An Extended Conceptual Modeling for ETL Processes in Privacy Preserving Data Mining

Kiran P., Sathish Kumar S., and Kavya N. P.

***Abstract*—Extraction transformation load is an important stage of a data warehouse process. It involves retrieval of information from multiple sources, applying transformation to map with the schema of the data warehouse. Privacy preserving data mining has become a prerequisite for all the present data warehouse projects. In this paper we provide a foundation on embedding privacy to conceptual modeling by indicating additional notations which ensures faster design.**

***Index Terms*—ETL, Data Warehouse, PPDM, Conceptual Modeling**

I. Introduction

Extraction Transformation Load (ETL) is an important component of data warehouse. Extraction deals with the retrieval of data from various sources. These sources of data can be of different format and different location in space [1]. Each of these locations must be analyzed before it is loaded on to the data warehouse. The most important and complex part of ETL is the Transformation phase. Transformation can be visualized as the change that must be made on to the source data before it is loaded on to the data warehouse. There are various methods that are used for transformation. Usual transformation include deletion of a column, conversion of values from one representation to the other, aggregation of values for better retrieval, removal of null values, converting multiple representations in to a single format and joining of multiple tables. The last phase is the loading phase which copies the information from the transformed data to the warehouse. The overall process consists of three major steps first identification of the source targets second is the actual data representation also called conversion and the last phase is moving resultant to the target location. These tasks can be done in intermediate stages or as a single logical execution. Each of these stages indicated separately will help improve the overall representation of Data Warehouse. The initial design of conceptual schema plays an important role and also propagates to other stages. The complexity of extraction transformation and load lies in the mapping of data from source to destination. As indicated by Alkis Simitsis [1] the environment of ETL process can be shown in fig 1. The left side we have source representation where in information is distributed. In the middle we have Data Staging Area (DSA) which encapsulates the major transformation and the last is the Data Warehouse representation.

Privacy Preserving Data Mining (PPDM) is a novel field in data mining. PPDM is concerned with extraction of information from data warehouse without revealing sensitive information of individuals and company privacy details [8], [9], [10]. Present industry consists of database which is distributed across multiple source locations. Most of them doesn't want their data to be revealed so that the confidentiality is lost which is of a great concern. The confidentiality of some attribute within the data base must be ensured. There are two major methods in PPDM first by using cryptographic representation and the other by using heuristic algorithms which ensures that sensitive data is not revealed. Most of the current industry require that there data be secured during transmission and also when the data is present in the data warehouse. In this paper we are using a cryptographic representation called Elliptic Curve Cryptography to ensure confidentiality which can be indicated to some attributes during design and data distortion in heuristic algorithm. ETL is an important stage in Data Warehouse and the processing activities can be indicated by means of diagrams as a part of design. Designing ETL and embedding PPDM representation is the main task of this research.

We present the details of our research as follows. We have discussed the related work in section 2, Different types of transformation methodologies presently available in most of the present ETL tools and additional notations for indicating privacy is discussed in section 3. In section 4, we have taken a motivation example to explain the representation of additional notation that we have used. we state our conclusions and future scope of extensions in section 5.

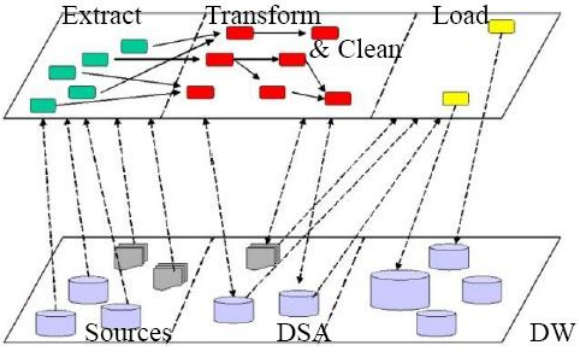

Fig. 1. The Environment of ETL processes.

II. Related Work

In this section we discuss the research efforts, ETL tools

Manuscript received April 1, 2012; revised May 15, 2012.

P. Kiran is with the Visvesvaraya Technological University, Belgaum, Karnataka, India (e-mail: kiranmys@rediffmail.com, Tel:+919845440059, fax:+9108028611882).

S. Sathish Kumar is with the MGR University, Chennai, Tamil Nadu, India.

N. P. Kavya is with the Department of MCA , RNS Instituite of Technology, Bangalore.

that are available along with related expertise. The front end of data warehouse has been driven by conceptual modeling. Conceptual modeling gives an abstract representation of the business scenario which helps the developers to built a robust model. Most of the research in conceptual modeling is characterized by conceptual characteristics of star schema, dimensional modeling and aggregate data marts [2]. Dimensional modeling [4], [5] which is a design concept used for building the data warehouse. Entity relationship modeling [7], [11], [12] is useful in designing conceptual schema of the database and was quite popular in the industry. Later stages we had influence of object-oriented concept, the result of which was UML modeling[14], [15]. Unified modeling Language uses lot of graphical notations for designing conceptual schema. Each of these modeling concepts provides a way to get in to the requirement and analyze it on to the data base without a clear winner. Dimensional modeling is featured by its minimality, understandability and direct mapping on to logical structures. Entity Relationship modeling and unified modeling Language are also used depending on the company and the requirement. Above related work was on conceptual modeling in data warehouse. Rather than concentrating on the entire warehouse few efforts was also made on conceptual modeling for ETL since most of its task are dependent on it. As a first attempt author [16] had separated warehouse conceptual schema and ETL conceptual schema. Author in [1] has given notation but has not indicated representations for PPDM. In this paper we propose a method that introduce additional notation in design which can be used to represent PPDM representations.

