2

Security Issues for Data Warehousing and Data Mining

B. M. Thuraisingham
The MITRE Corporation
Burlington Road
Bedford, MA 01730
U.S.A.
Tel: 617-271-8873
Fax: 617-271-2780
Email: thura@mitre.org

Abstract

This paper describes security issues for data warehousing and data mining. It first provides an overview of data warehousing and security issues for data warehouses. Then a discussion of data mining , security implications of data mining, as well as data mining as a tool to handle security problems are given.

Keywords

Data warehouse, Data mining, Security, Heterogeneous database integration, Statistical databases, Inference problem, Auditing

1 INTRODUCTION

Data warehousing and data mining are two terms that have become an essential part of data management technology. Having a data warehouse for managing the data is becoming a necessity with many enterprises. Several organizations are building their own data warehouses. Commercial database system vendors are marketing data warehousing products. In addition, some companies are specializing in developing data warehouses. The idea behind a data warehouse is that it is often cumbersome to access data from multiple and possibly heterogeneous databases. Several processing modules need to cooperate with each other for processing a query in a heterogeneous environment. Therefore, a data warehouse will bring together the essential data from these diverse data sources. This way the users need to query only the warehouse. In addition, a data warehouse also often

P. Samarati et al. (eds.), *Database Security*

contains information such as summary reports and aggregates that are determined by the applications using the warehouse and the types of queries posed.

A related technology, which is used to convert the data in the warehouse as well as in other databases into some useful information is data mining. That is, data mining is the process of posing a series of appropriate queries to extract information, often previously unknown, from large quantities of data in the database. Data mining technology is a combination of various other technologies including statistics, machine learning, database management, and parallel processing. Types of data mining include classification, association, and sequencing. For example, data mining by association implies detecting the following pattern: whenever John travels to London, Peter also travels with him.

The developments in data warehousing and data mining technologies have resulted in additional security concerns. For example, can information be deduced from the use of various data mining tools? What are the appropriate auditing procedures for data warehouses? This paper will discuss security issues for data warehousing and data mining. First it will describe security issues for data warehouses. In particular, security for building the warehouse, as well as querying the warehouse will be addressed. The second half of the paper will address data mining. In particular, the security threats due to data mining, some techniques for handling these threats, as well as the use of data mining as a tool to handle security problems will be presented.

The organization of this paper is as follows. In section 2 we will first provide an overview of data warehousing and then describe security issues. In section 3 we will first describe data mining and then discuss the relationships between data mining and security. In particular, security implications of data mining with respect to the inference problem as well as the use of data mining techniques to handle security problems will be discussed. Summary and future considerations are given in section 4. For some background information on data warehousing we refer the reader to [INMO93]. An overview of data mining is given in [IEEE93]. This paper is based on the invited talk presented at the 10th IFIP Working Conference in Database Security in July 1996. This talk will also be given as a panel presentation at the 19th National Information Systems Security Conference in October 1996. This paper is an enhanced version of the extended abstract published in [THUR96].

2. DATA WAREHOUSING AND SECURITY

2.1 Overview of Data Warehousing

As stated by Inmon [INMO93], a data warehouse is subject-oriented, persistent, and time variant. It is used to help in the decision support process. There are two aspects to data warehousing. One is building the warehouse and the other is querying the warehouse. Many of the commercial tools focus on structuring the warehouse in such a way so that query processing can be facilitated. Building the warehouse from heterogeneous data sources is in the research stage. Figure 1 illustrates an example data warehouse. Here, data from four databases are integrated into the warehouse. Each database is managed by a

database management system (DBMS). Assuming that the DBMSs A, B, C, and D have information on employees, projects, medical benefits, and salaries,. respectively, the warehouse may have correlations between say employees, projects, medical benefits, and salaries. Also, as illustrated in figure 1, a warehouse is managed by a warehouse manager.

