General Considerations on Data Warehouse Security

Rodolfo Villarroel[1], Eduardo Fernández-Medina[2], Mario Piattini[2]
(1) Departamento de Computación e Informática, Universidad Católica del Maule,
Avenida San Miguel 3605 Talca (Chile)
rvillarr@spock.ucm.cl
Phone: +56 71 203525, Fax +56 71 260278
(2) Departamento de Informática, Universidad de Castilla-La Mancha,
Paseo de la Universidad, 4 13071 Ciudad Real (Spain)
{Eduardo.Fdezmedina, Mario.Piattini}@uclm.es
Phone: +34 926 29 53 Ext.: 3744, Fax +34 926 29 53 54

Abstract

There is plenty of information regarding research works on Data Warehouse (DW) management systems aimed at improving several aspects such as data modeling techniques, the physical level of modeled data, transactional processing... However, not too many efforts have been made related to security aspects. At present, there are methodologies to design DW, but security is not taken into account. On the other hand, there are techniques to design security but these techniques do not consider DW. Therefore, it is necessary to study the way to design secure DW. This paper puts forward general considerations on security (specifically, confidentiality) when designing DW.

Keywords: **Data Warehouse security, confidentiality, security model, security design, security requirements**

1.0 Introduction

Security is a "horizontal" aspect of software development that affects very closely each component of an application and, its integration in the process of software development is not appropriately understood [14]. In the field of security, we can distinguish between the concept of security (capability of a system to manage, protect and distribute sensitive information) and the concept of safety (absence of catastrophic consequences to the environment). The issue of DW security will be dealt with in this paper.

According to ISO/IEC 15408-1 [7], the concept of security refers to the capability of a software product to protect data and information in order to avoid that unauthorized individuals or systems are able to read and modify them and not to deny access to authorized staff. Castano et al. [4] refer to computing security such as the protection of information against unauthorized queries, inappropriate modifications or the lack of availability of a service in a given moment. We can see that both definitions of security are basically similar according to the following components: confidentiality (to prevent, to detect, to avoid the improper revelation of information), integrity (to prevent, to detect, to avoid the undue modification of information) and availability (to prevent, to detect, to avoid the denial of access to the services provided by the system).

We have to take always into account risk factors that can alter the security conditions [13]. Table 1 allows us to see risk factors according to security conditions.

The three security conditions mentioned in Table 1 (confidentiality, availability and integrity) have been studied by several authors, mainly in relation to databases. Concerning DW, security aspects have been dealt with by Gupta and Widom [6] from the viewpoint of Integrity. Besides, they have been studied by Labio and García Molina [12] from the viewpoint of Availability. Nevertheless, a very critical aspect such as DW confidentiality has not been appropriately studied. For instance, we have to be especially worried about the security of DW personal data since the capability to cross and analyze information coming from many sources, specifying profiles and patterns may be a threat to personal privacy.

Table 1. Risk factors according to security conditions

SECURITY CONDITIONS	RISK FACTORS				
	Intruder	Robbery	Catastrophe	Sabotage	Computer virus
Confidentialiy	X	X			
Availability	X	X	X	X	X
Integrity				X	X

A DW is better evaluated when it allows users an easier access to information, but if this information is accessed by unauthorized staff, it will loose all its value. An important aspect to be taken into consideration when studying DW is that information is not treated statically but its evolution through the time, in other words, its history, becomes more important. For this reason, mechanisms to allow confidentiality of such quantity of information must be established.

It is important to consider DW security from a methodological approach that allows us to design DW taking into account security aspects from the earliest stages to the end of the process of development. This approach should be an extension of existing modeling methodologies and standards to avoid that companies interested in DW security have to make an additional effort to learn other methodology.

2.0 Introduction to DW and the problem of Associated Security

A DW is a subject-oriented, integrated, time-varying, nonvolatile collection of data organized to support the decision making process [8]. This definition indicates that data are not oriented to functional processes as in classical applications. On the contrary, they are oriented to subjects, providing a unique and integrated view of the organization, which is understood in a transversal way. Moreover, information is not treated statically but its evolution through the time gets more important.

It is very common the fact that storage is not monolithic and it could be divided into subsets, not necessarily unjoint, that give us a partial view of the organization. These elements, called data marts, can be defined as databases oriented to the subject, available for users, for decentralized decision making, and comprising a less wide field than DW.

