SECURITY OF DATAWAREHOUSING SERVER

Mr. V.M.Navaneethakumar,
Assistant Professor,
Department of Computer Applications,
K.S.R. College of Engineering, Tiruchengode.
Email: kumarvmn@gmail.com

Dr. C. Chandrasekar,
Reader & Associate Professor,
Department of Computer Science,
Periyar University, Salem – 636011.

ABSTRACT

The aim is to study the data warehouse specific security needs which means focusing on the long life cycle of the data and the data security and privacy issues. These requirements need to be fulfilled despite of the unpredictable future uses of the data. The basic assumption is that data warehouse server is running on the dedicated server, and the basic information system security is properly taken care of. Thus this paper covers only the basics of the regular information system security approaches used in the server and database environments and attention is paid on data warehouse specific issues that come from the sensitivity, business value and the long and unpredictable lifespan of the data. The aim is to present a data warehouse security framework, based on the current state of the art that concentrates on the data warehouse specific security needs from the data content point of view and gives a useful tool to help dealing with data warehouse security process.

***Keywords*: -** Data warehouse, Dwh, Security, Privacy, Framework.

1. INTRODUCTION

Data warehouses have become increasingly popular as information in general has been available in the electronic form and increased computation and network resources have made it possible to process large quantities of the data in an efficient manner. The concept of data warehouse is based in the idea of long time structured storage where data is kept archived and available for future use with a structure that supports efficient analyses. The typical use scenario of the data warehouse contains complex cross-analyses of the data, in some cases even using multiple data sources in order to derive additional information from the data. This allows very broad and detailed analyses of the information contained within the source data. The cross-use of the multiple sources can increase the value and level of the information available.

The basic assumption is that the information system implementing the data warehouse uses all the necessary means for protecting the integrity and availability of the underlying database, as well as the confidentiality and privacy of the information. However, the data mining use of the information system content can not necessarily be predicted at the time when the data warehouse is set up. Further, the future cross-referencing and combining of the multiple data sources can create the situations where the level of the information that can be extracted from the data warehouse strongly exceeds all that was originally predicted.

This paper examines the approaches on securing the information in data warehouse. The basic practices for protecting the integrity and the privacy of the information system (encryption, backups, user access control, etc.) and the approaches on securing the actual operating environment (network, servers, applications, etc.) used to run a data warehouse are left outside of the scope. This paper concentrates into the “C” of the information system security often referred as CIA (Confidentiality, Integrity, and Availability) and how to implement that as a part of the overall security design of the data warehouse. It is assumed that regular information system security approaches that apply in any networked information system are properly taken care of, and this paper concentrates on data warehouse specific issues. More general purpose information security assurance activities that apply to IT-systems in general are well documented for example in SSE-CMM [1], but as has been described in the previous sections the nature of the data warehouse calls for additional security steps.

2. RESEARCH METHODOLOGY

The paper is based on the literature research. The intention is to provide an overview over the current state of the art and use that as a base for presenting a data warehouse security planning framework that emphases the data warehouse specific security needs.

3. BASIC APPROACHES ON SECURING THE DATA WAREHOUSE

When building a data warehousing system there are several things to be considered regarding the security of the resulting system. Several steps are common in any information system related development task, but the nature of the data warehousing system requires special attention on the data itself. In the literature there are several “task-list” approaches to guide the execution of a data warehousing project. One example of such a “task-list” is presented in the work of Warigon [9] already in 1998. The source lists number of basic

actions that needs to be taken care of already during the planning phase of the data warehouse.
The steps presented are:

- **Identifying data**; this means creating an inventory of the data that is made available to data warehouse users.
- **Classifying data**; creating the initial classification of the sensitivity and the type of the data stored in the data warehouse. This, together with item 1 emphases the importance of understanding and defining the nature of the data as early as possible in the planning phase. These items together create the base for understanding the data ontology, which is crucial both for the efficient use of the data and for properly protecting the data, as has been stated e.g. in the work of N. Szirbik et.al. [8].
- **Quantifying the value of data**; this is done to provide the base for some estimates in the potential cost of recovering security preaches (that being of corruption and/or loss of data, or loss of confidentiality etc.). The actual financial value may be hard to estimate as the inconsistency of the data warehouse content may lead to erroneous business decisions.
- **Identifying data protection measures and their costs**; for all the identified threats the potential remedies are defined and priced.
- **Selecting cost-effective security measures**; the identified security measures are leveled with the value of the data and the severity of the threat.
- **Evaluating the effectiveness of security measures**; finally the effectiveness of the security measures needs to be addressed. These are an example of the basic steps to be taken when planning the data warehouse. All the steps are required and can be seen essential, but especially the steps 1, 2 and 4 create the base the sophisticated data protection approaches. Also the works of Fernandez-Medina et.al. [3] and Szirbik et.al. [8] Emphasize the importance of the early stage planning in order to implement efficient security controls. The early stage planning is important not only for the access control methods, but is required also for the proper implementation of the audit methods of the data warehouse implementation. Proper auditing controls needs to be defined as a part of continuous security process. Generic information system security approaches with basic steps are described in several sources, but System Security Engineering Capability Maturity Model (SSE-CMM) [1] offers a good example of the collection of basic practices on securing the information system. However, SSE-CMM does not deal with the data warehouse specific issues, but the essentials are always the same: you need to know what to protect, why, and against what, and then define the approach based on these corner stones.