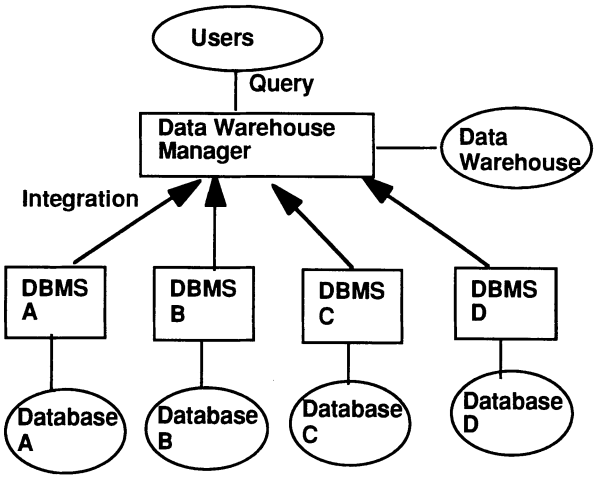

Figure 1 Example Data Warehouse

Several technologies have to be integrated to developing a data warehouse. These include data modeling, distributed databases, heterogeneous database integration, statistical databases, parallel processing, database design, access methods, integrity, and security. Inmon [INMO93] has outlined multiple steps to developing a data model for a warehouse. For example, a three level modeling process can be carried out as follows. At the highest level, enterprise models as well as corporate models have to be developed. At the middle level subject dependent models have to be developed. At the lowest level, the physical models have to be developed. Data distribution technologies can be used to distributing the data warehouse. Such a warehouse is needed if the warehouse data is distributed. Heterogeneous database integration technologies can help in integrating the data from the multiple databases into the warehouse. Statical database technology may be used to process queries based on aggregates, averages, and sums. Parallel processing techniques may be used to enhance the performance of the query processing strategies. Database design techniques and access methods may be used to structure the data warehouse in such a way to facilitate retrievals. Integrity techniques are needed to determine the quality of the data in the warehouse. Security techniques ensure that the data in the warehouse is accessed by authorized individuals.

A discussion of many of these technologies to building a warehouse are provided in [INMO93]. An overview of data warehouse in relationship to a data management

framework is given in [THUR97]. In section 2.2 of this paper we will focus on security technology for a data warehouse.

2.2 Security for Data Warehouses

In this section we discuss some of the security issues for a data warehouse as given in [THUR96]. Security has to be major consideration in both integrating the databases to form the warehouse as well as to retrieve information from the warehouse. This is illustrated in figure 2. In this example, it is assumed that two of the four DBMSs are multilevel secure (MLS) DBMSs.

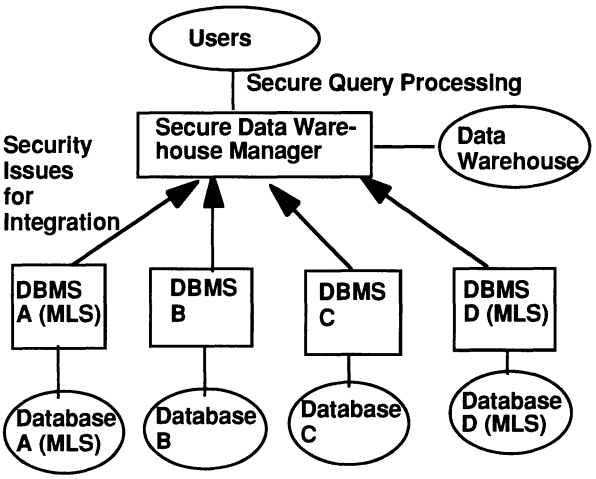

Figure 2 Security for Data Warehouse

Research on security for integrating heterogeneous databases will contribute significantly toward exploring security for building a warehouse. For example, when integrating multiple heterogeneous databases to build a warehouse, one may have to deal with multiple security policies. A major issue here is in dealing with inconsistent policies. One needs to resolve various conflicts and generate an appropriate security policy for the warehouse. Work has been reported in [BLAU95] on security for federated database management. One needs to examine such work in developing a security policy for the warehouse. Other issues include the security impact on (1) the data model for the warehouse, (2) generating appropriate update requests on the warehouse from the updates made to the individual databases, (3) developing the metadata for the warehouse, and (4) maintaining security for mappings and transformations between the individual data sources and the warehouse. Also, if some of the individual database management systems are multilevel secure as illustrated in figure 2, then there may be some additional security concerns such as the trusted computing base for the warehouse.