Nowadays, we must work in operational as well as analytical environments. These environments have very different functions. The mission of operational systems is to be the infrastructure of day to day business functions; hence, they only contain the necessary data to fulfil the business daily requirements. If we tried to use these systems to process consistent, integrated, well-defined and time dependent information for purposes of analysis and decision making, we would notice that data available from operational systems do not fulfil these requirements. To solve this problem, we must work in an analytical environment strongly supported by the use of multidimensional models to design DW.

Each dimensional model is composed of a table with a composite primary key called fact table and a set of smaller tables called dimension tables. The structure that they form has the shape of a star, see Figure 1, and that is the reason why this modeling system is called star schema.

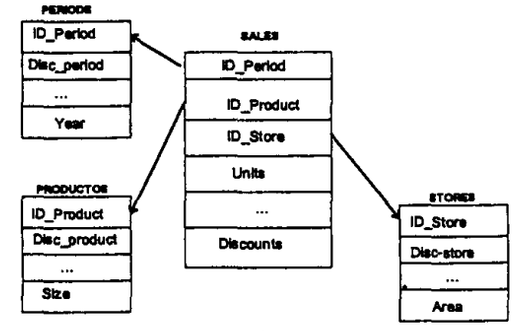

Figure 1: Example of a Star Schema

The star schema is highly denormalized, so it is very different from database modeling. The objective of this denormalization is to increase the query efficiency because a fact table is related to lots of dimension tables. The snowflake schema is a variant of the star schema and within this modeling system, dimensions are normalized by creating hierarchies along dimensions and maintaining fact tables, which are the essential part of the star schema.

In an operational environment, just a small quantity of information is accessed (the information used in a transaction) due to the fact that users know what they are looking for. In contrast, in a DW, users carry out very large searches throughout milliards and millions of registers to answer a unique query and very often, they do not even know what they are looking for before starting the search. For that reason, fact tables are formed mainly by numeric and additive indicators that allow users to obtain register summaries useful to improve the query efficiency.

In DW architecture, data are extracted from several sources, after that, they are cleaned and depurated. Later on, they are stored to be finally accessed by users for decision making [8].

Hence, we can distinguish three main DW processes, as shown in figure 2. These processes are the following:

- Acquisition: Set of processes whose purpose is to extract information from the origin systems (operational systems, external systems, etc.), to integrate it, to transform it, to depurate it and finally, to load it into DW according to a previously established design.
- Storage: It is the main part of DW, the place where all data, ready to be exploited later, are located. Storage is independent of the later use to be done from it, not only from the viewpoint of users but also from that of applications.
- Access: Set of processes whose objective is to capture and exploit DW content to provide users the information needed for decision making.

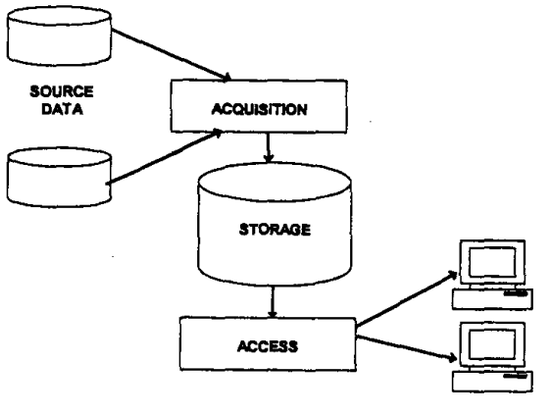

Figure 2. Main processes of a Data Warehouse

Some security aspects to be taken into account in the Acquisition process are the following:

- Security measures must be established to avoid unauthorized access to extraction, cleaning, depuration and loading of data.
- Temporary files and loading areas used during the Acquisition process must be protected and, at the same time, it must be guaranteed that they will be deleted and destroyed when they are no longer useful.
- Integrity in the extraction of temporary information must be kept. This fact will allow a future synchronization with information coming from other sources.

Some security aspects to be considered in the Storage process are the following:

- If there have been introduced data marts, they must fulfil the same security restrictions than DW, independently of the fact that they are located within the same system or within a different one. Other option is to use data marts as places to store the most sensitive information and locate them in an even better protected place than DW, since, usually, the most confidential information must be only accessed by a small set of individuals.
- Security of metadata or "data about the data" must be taken into consideration, a repository that must contain information such as: the data source, the transformation suffered by each data, the data model and its relation to the DW, integrity rules, rules to manage views, definitions of roles and access, information quality and metrics.
- The encryption of critical information must be considered.
- Security must be considered, taking into account the level of granularity, in other words, the disaggregation of data. For example, the total amount of daily, weekly or monthly sales that will affect the size of DW, a critical factor by itself because of its big increase. Level of frequency and availability are important to be considered within security mechanisms.