Just like with regular information systems, the security of the data warehouse is not something that is implemented once and then keeps protecting the system through it's entire life span. There needs to be processes in place to monitor the changes both in the system and it's environment and making sure that necessary adoption to the changes takes place also in the security measures [1]. Further, the auditing methods for the follow-up of the current security measures and policies needs to be considered when planning the data warehouse implementation [3]. When the project is dealing with the data warehouse there are data security issues that are specific to the data warehouse or multidimensional databases (MD) alike, but can not be comprehensively covered by general information system security practices. These considerations have been examined for example in the work of E. Fernandez-Medina et.al. [4] and also in the white paper material of a database vendor [6].

Ensuring the security is also about to manage the complexity: the more complex the system is, the harder it is to keep the whole concept properly managed and thus properly secured. This implies that the collection of several data marts is harder to secure than is one central data repository [6]. However, the centralized system may bring other issues regarding the availability of the service and also one central collection of the data means that all the valuable information is located in the same system and impact of the security breach can be more devastating that it would in a case of single data mart.

4 Specific issues on Data warehouse and addressing them

Section 4 first examines the data warehouse specific security needs that can be identified. The needs are classified into specific security areas that are then organized as a framework that can be used to guide the security work with data warehouses. Last part of this chapter gives a brief overview on the techniques that can be applied into identified security areas.

4.1 Specific security needs on Data warehouse

We can agree that the approaches for data warehouse security must extend beyond the usual measures used to protect the conventional databases and IT-systems. The base information system security practices, however, cannot be omitted as they provide the base also for the data warehouse security. Even the most novel data warehouse level protection measures are rendered useless if the database itself can be compromised. As was discussed already in Section 3 there are security issues that are specific for data warehouses but do not concern regular databases, or at least cannot be solved with approaches that are suitable for the conventional databases.

These special considerations root from the fact that regular databases base their security mainly on the access restrictions that are defined on the very basic level of the database structures (e.g. per table or table row). This approach may either restrict the use of data too much, or be inadequate after cross analyses have been run. The data item that itself is not considered as a confidential may, when combined with other data, change it's nature completely for example by allowing identification and linking of a single person to confidential health information.

In the data warehouses the most often used action is to read the data, although in some cases it may be useful to be able to apply update actions. However, considering the nature of the data warehouses and their typical use with together of the intention of protecting the data security and privacy it is justified to narrow the scope to read of the data like was done in the work of Férnandez-Medina et.al. [4]. We can also argue that when we're dealing with the data security, we can start with securing the access to the data independent if that access is to read, modify and store the data. After access rights to the data have been solved, then the type of the action (read, write or modify) can be solved as a second step. However, when analyzing the data any modifications that would write back to the source data are rarely needed, as was stated by E. Fernandez-Medina et.al. [4]. It seems that in several sources the classification of the data plays crucial role as a base for the mechanism on protecting the data. Classification of the data is important so that the nature of the information is understood. This is required both for the utilization of the data in analyses, but also as a base for security planning. In order to understand what restriction needs to be applied one needs to understand the nature of the data. Further, the shared ontology between potentially heterogeneous data sources makes it easier to make meaningful combinations of information, and to ensure that data security aspects are understood and valued similarly by data source owners and the people who are implementing and using the data warehouse system. For example the work of N. Szirbik et.al. [8] gives a good overview on this topic.

In the literature there exists different kind of classifications for technologies that are used to protect privacy within the data. For example the work of Y. Liu et.al. [5] introduces the classification in three categories: Query restriction, Data

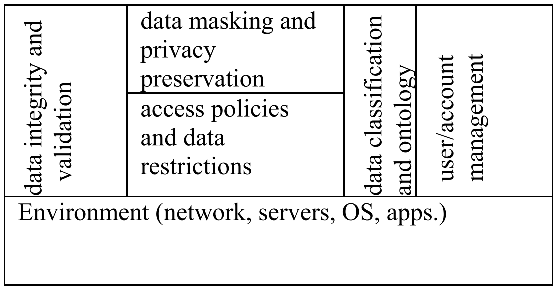
data integrity and validation
data masking and privacy preservation
access policies and data restrictions
data classification and ontology
user/account management
Environment (network, servers, OS, apps.)