III. Conceptual Model

This section gives different notations for conceptual model for ETL activities. These notations are used in high level, user-oriented entities to capture the process of ETL existing notations are shown in fig. 2 PPDM requires additional notation for faster representation which cannot be indicated just as a function, so a specialized PPDM notations are introduced and indicated in fig. 3.

A. Attribute and Concepts

Attribute represents property of an entity and has the same characteristics as in case of ER modelling and the symbol that is used is oval.

Concepts represent an entity with reference to a data base. It is used to represent source data or a file from which information is retrieved. A concept is identified by a name and contains set of attributes.

B. Transformation, Constraints And Notes

Transformation is the process of changing the contents of the data base to a format suitable with respect to the data ware house. Transformation include filtering of unnecessary information like violation of primary key and foreign key references. It is also used for aggregation of data. In general transformation is characterized by finite set of input and finite set of output attributes. The notation for transformation is hexagon named with its corresponding symbol. Constraints are used to apply some conditions on the data if the condition is satisfied then it is moved on to the next level. Constraint is graphically depicted as a set of solid edges starting from the involved attributes and targeting the facilitator transformation. Notes are used to capture extra comments that the designer wishes to include for extra information. The notation used for notes is rectangles with a dog-eared corner.

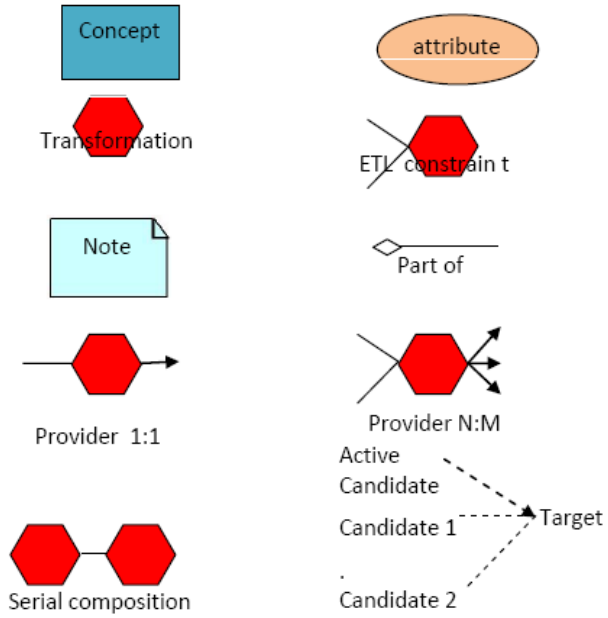

Fig. 2. Existing Notation for the conceptual modelling of ETL activities.

C. Provider Relationship and Serial Composition

A provider relationship maps a set of input attributes to a set of output attributes. There are two types of provider relationship 1:1 and N: M. A 1:1 provider maps all the attribute of source side to target side and the notation is a solid bold directed arrow from the input towards the output attribute. In certain scenarios not all attributes are required either there will be increase in number of attributes or decrease in the attributes depending on the requirement of a data warehouse to facilitate it N:M provider is used and the notation is a solid arrows starting from the providers and targeting the consumers, all passing through the facilitator transformation. Serial transformation is used when there is multiple transformations that must be made on an attribute which cannot be indicated as a single transformation. The notation is a series of transformation symbols interconnected by bold lines.

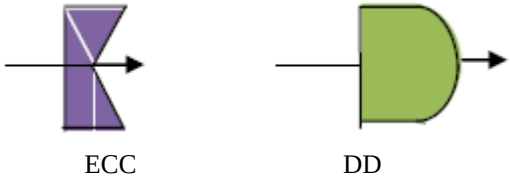

Fig. 3. Additional Notations for conceptual modelling in ETL for PPDM approach.

D. Part-of and Candidate Relationship

Part-of is used similar to UML representation to indicate inclusive but does not have lot of significance in this model. The notation is a edge with a small diamond at the side of

the container object. Candidate relationship is used to indicate which source is used as the candidate for the target representation since multiple sources will have multiple candidates. Among them one of them is called as a active candidate which will help the developer while updating the contents of the warehouse. The notation is with bold dotted lines between the candidates and the target concepts and an active candidate relationship with a directed bold dotted arrow from the provider towards the target concept.

E. ECC and DD Transform

ECC Transform accepts a single attribute and does a cryptographic transformation on an attribute to have resultant as a encrypted form of an attribute. It is applied on an individual attribute separately. The notation for ECC transform is a 'E' with right edges connected. DD transform is similar to ECC only difference is that DD is a data distortion which is denoted by 'D'. these two notation will help the designers to represent privacy issues effectively during design.