Some security aspects to be considered in the Access process are the following:

- Access, executed by specific applications or any other kind of tools must be subjected to

security restrictions established in data model.

- Security in the access of end users as well as people in charge of extracting information from source systems must be considered.
- Multidimensional aspects in the access to DW, with new concepts to consider in the security such as dimension, access path, etc., must be taken into account.
- One of the security problems that must be considered in DW is that one related to inference. Inference attacks make it possible that DW protected information can be obtained by other means. For instance, if we want to know the wage of the president of the company, the most reasonable thing to do is to protect this information against the access of not authorized people. However, a statistic query, such as that one to obtain the maximum of wages can be unprotected.
- Queries executed using Data Mining tools need a harder access control with respect to the confidentiality of the data to be given.

Decisions taken to preserve security must take into account the special features of DW. These features can be summarised as follows [19]:

- DW stores data derived from the local data.
- Queries are processed at the DW, without need of consulting the local databases from which the data stored at the DW have been obtained.
- DW must respect the autonomy of the local databases.

Moreover, regarding DW security and paying special attention to DW access control, some questions could be asked such as:

- Should access restrictions specified in DW be the same than those specified at a local level?
- Are DW able to allow access to some data when the local sources, from data have been obtained, do not allow it?
- Are DW able to deny access to data although local sources, from data have been obtained, allow it?
- Who is able to specify access restrictions to data stored in DW that have not a correspondence with data of the local system? (For example, historical data, or data obtained in the DW, aggregating data from different sources).

Security and specifically, confidentiality is a very important aspect for DW, due to the fact that constant changes in user queries and data sources force DW to be more flexible, but also to carry out a harder control in confidentiality of information.

3.0 Security proposals related to Data Warehouses

In the literature, we can find several initiatives to include security in DW. Many of them focus on interesting aspects related to access control, multilevel security, its application to federated databases, applications using commercial tools, etc. These initiatives will be analyzed below.

In [11], a security model based on a mandatory access control for OLAP cubes is proposed. Security restrictions are defined for each role within a DW environment, in a way that each role defines a subcube of the DW's N-dimensional cube. The advantage of this approach in relation to security management is its flexibility to assign roles to the different virtual subcubes. This model attacks an important but punctual aspect, related to access control. It does not consider a broader approach, comprising from the earliest stages of development, and it does not refer to modeling systems such as the conceptual and logical ones to design DW either.

[9] states a model for DW security based on metadata. This model allows us to assign a reduced view of a DW for each group of users making it possible that all users can freely navigate through the DW reduced data. Moreover, this model avoids threats to the security regulations of the information system. This proposal starts analyzing the requirements and impacts in the selection of an appropriate security model for DW, Later, the model is stated and applied through a prototype. However, it does not offer a work analyzing a modeling system; it only indicates what aspects (legal, auditorial related to networks, etc) must be considered in security requirements.

In [2], it is indicated that access to DW can be restricted or modified using user profiles. Furthermore, this initiative is focused on ideas that can contribute to DW security such as control of replication, aggregation and generalization, exaggeration and misleading, anonymity, and user profile based security. There is not a formal proposal. On the contrary, ideas that can contribute to DW security are studied in a global way. So, they are only general recommendations that can be helpful to establish formal proposals of security design.

[17] suggests taking view security as a base of DW security. It states a proposal that allows automated inference of lots of permissions in DW in order to minimize the learning curve for administrators as well as the amount of new software that vendors would need to implement. To do so, it puts forward an extension of SQL standard model for systems with redundant and derived data. In contrast, this proposal does not take into account aspects related to DW conceptual model nor DW security in conceptual models. It is more focused on the logical aspects that consider a compatibility with relational technology and at the same time, it considers that one of the unsolved problems is the creation of a tool based on theory.

In [16], research is focused on authorization and access control. For that reason, it defines different OLAP access control requirements and compares the implementation of commercial ROLAP (Relational OLAP) products, such as Microsoft SQL Server 2000, MicroStrategy, Cognos Powerplay, and Oracle Express. It underlines the practical aspect of implementation in commercial systems and a later comparison of capabilities to implement security requirements. Moreover, it indicates globally that the design methodology of classical databases (requirement analysis, conceptual design, logical design and physical design) should be applied to OLAP applications security too. But, it recognises that multidimensional conceptual data modeling and OLAP security mechanisms are significantly different from the capabilities of relational database management systems. This is a very general proposal and does not study the way to carry out each DW modeling including security; it only indicates that requirements will be formalized in the future by defining a security restriction language.