Figure.1 Overview of the security areas of data warehouse implementation.

In this paper we consider the first approach as a part of "access policies and data restrictions" presented in Figure 1, and latter two categories as a part of "data masking and privacy preservation". The classification used in the work of Y. Liu et.al. [5] is based entirely on examining the phase when the data mining query is performed, and the classification presented in this paper takes a view from the overall security implementation process perspective.

4.2 Framework for security areas

Data warehouse security, just like the security of any other information system, consists of several layers that all needs to be taken care of in order to achieve proper security level. However, as was described in Section 4.1 data warehouse has some aspects that are specific for this type of the system and needs to be considered in addition to the regular information system security practices. These needs can be classified into four security areas marked within the dotted line in Figure 1 and described in the following chapters. Figure 1 gives an overview of the security areas that needs to be taken into account the data warehouse implementation. The dashed box isolates the security areas that do require considerations specific to the data warehouse environment. The security areas are described below:

- **Data integrity and validation**; this contains the need for ensuring that the data fed to the warehousing system is valid and accurate. This covers also the actions that need to be taken care of when combining information from multiple sources (e.g. confirm compatible semantics and scaling of data values). As the life cycle of the data warehouse is expected to be long special care must be taken in protecting the integrity and validity of the data content.
- **Data masking and privacy preservation**; this refers to the need of ensuring that the privacy and confidentiality needs are fulfilled and only proper level of data details are made available from the data warehouse instead of exposing any more details than have been defined. This

may also mean that the level of the data details is already filtered when the data is brought into the data warehouse.

- **Access policies and data restrictions**; this refers to more basic approaches where the protection of the data is done with access limitations. Typically this means that the data warehouse itself will contain all the data and the protection is based on the access policies and trust on the data warehouse administration. The access policies and data restrictions are also base for the auditing methods.
- D**ata classification and ontology**; understanding the nature of the data stored within the system and applying proper classification is the base for implementing all security needs and maintaining the desired security level through the life cycle of the system. The user/account management and basic environment related security issues not included within the dashed rectangle of Figure 1 can not be forgotten in data warehouse design, but they are considered as a base activities for all information system security. The extension of the user profiles regarding their roles and connection of the roles to the data are included in the "access policies and data restrictions" and "data classification and ontology" areas. The security area named as "users/account management" in Figure 1 refers to the environment level accounts, which often are different from the data warehouse user accounts but still can not be omitted.

4.3 Applying techniques to security areas

As has been stated for example in the work of A. Rosenthal et.al. [7], the most simple approach to the securing the data is to derive the data warehouse access permissions from the source data and treat the data warehouse and source database as a one distributed database. This approach, however, may not be sufficient to protect the confidentiality of the data as it permits the transfer and copying of the complete data, which may also conflict with the data protection legislation. Also, after cross referencing and combining the data resulting information typically exceeds the level of the originating data sources. Further, the approach relies heavily to honesty of the data warehouse operators. It is also possible to apply cryptographic approaches for masking the data content, but those appear to be computationally challenging and thus often impractical when analyzing large quantities of data [2]. Another rather primitive approach is to use trusted third party to aggregate and analyze the data, and then return only the results to data warehouse customers. This approach adds some protection to the data as it's not shared among the parties, but raises some concerns on the level of trust into third parties as was discussed in the work of F. Emekci et.al. [2].

It is also possible to share the information among multiple sources and still retain some level of privacy. Mechanisms that allow this approach are based on either masquerading the data with random substitutes (e.g. replacing personally identifiable information like social security numbers with randomized IDs) and by deriving the classification or summary information from the data itself and then exchanging only the summary information. For the data classification problem the work on F. Emekci et.al. [2] offers a decision tree implementation solution that allows a decision tree creation over multiple sources without compromising the privacy of the source data. When applying perturbation techniques (e.g. by adding random noise in order to masquerade the details or using data swapping techniques) care needs to be taken to protect the results from the statistical biases or even actual errors that may occur as a result of source data manipulations. There are, however, techniques that maintain the statistically accurate while protecting the data privacy details, like is presented in [5] where approach is based on creating summary data for analyses.

All the techniques used to protect the actual data do require that the semantics (ontology) of the data is well-defined and understood. The classification of the data creates the base against which the decisions regarding the sensitivity and required privacy of the data can be based on. In order to be really useful the classification must be based on the shared structures and logic that can be equally interpreted by different parties. Work of Fernandez-Medina et.al. [3] and [4] provide generalized approach in a form of UML extensions for defining the security related classification of the data. The approach takes into account both the nature of the data and the role of the user utilizing the data. Understanding the potential changes in the data security requirements through the time is possible only when the semantics of the data is documented. The periodic re-evaluation of the potential changes in the security needs, threats and then updating the risk-analyses and protection methods alike is part of the system security process. Again, for the more detailed description of the general information system security process I refer to the SSE-CMM documentation [1]. Well defined data ontology creates the base also for ensuring that data integrity, although the integrity also depends on the validation of the data sources used to fed the information into the data warehouse.

