Modelling Extraction Transformation Load embedding Privacy Preservation using UML

Kiran P
Research Scholar
VTU,Belgaum
Karnataka, India

S Sathish Kumar
Research Scholar
MGR University, Chennai
Tamil Nadu,India

Kavya N P
Prof & Head
Dept of MCA,RNSIT
Bangalore,Karnataka,India

ABSTRACT
Extraction Transformation Load plays an important phase in development of data warehouse due its complexity of selecting data from different location and having different structures. The recent industry of data warehouse is driven by Privacy Preserving Data Mining which ensures privacy of sensitive information during Mining and is a requirement of most Data Bases. Current approaches to modelling extraction transformation load do not include privacy representation in Conceptual Modelling. This paper proposes object-oriented approach to model Extraction Transformation Load embedding privacy preservation. The major components of extraction include Data Source, Source Identifier, Retrieval, Join, Privacy Preserving Area and Data Staging Area. All the above mentioned components have been modelled using Unified Modelling Language.

Keywords
Privacy Preserving Data Mining, ETL, Data Stage Area, Privacy Preserving Area , Data Warehouse.

1. INTRODUCTION
Providing security to the data which is retrieved from distributed heterogeneous databases and data mined after it is integrated are one of the leading issues in database research and industry. Extraction Transformation Load (ETL) is an important stage in developing the Data Warehouse. The Extraction process starts with identification of the heterogeneous databases and their format of representation. This information must be transformed, which may include elimination of data based on some conditions, change of format from one representation to the other also called filtering, merging of data from multiple sources and usually in databases we use different primary keys but in data warehouse we require additional keys called surrogate keys so there is a requirement to have an efficient mechanism to assign surrogate keys. The next important task is aggregation of information in order to help the queries and to improve the performance of the Data Warehouse. All these transformed data must be placed in the Data Warehouse which is done in Load phase.

ETL processes are one of the important components of DWs, because erroneous or ambiguous data will give improper results which will affect business decisions made by the manager. Therefore, there is a requirement to design ETL correctly at the initial stages to improve data quality which will improve decision making. Embedding privacy on to the ETL representation plays an important task since most of the present DW requires some of the information to be kept confidential. In this paper, we have used Unified Modelling Language to represent modelling of ETL embedding privacy preservation.

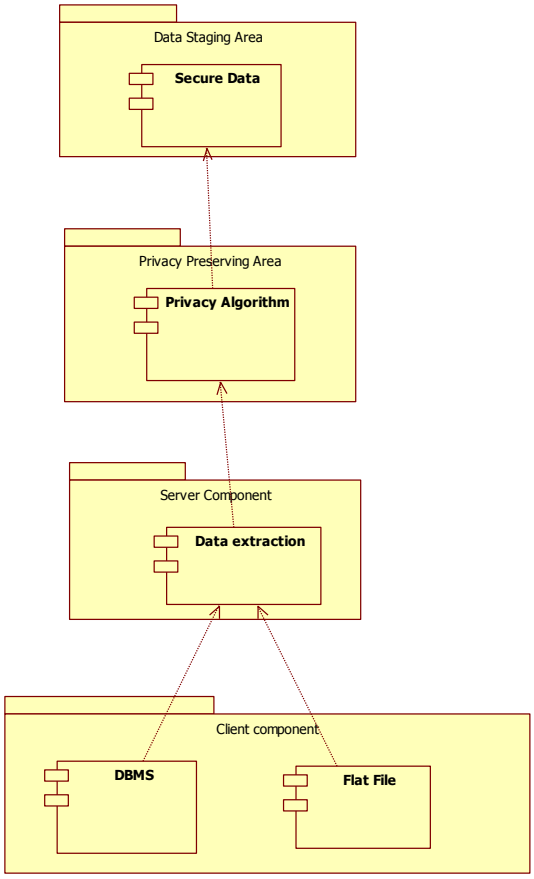

Fig. 1 Deployment Diagram for Architecture of Data Extraction Embedding Privacy Preservation

We present the details of our research as follows. We have discussed the related work in section 2, Architecture of data extraction embedding privacy has been discussed in section 3. In section 4, UML modelling has been indicated. Algorithm for Data extraction embedding privacy has been discussed in section 5. We have taken a motivation example of medical data set to explain the representation and experimental results has been indicated in section 6. We state our conclusions and future scope of extensions in section 7.