[21] develops a basic authorization model for DW and OLAP, focused especially on expressivity and usability in DW context. The purpose of this model is based upon the fact that, despite many DW are implemented in relational databases that offer a wide variety of security mechanisms; it is not trivial to grant permissions in relational tables. This is due to the fact that SQL sentences do not allow us to define access privileges intuitively. Therefore, this proposal states a very basic algebraic formal model and besides, it indicates that an informal authorization language can be used to illustrate how to formulate intuitively access restrictions for operations in a DW. The proposed solution is limited and does not take into account the different types of DW design with their security restrictions.

[1] refers to an architecture of Information Systems for secure DW. It suggests finding DW schemas in an architecture for federated databases. This proposal studies the Mandatory Access Control mechanism (MAC) to protect data from unauthorized access. Special attention is paid to the architecture of DW schemas and their conceptual design. Finding the different DW schemas in an architecture for federated databases, authors indicate that an integrated architecture comprising both areas has been achieved. This proposal considers an integrated access in real time and it establishes bases to access not only to databases but also to other sources of information. As it is based upon an architecture of federated databases, all the results of the research must be considered with this vision.

We can conclude that the analyzed proposals refer to punctual aspects that allow us to improve DW security in Acquisition, Storage and Access aspects. However, neither of them studies the security aspect comprising all stages of the system development cycle nor considers the introduction of security in DW design (conceptual design, logical design and physical design). As there is not a methodological approach integrating security in DW design, we can state that the problem of DW security remains unsolved.

4.0 Conclusions and future work lines

There is a mature field in relation to methodologies and techniques associated to DW modeling [3], [5], [10], [15], [18], [20]. However, there is an immature field regarding to the fact of taking into account security aspects in these methodologies and techniques. Some security proposals associated to DW have been developed but they are punctual solutions that partially comprise the necessary security requirements for the process of Acquisition, Storage and Access for DW. Furthermore, none of these proposals considers a methodological approach formally including security in the process of DW design from the conceptual, logical and physical points of view.

Taking into consideration the fact that information technologies must provide new ways to manage the huge amount of typical data of current DW, we have to pay special attention to information confidentiality aspects. To do so, we must start with the creation of a rule to classify information according to the level of confidentiality to be applied, from completely public to very secret information. After that, we can define models, methods and tools that enable us to design secure DW in an integrated way, by using UML and the multilevel security model. To be able to do that, we think that a methodological approach that allows us to build DW taking into account security aspects from the earliest stages of development until the end of the process should be developed. This methodological approach should be an extension of existing modeling methodologies and standards because, otherwise, organizations that are really interested in DW security would have to make a big effort to adapt to the new technology. At present, we are progressing in the definition of these methodological aspects.

The more widely spread modeling standard is UML. Hence, it would be interesting to provide UML with security features to be able to develop modeling including, on the one hand, the UML syntax and power and on the other hand, the new security features, ready to be used, when the application includes security requirements that need these features. UML has been widely accepted as the modeling language oriented to standard objects to model several aspects of software systems. So, any approach using UML will minimize the effort of developers to learn new notations or methodologies for each subsystem to be modeled. UML is an extending language since it provides mechanisms (stereotypes, tagged values and restrictions) in specific domains, if necessary, such as web applications, business modeling, software development processes and so on. We consider it appropriate the use of a DW design methodology that uses UML extensions for security aspects to be added. Moreover, we think it is essential the use of an OCL (Object Constraint Language) based language to be able to specify DW security restrictions precisely together with Class Diagrams and the development of a CASE tool, integrated in Rational Rose, to support the process of DW design in a secure way for the later validation of the proposal by applying it to real situations. The biggest cost of software system is maintenance and this is a consequence of imprecise, incomplete and arbitrary documentation. With an UML extension that allows us to model DW security requirements, a more robust specification will be achieved.

Moreover, applying the methodological approach, it would be desirable that information systems that are developed fulfil the necessary protection requirements if they store personal or sensitive data. In many countries, these requirements are demanded and determined by the existence of laws dealing with the protection of personal data.

Acknowledgements

This research is part of the CALIPO project, supported by the Dirección General de Investigación of the Ministerio de Ciencia y Tecnología (TIC2003-07804-C05-03), and the MESSENGER project, supported by the Consejería de Ciencia y Tecnlogía of the Junta de Comunidades de Castilla-La Mancha (PCC-03-003-1).