There are many other important security considerations in building a warehouse. This is due to the fact that when integrating heterogeneous databases, one does not assume the development of a data repository whereas in the case of a warehouse, there is usually a physical data repository. An example security concern is the following. A warehouse database may give summary information. This summary information is often derived from the data in the heterogeneous databases. It is important that one does not deduce sensitive information in the heterogeneous databases from the summary information in the warehouse. Therefore, statistical database security as well as the developments on the inference and aggregation problems will also play an important role in securing the warehouse.

The previous discussion focussed on the security issues for building a warehouse from heterogeneous data sources. Security should also be maintained while the warehouse is in operation. For example, actions on the warehouse need to be audited. The question is can the traditional database auditing techniques be used for the warehouse? Other issues include the following. Should there be a special warehouse administrator and warehouse security officer? What is the relationship between the warehouse administrator / security officer and the administrators / security officers of the heterogeneous databases used to develop the warehouse? Can appropriate query modification techniques be developed for the warehouse? Should the access control rules enforced on the warehouse be taken into consideration when structuring the warehouse depending on the queries? What is the security impact on the access methods and index strategies? How can views be used as a protection mechanism for the warehouse? Research is being conducted on addressing some of these issues. For example, Stanford University's Data Warehousing project [ZHUG95] is investigating techniques for materialized views for the warehouse as well as maintaining the views as the data sources get updated. Enforcing security through views in a warehousing environment needs more work.

Administering a data warehouse has some problems. For example, consider the integration of multiple heterogeneous data sources in building a warehouse. When the databases get updated, one may want trigger mechanisms to be enforced in such way that the warehouse gets updated automatically. However, in many cases due to autonomy, the individual database administrators may bot want triggers to be enforced on their data. This poses a major problem for updating a warehouse. With respect to auditing, Inmon [INMO93] has stated that it may not always be a good idea to audit a warehouse. For example, if a warehouse has to be audited, then some irrelevant data may have to be stored in the warehouse.

Many of the issues discussed here show that security for data warehousing is a combination of security for database management systems, statistical databases and integrating heterogeneous databases. More research is needed to determine the security issues specific to the warehouse before solutions for securing a warehouse can be developed.

3. DATA MINING AND SECURITY

3.1 Overview of Data Mining

Data mining is a technology that builds on data warehousing technology. While a data warehouse structures the data in such a way to facilitate query processing, data mining tools can be applied on a data warehouse. data mining is the process of obtaining information previously unknown by posing various queries. The relationship between data warehousing and data mining is illustrated in figure 3. Note that one does not have to build a warehouse to carry out data mining. However, a warehouse helps to structure the data in such a way to help the data mining process.

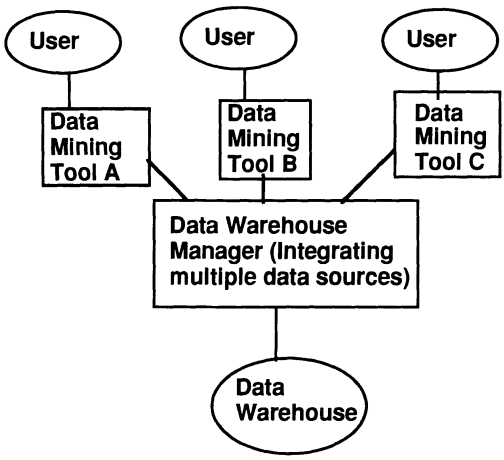

Figure 3 Relationship between Data Warehousing and Data Mining

There are various types of data mining. These include classification, association, and sequencing. An example of classification is as follows. Suppose a marketing organization has information that everyone in its list who live in San Francisco own a house costing more than 200K. Suppose John lives in San Francisco and the cost of his house is unknown. Then the marketing organization can classify John into the same group and assume that John's house also costs more than 200K Association is the commonly used type of data mining Here association and correlations are made between the data. For example, if John and Paul travel to London, then Mart will also be traveling to London. An example of sequencing is as follows. John always goes to the chemist after he goes to the bank on Saturdays. Note that the types of data mining are different from data mining techniques. A data mining technique is a particular method used to mine the data. Some techniques include inductive logic programming, fuzzy logic, and rough sets.