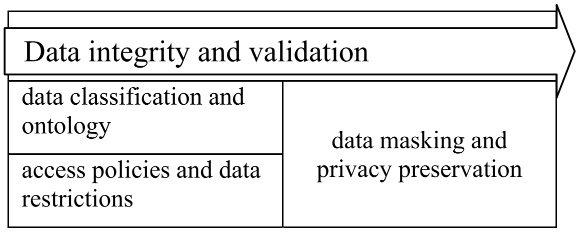
Data integrity and validation
data classification and ontology
access policies and data restrictions
data masking and privacy preservation

Figure .2 Relational timing of the security framework's data warehouse specific activities.

Data integrity and validation needs to run in parallel with above-mentioned activities. The complexity of the validation and integrity depends very much on the number of the data sources used to collect the data warehouse content.

Figure 2 gives an overview of the relational timing (running from left to right) of the security framework areas and activities that are considered specific for the data warehouse implementation. Figure 2 emphasis the facts that the data integrity and validation needs to be taken into account through the whole life span of the data warehouse project. Also, the data classification and access policies/data restrictions needs to be considered right from the beginning, and these two areas form the base for data masking and privacy protection work.

6. CONCLUSIONS

In previous section we have seen a suggested overall security framework approach for securing the data warehouse server, and especially the information content of it. Also the relational timing of the security related activities have been presented. The classification of the data is crucial as it creates the base for all data warehouse specific security activities. Together with the data also the users needs to be classified. Combination of these two classifications do create sound base for defining the access policies and data restrictions. In the examined literature there are approaches for both of these tasks, but in order to be really useful a commonly shared classification approach needs to be used. Unification of the data ontology and user classification is a starting point for a comprehensive data warehouse security. Various data masking methods are required when the data content is shared among the entities that do not completely trust each others, or when for example legislation requires hiding of the specified data details. Data integrity and the validation of the data must be considered throughout the whole implementation process. In addition to the data warehouse specific security considerations that often deal only with the data and it's semantics the basic information system security needs to be taken care of. This means also defining the security maintenance process for the data warehouse, and this process needs to be build in a way that takes also data warehouse specific security requirements into consideration as a integral part of the process. The security areas presented in Section 4.2 and relational timing (as presented in Figure 2) form a general framework that supports the work with data warehouse security. This framework itself is not sufficient, but it helps to structure and organize the security related work.

REFERENCES

- Systems Security Engineering Capability Maturity Model SSE-CMM - Model Description Document v3.0. Carnegie Mellon University, 2003. http://www.ssecmm. org/model/model.asp.
- F. Emekci, O. Sahin, D. Agrawal, and A. E. Abbadi. Privacy preserving decision tree learning over multiple parties. Data & Knowledge Engineering, 63:348–361,2007.
- E. Fernandez-Medina, J. Trujillo, R. Villarroel, and M. Piattini. Access control and audit model for the multidimensional modeling of data warehouses. Decision Support Systems, 42:1270–1289, 2006.
- E. Fernandez-Medina, J. Trujillo, R. Villarroel, and M. Piattini. Developing secure data warehouses with a uml extension. Information systems, 32:826–859, 2007.
- Y. Liu, S. Y. Sung, and H. Xiong. A cubic-wise balance approach for privacy preservation in data cubes. InformationSciences, 176:1215–1240, 2006.
- Oracle. Security and the data warehouse.Oracle While Paper, 2005. http://www.oracle.com/technology/products/bi /db/10g /pdf/twp_bi_dw_security_10gr1_0405.pdf.
- Rosenthal and E. Sciore. View security as the bases for data warehouse security. In Proceedings of the International Workshop on Design and Management of DataWarehouses (DMSW 2000), pages 8:1–8:8, 2000.
- N. Szirbik, C. Pelletier, and T. Chaussalet. Six methmodological steps to build medical data warehouses for research. International Journal of Medical Informatics, 75:683–691, 2006.
- S. Warigon. Data warehouse control & security - seven-step program to secure database warehouses. findarticles.com, 1998. http://findarticles.com/p/articles/mi_m4153/is_ n1_v55 /ai_20568160.

BIOGRAPHY

V.M.Navaneethakumar is currently working as Assistant Professor in the Department of M.C.A, K.S.R College of Engineering, Tiruchengode. His area of interest is Data Warehousing and Mining, Network Security. He has attended various workshops regarding data mining. He has presented papers are various seminars and conferences in the field of data mining.