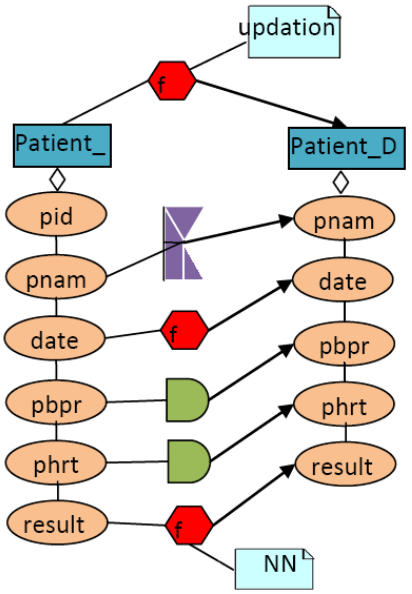

Fig. 4. Conceptual modelling for the motivation example.

IV. Designing Conceptual Modelling for PPDM an Example

We are using a simple patient data base for our modelling. In this transformation we are concentrating more on PPDM representation in modelling. Let us assume that the following patient schema is preset in the source data base.

Patient_S(pid,pname,date,pbpr,phrt,result)

The details are patient-id, patient name, patient blood pressure, patient heart rate and result of the diagnosis. This information is passed on to the data warehouse with the target schema as

Patient_DW(pname, date, pbpr,phrt,res)

The details of transformation as follows pname is to be cryptic so that it is not revealed to the end user. pbpr and phrt to be data distorted before it is stored on to the data base. The conceptual modelling is shown in figure 4. Transformation from source to target is triggered by the updation process. NN function indicates that the attribute is accountable if the result is Not Normal. Date transformation function is required to convert date into the appropriate representation of Data warehouse schema. This representation is simple to design and understand.

V. Conclusion

Conceptual modelling is an important task in software engineering cycle which facilitates better understanding and faster development. In this paper we are using additional symbols to facilitate privacy representation in design which earlier researchers had not considered. This method gives a better way of representing privacy issues. Future directions may concentrate on designing specific symbols for privacy preserving data mining itself and mapping of conceptual to logical representation. Additional research may concentrate on standardization of these notations.

References

[1] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual Modeling for ETL Processes," In *Proc. of the 5th ACM Int*. Workshop on Data Warehouising and OLAP, McLean, USA, pp. 14-21, Nov 2002.

[2] A. Tsois, "MAC: Conceptual data modeling for OLAP," In *Proc. DMDW*, Interlaken, Switzerland, pp. 5.1-5.13, 2001.

[3] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Modeling ETL Activities as Graphs," In *Proc. of the 4th Int. Workshop on Design and Management of Data Warehouses*, Toronto, Canada, pp. 52-61, May 2002.

[4] R. Kimball, "A Dimensional Modeling Manifesto," *DBMS Magazine*. August 1997.

[5] R. Kimbal, L. Reeves, M. Ross, and W. Thornthwaite, "The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing," *Developing, and Deploying Data Warehouses*. John Wiley and Sons, February 1998.

[6] T. B. Nguyen and A. M. Tjoa, "Conceptual Multidimensional Data Model Based on Object Oriented MetaCube," In P*roc. of the 2001 ACM symposium on Applied Computing*, Las Vegas, Nevada, USA, pp.295-300, Mar. 2001.

[7] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Information integration: Conceptual modeling and reasoning support," In *Proc. COOPIS*, New York, USA, pp. 280-291, 1998.

[8] S. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Record*, vol.33, pp 50-57March 2004.

[9] V. S. Miller, "Use of Elliptic Curves in cryptography," A*dvances in Cryptology- Proceedings of CRYPTO'85*, Springer Verlag, pp. 417-426, 1986.

[10] W. L. Du and Z. J. Zhan, "Using randomized response techniques for privacy-preserving data mining," In: *The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 505-510, 2003.

[11] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "A principled approach to data integration and reconciliation in data warehousing," In *Proc. DMDW'99*, Heidelberg, Germany, 1999.

[12] B. Husemann, J. Lechtenborger, and G. Vossen, "Conceptual data warehouse modeling," In *Proc. DMDW, Stockholm*, Sweden, pp. 6.1 –6.11, 2000.

[13] D. L. Moody and M. A. R. Kortink, "From enterprise models to dimensional models: a methodology for data warehouse and data mart design," In *Proc. DMDW*, Stockholm, Sweden, June 2000.

[14] T. B. Nguyen, A. Min Tjoa, and R. R. Wagner, "An Object Oriented Multidimensional Data Model for OLAP," In *Proc. WAIM*, Shanghai, China, June 2000.

[15] J. C. Trujillo, M. Palomar, and J. Gómez, "Applying Object- Oriented Conceptual Modeling Techniques to the Design of Multidimensional Databases and OLAP Applications," In *Proc. WAIM*, Shanghai, China, pp. 83-94, jun. 2000.

[16] M. Bouzeghoub, F. Fabret, and M. Matulovic, "Modeling Data Warehouse Refreshment Process as a Workflow Application," In *Proc. DMDW'99*, Heidelberg, Germany, 1999.