Like data warehousing, there are several technologies that have to be integrated for data mining. These include artificial intelligence, database management (including data warehousing), machine learning, statistics, and parallel processing. For example, parallel processing techniques speed up the data mining algorithms and machine learning techniques are used for detecting patterns through inductive inference.

There are several challenges to data mining., These include performance, dealing with redundant and incomplete data, and security. Another challenges is to determine which data mining technique to use. An overview of data mining in relationship to a data management framework is described in [THUR97]. In section 3.2 we describe the relationships between data mining and security as given in [THUR96].

3.2 Data Mining and Security

Recently there has been much interest on exploring the relationship between data mining and security. Some preliminary ideas were discussed at the data mining special session that took place at the Ninth IFIP 11.3 Working Conference on Database Security in 1995 [LIN95]. More details on this topic have been given in [MARK96]. There are two aspects to data mining and security. One is that data mining techniques can be be applied to handle problems in intrusion detection and database auditing. In the case of auditing, the data to be mined is the large quantity of audit data. One may apply data mining tools to detect abnormal patterns. For example, suppose an employee makes an excessive number of trips to a particular country and this fact is known by posing some queries. The next query to pose is whether the employee has associations with certain people from that country. If the answer is positive, then the employee's behavior is flagged.

Current data mining tools are sufficiently advanced so that one could start applying them to detect intrusions and abnormal behavior. However, many of these tools work on structured databases such as relational databases. Therefore, the data to be examined has to be first converted to structured format so that these tools can be applied. Recently, an investigation was reported in [GRIN96] where the idea is to place network intrusion data to be mined in various repositories. This will enable researchers as well as developers to test the algorithms and tools on these common repositories to see if suspicious behavior could be determined. In other words, the network intrusion data sets to be explored (e.g. mined, visualized, etc.) will enable researchers to compare various approaches to data exploration. Data mining, visualization, and any other collection of tools as well as the human expert may be used in this process. The goal is to determine what tools can help in discovering real time suspicious behavior. Research is also beginning on data mining for unstructured data such as text and images. As developments are made, one could expect to have tools to apply on unstructured audit data.

The second aspect to data mining and security is the inference problem. That is, while the previous example shows how data mining tools can the used to detect intrusions and abnormal behavior, the next example shows how data mining tools can be applied to cause security problems. Consider a user who has the ability to apply data mining tools. This user can pose various queries and infer sensitive hypothesis. That is, the inference problem occurs via data mining. There are various ways to handle this problem. One

approach is as follows. Given a database and a collection of data mining tools, apply the tools to see if sensitive information can be deduced from the unclassified information legitimately obtained. If so, then there is an inference problem. Such an approach may be carried out periodically as the database gets updated. There are some issues with this approach. One is that we are applying only a limited set of tools. In reality, the user may have several other data mining tools available to him. Furthermore, it is impossible to cover all ways that the inference problem could occur.

Another approach which is much harder to accomplish is to apply a data mining-based inference controller during run time as illustrated in figure 4. This means when a user poses a query, determine whether by releasing the results an inference problem could occur. The inference controller in this approach will be based on a collection of data mining techniques such as classification, association, and sequencing. For example, suppose we want to protect the fact that whenever Peter travels to London, so does John. This may be due to the fact that Peter is working on a classified project and we want to hide the fact that John also works on the same project. By observing the pattern that Peter and John always travel together to London, one may infer the sensitive fact through association. The inference controller should detect the fact that a user may be able to infer this sensitive information and not release certain responses to the user.

