Security in Data Warehousing*

(Invited Talk)

Bharat Bhargava
CERIAS and Computer Science Department
Purdue University, West Lafayette,
Indiana 47906
bb@cs.purdue.edu

Data warehouse [2,4,5,6] is an integrated repository derived from multiple source (operational and legacy) databases. The data warehouse is created by either replicating the different source data or transforming them to new representation. This process involves reading, cleaning, aggregating and storing the data in the warehouse model. The software tools are used to access the warehouse for strategic analysis, decision-making, marketing types of applications. It can be used for inventory control of shelf stock in many departmental stores.

Medical and human genome researchers can create research data that can be either marketed or used by a wide range of users. The information and access privileges in data warehouse should mimic the constraints of source data. A recent trend is to create web-based data warehouses and multiple users can create components of the warehouse and keep an environment that is open to third party access and tools. Given the opportunity, users ask for lots of data in great detail. Since source data can be expensive,
its privacy and security must be assured. The idea of adaptive querying can be used to limit access after some data has been offered to the user. Based on the user profile, the access to warehouse data can be restricted or modified.

In this talk, I will focus on the following ideas that can contribute towards warehouse security.

1. Replication control

Replication can be viewed in a slightly different manner than perceived in traditional literature. For example, an old copy can be considered a replica of the current copy of the data. A slightly out-of date data can be considered as a good substitute for some users. The basic idea is that either the warehouse keeps different replicas of the same

* This research is partially supported by CERIAS and NSF under grant numbers 9805693-EIA, CCR-9901712.

Y. Kambayashi, M. Mohania, and A M. Tjoa (Eds.): DaWaK 2000, LNCS 1874, pp. 287-289, 2000.

items or creates them dynamically. The legitimate users get the most consistent and complete copy of data while casual users get a weak replica. Such replica may be enough to satisfy the user's need but do not provide information that can be used maliciously or breach privacy. We have formally defined the equivalence of replicas [7] and this notion can be used to create replicas for different users. The replicas may be at one central site or can be distributed to proxies who may serve the users efficiently. In some cases the user may be given the weak replica and may be given an upgraded replica if wiling to pay or deserves it. Another idea related to this is the idea of witness that was discussed in [1].

2. Aggregation and Generalization

The concept of warehouse is based on the idea of using summaries and consolidators. This implies that source data is not available in raw form. This lends to ideas that can be used for security. Some users can get aggregates only over a large number of records where as others can be given for small data instances. The granularity of aggregation can be lowered for genuine users. The generalization idea can be used to give users a high level information at first but the lower level details can be given after the security constraints are satisfied. For example, the user may be given an approximate answer initially based on some generalization over the domains of the database. Inheritance is another notion that will allow increasing capability of access for users. The users can inherit access to related data after having access to some data item.

3. Exaggeration and Misleading

These concepts can be used to mutilate the data. A view may be available to support a particular query, but the values may be overstated in the view. For security concern, quality of views may depend on the user involved and user can be given an exaggerated view of the data. For example, instead of giving any specific sales figures, views may scale up and give only exaggerated data. In certain situations warehouse data can give some misleading information [3]; information which may be partially incorrect or difficult to verify the correctness of the information. For example, a view of a company's annual report may contain the net profit figure including the profit from sales of properties (not the actual sales of products).

4. Anonymity

Anonymity is to provide user and warehouse data privacy. A user does not know the source warehouse for his query and warehouse also does not know who is the user and what particular view a user is accessing (view may be constructed from many source databases for that warehouse). Note that a user must belong to the group of registered users and similarly, a user must also get data from only legitimate warehouses. In such cases, encryption and middleware are to be used to secure the connection between the users and warehouse so that no outside user (user who has not registered with the warehouse) can access the warehouse.

5. User Profile Based Security

User profile is a representation of the preferences of any individual user. User profiles can help in authentication and determining the levels of security to access warehouse data. User profile must describe how and what has to be represented pertaining to the users information and security level authorization needs. The growth in warehouses has made relevant information access difficult in reasonable time due to the large number of sources differ in terms of context and representation. Warehouse can use data category details in determining the access control. For example, if a new employee would like to access an unpublished annual company report, the warehouse server may deny access to it. The other alternative is to construct and return a view to her, which reflects only projected sales and profit. Such a construction of view may be transparent to the user. A server can also use the profile to decide whether the user should be given the access to associated graphical image with the data. The server has the option to reduce the resolution or the quality of images before making them available to users.

References

[1] Arnon Rosenthal and Edward Sciore, View Security as the Basis for Data Warehouse Security, Proceedings of the International Workshop on Design and Management of Data Warehouse (DMDW'2000), Sweden, June, 2000.

[2] Arun Sen and Varghese S. Jacob, Introduction – Guest Editorial, Communication of ACM, Vol. 41, No. 9, Sept. 1998.

[3] Demillo, R. A., D. P. Dobkin, R. J. Lipton, Even Databases that Lie can be Compromised, IEEE Transaction on Software Engineering, SE 5:1, pp. 73-75, 1978.

[4] Jim Scott, Warehousing Over the Web, Communication of ACM, Vol. 41, No. 9, Sept. 1998.

[5] Stephen R. Gardner, Building the Data Warehouse, Communication of ACM, Vol. 41, No. 9, Sept. 1998.

[6] Sunil Samtani, Mukesh Mohania, Vijay Kumar, Yahiko Kambayashi, Recent Advances and Research Problems in Data Warehousing, Advances in Database Technologies, ER Workshop, LNCS Vol. 1552. 1998.

[7] Meeliyal Annamalai and Bharat Bhargava, Defining Data Equivalence for Efficient Image Retrieval in a Distributed Environment, In Proceedings of the Second World Conference on Integrated Design and Process Technology, Austin, Texas, Dec. 1996.