5.0 References

[1] A. Abelló, M. Oliva, J. Samos y F. Saltor. Information System Architecture for Secure Data Warehousing. Technical Report LSI-00-26-R. Dep. of Information Systems, University of Catalunya, Spain. April 2000. http://citeseer.nj.nec.com/article/abello00information.html

[2] B. Bhargava. Security in Data Warehousing (Invited Talk). Proceedings of the 3rd. Data Warehousing and Knowledge Discovery (DAWAK'00). 2000.

[3] F. Carpani. Multidimensional Models: A State of Art. Reporte Técnico RT 00-12, Universidad de la República, Uruguay, 2000. http://www.fing.edu.uy/inco/pedeciba/bibliote/reptec/tr0012.pdf

[4] S. Castano, M. Fugini, G. Martella y P. Samarati. Database Security. Addison Wesley. 1995

[5] M. Golfarelli, D. Maio y S. Rizzi. Conceptual Design of Data Warehouses from E/R Schemes, International Conference on System Science HICCS'98, IEEE, Hawai, 1998.

[6] A. Gupta y J. Widom. Local Verification of Global integrity Constraints in Distributed Databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. United States, Washington, D.C., pp. 49-58.

[7] ISO/IEC 15408-1. Information Technology. Security Techniques. Evaluation Criteria for TI Security. Part I: Introduction and General Model.

[8] H. Inmon. Building the Data Warehouse. John Wiley & Sons. Third Edition. 2002.

[9] N. Katic, G. Quirchmair, J. Schiefer, M. Stolba, y M. Tjoa. A Prototipe Model for Data Warehouse Security Based on Metadata. Proceedings of the 9th. Int. Conf. On Database and Expert Systems Applications (DEXA'98), Vienna. Austria. IEEE Computer Society, vol 8, pp. 300-308, 1998.

[10] R. Kimball, L. Reeves, M. Ross y W. Thornthwaite. The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, 1998.

[11] R. Kirkgoze, N. Katic, M. Stolba y A. Tjoa. A Security Concept for OLAP. Proceedings of the 8th. International Workshop on Database an Expert System Applications (DEXA '97). IEEE Computer Society. 1997.

[12] W. Labio y H. García-Molina. Efficient Snapshot Differential Algorithms in Data Warehousing. Technical Report, Dept. Of Computer Science, Stanford University, 1995.

[13] A. Lardent. Sistemas de Información para la Gestión Empresaria: Procedimientos, Seguridad y Auditoría. Prentice-Hall. 2001.

[14] T. Lodderstedt, D. Basin, J. Doser. SecureUML: A UML-Based Modeling Language for Model-Driven Security. 5th International Conference on the Unified Modeling Language, Dresden, Germany, LNCS 2460, Springer-Verlag. 2002.

[15] S. Luján-Mora, J. Trujillo y I. Song. Extending the UML for Multidimensional Modeling. 5th International Conference on the Unified Modeling Language, Dresden, Germani, LNCS 2460, Springer-Verlag. 2002.

[16] T. Priebe, and G. Pernil.Towards OLAP Security Design – Survey and Research Issues. Proceedings of the 3rd ACM International Workshop on Data Warehousing and OLAP (DOLAP 2000), Washington DC, November 2000.

[17] A. Rosenthal, and E. Sciore. View Security as the Basic for Data Warehouse Security. Proceedings of the International Workshop on Design and Management of Data Warehouse (DMDW'2000), Sweden. June 2000.

[18] C. Sapia, M. Blaschka, G. Höfling y B. Dinter. Extending the E/R Model for the Multidimensional Paradigm. Advances in Database Technologies, LNCS Vol 1552, Springer-Verlag. 1999.

[19] B. Thuraisingham, L Schillper, P. Samarati, T. Y. Lin, S. Jajodia y C. Clifton. Security Issues in Data Warehousing and Data Mining: Panel Discussion, Database Security XI – Status and Prospects, T. Y. Lin ans S. Qian (eds.), Chapman & Hall, pp. 3-16, 1998.

[20] J. Trujillo, M. Palomar, J. Gómez y I Song. Designing Data Warehouses with OO Conceptual Models. IEEE Computer, Vol 34, número 12, 2001.

[21] E. Weippl, O. Mangisengi, W. Essmayr, F. Lichtenberger, and W. Winiwarter. An Authorization Model for Data Warehouses and OLAP. http://citeseer.nj.nec.com/522638.html