Figure 4 Inference Controller

Building an inference controller based on the second approach is extremely difficult as theory and foundations for data mining are yet to be developed. While there is some work on the relationship between inductive logic programming and data mining, the research is still in the preliminary stages. Current data mining techniques are rather ad-hoc and therefore it is nearly impossible to build such an inference controller. Note that the work reported in [THUR95] takes a similar approach to handle the inference problem, but focuses only on deductive reasoning. Data mining techniques are far more complex than deductive reasoning.

The research reported in [CLIF96] shows much promise on developing techniques to handle the inference problem based on the first approach. For example, it has been shown that by applying various data mining tools that exist today, one could deduce some potentially sensitive information. The challenge then is to develop techniques to handle this problem. Some of the methods that are being explored include giving partial answers to queries, introducing additional information and noise into the responses, and giving answers to different but related queries. Research in this area is just beginning.

4. SUMMARY AND FUTURE CONSIDERATIONS

This paper has provided an overview of data warehousing as well as data mining, and described some security issues. For data warehousing, a discussion on security for integrating the databases to form the warehouse as well as security for querying the warehouse are given. For data mining, both security implications of data mining as well as using data mining tools to handle security problems are discussed.

The security issues discussed in this paper for data warehousing and data mining are based on a very preliminary investigation. However, the paper has identified some security challenges that have to be addressed for an operational data warehouse. With respect to security implications of data mining, the problems are much harder. This paper has only identified the problem. Much research is needed before viable solutions are developed. However, work on the use of data mining tools to address security problems can begin with existing technology. In summary, security for data warehousing and relationships between data mining and security have numerous research issues that need to be investigated.

REFERENCES

[BLAU95] B. Blaustein, C. McCollum, A. Rosenthal, K. Smith, and L. Notargiacomo, "Autonomy and Confidentiality: Secure Federated Data Management," Second International Conference on Next Generation Information Technologies and Systems, Naharia, Israel, June 1995.

[CLIF96] C. Clifton, and D. Marks, "Security and Privacy Issues for Data Mining," Proceedings of the ACM SIGMOD Conference Workshop on Data Mining, Montreal, Canada, June 1996.

[GRIN96] G. Grinstein, "Data Exploration through Mining and Visualization," To be published in the Proceedings of the IEEE Visualization'96 Conference, San Francisco, CA October 1996.

[IEEE93] Special Issue on Data Mining, IEEE Transactions on Knowledge and Data Engineering, December 1993 (Ed: N. Cercone and M. Tsuchiya)

[INMO93] W. H. Inmon, "Building the Data Warehouse," John Wiley and Sons, 1993

[LIN95] T.Y. Lin, D. Marks, T. Hinke, and B. Thuraisingham, "Data Mining and Security," Special Session at the 9th IFIP 11.3 Database Security Workshop, N.Y. August 1995.

[MARK96] D. Marks, "Inference in MLS Databases," IEEE Transactions on Knowledge and Data Engineering, February 1996.

[THUR95] B. Thuraisingham and W. Ford, "Security Constraint Processing in a Multilevel Secure Distributed Database System," IEEE Transactions on Knowledge and Data Engineering, April 1995.

[THUR96] B. Thuraisingham, "Data Warehousing, Data Mining, and Security," Published in the Proceedings of the 10th IFIP Working Conference in Database Security 1996; also published in the Proceedings of the 19th National Information Systems Security Conference, 1996.

[THUR97] B. Thuraisingham, "Data Management Systems Evolution and Interoperation," To Appear, CRC Press, 1997.

[ZHUG95] Y. Zhuge, H. Garcia-Molina, J. Hammer, and J. Widom, "View Maintenance in a Warehousing Environment," Proceedings of the ACM SIGMOD Conference, San Jose, CA, May 1995.

ACKNOWLEDGMENTS

I thank Don Marks of the Office of INFOSEC Computer Science, Department of Defense, and Cathy McCollum, Chris Clifton, Georges Grinstein, and Ken Smith all of The MITRE Corporation for comments and/or inputs to this abstract.

DISCLAIMER

The views and conclusions reported in this paper are those of the author and do not reflect the policies and procedures of the MITRE Corporation or of the U.S. Government.