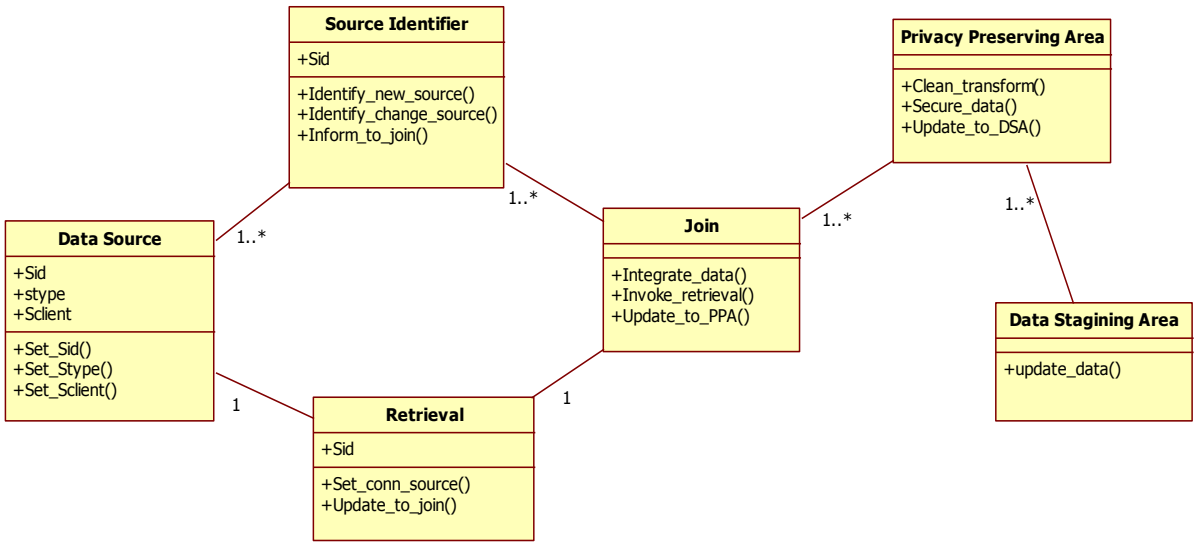

Fig. 2 Class Diagram for ETL embedding Privacy Preservation

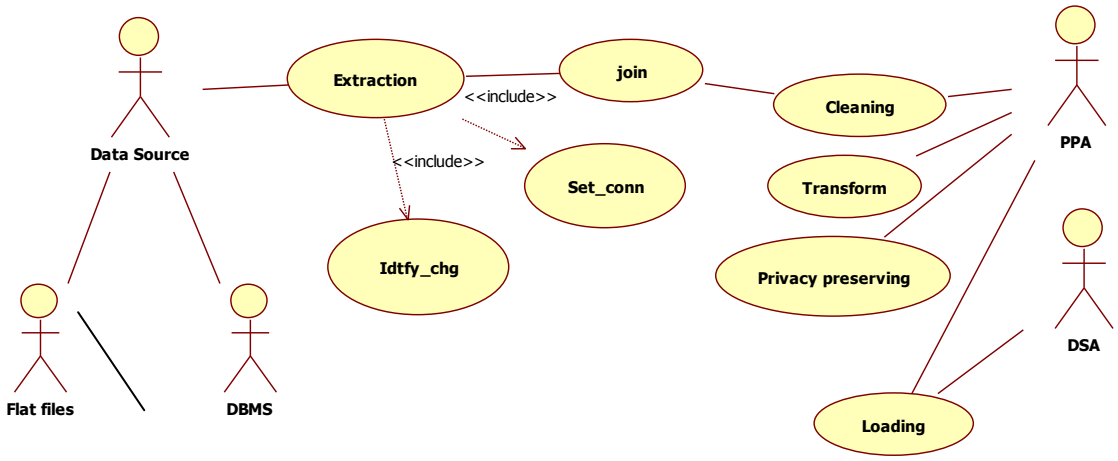

Fig. 3 Use case diagram for ETL embedding privacy preservation

2. RELATED WORK

ETL is an important foundation of building data warehouse[1]. Users extract the required data from the data source, clean the data in accordance with pre-defined model schema and then data will be loaded into the data warehouse. A common management and design tool for ETL , system structure and program framework has been discussed in [1]. Traditional methods of ETL development are very much difficult to meet business requirements so the authors have demonstrated an extended UML mechanism by which a faster representation is possible [2]. In [3] the authors have studied how the meta data can be used such that higher performance can be achieved and has also shown optimization using meta data concept gives better performance.

Time oriented retrieval plays an important while retrieving data from data warehouse and while achieving speed it's also important to achieve accuracy which is a challenging task. To achieve and accelerate data warehouse processes the author [4] has suggested methods to increase the speed during join and aggregation. A method called change data capture which is used to increase the performance in Extraction has been discussed.

A survey of existing modelling techniques, their characteristics notation and activities has been analyzed in [5]. This study states that there is a clear classification of ETL process modelling approaches but there is no clear and enough research in this area.

