2004). The uconabc usage control model. *ACM Transactions on Information Security*, 7(1), 128–174.
[20] Priebe, T., & Pernul, G. (2004). Sicherheit in Data-Warehouse- und OLAPSystemen. *Rundbrief der Fachgruppe Modellierung betrieblicher Informationssysteme (MobIS) der Gesellschaft für Informatik e.V. (GI)*.
[21] Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*.
[22] Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996, February). Role-based access control models. *IEEE Computer*, 29(2), 38–47. doi: http://csdl.computer.org/comp/mags/co/1996/02/r2toc.htm.
[23] Sandhu, R.S., Ferraiolo, D., & Kuhn, R. (2000, July). The nist model for rolebased access control: Towards a unified standard. In *Proc. of 5th ACM Workshop on Role-Based Access Control*, Berlin, Germany: ACM, ACM Press.
[24] Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 557– 570.