Logical optimization of ETL processes based on state-space search problem has been discussed in [6]. Authors have also indicated heuristic algorithms towards minimization of the execution cost of an ETL workflow.

Real-time ETL is method which allocates the work and takes decisions based on the amount of data it receives and partitioning of it to achieve better response time using parallel architectures has been proposed in [7].

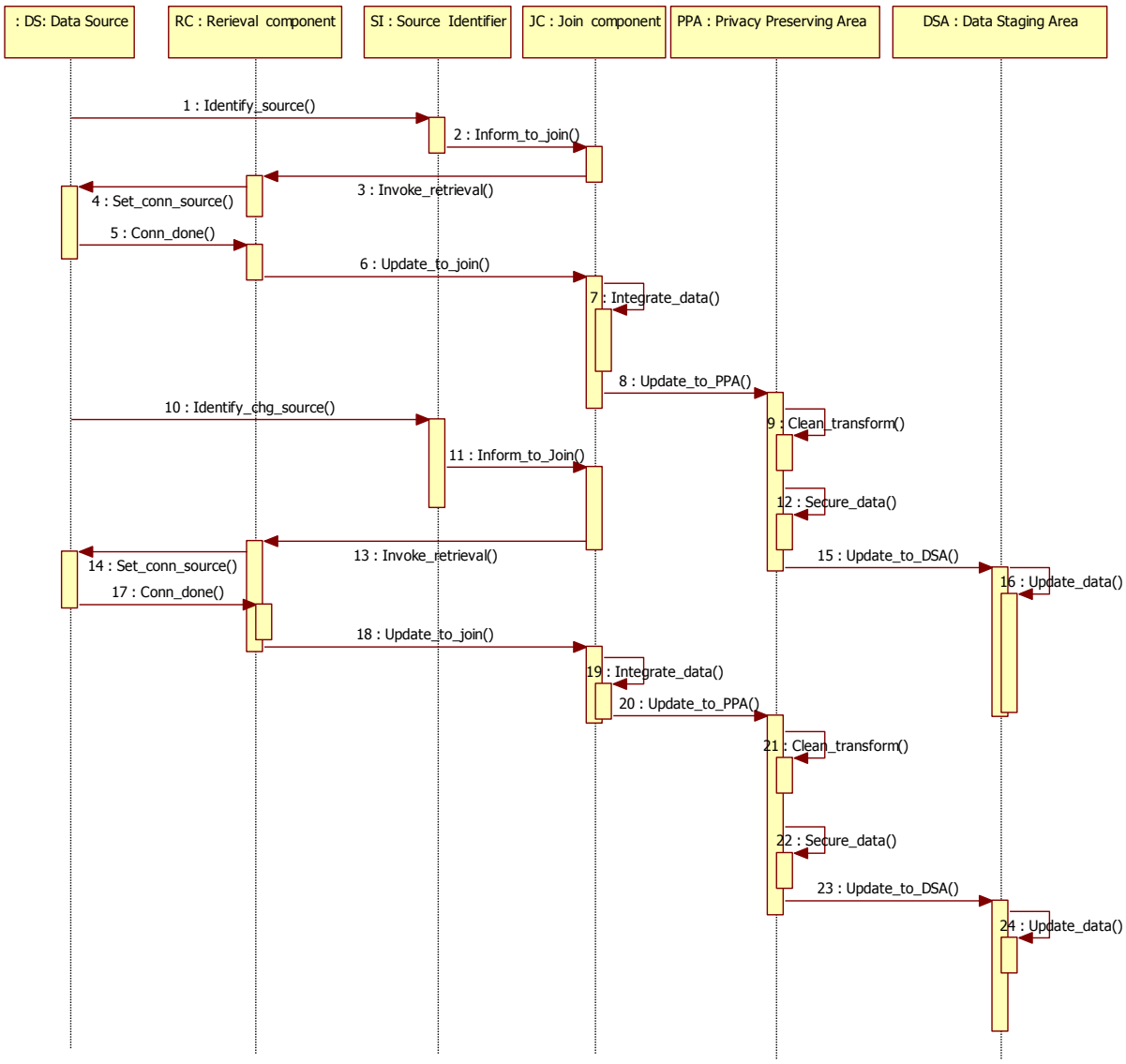

Fig. 4 Sequence Diagram for ETL embedding Privacy Preservation

In [8] an object-oriented approach has been used to represent ETL process by means of Unified Modelling Language (UML) and gives a brief idea regarding the class diagrams, use-case and sequence diagram that is required for ETL representation. Author in [9] have extended the earlier representation of [8] and has given a new dimension in representing secure data extraction through UML, but in none of the papers there is no representation of ETL embedding privacy. In this paper we are representing ETL in Privacy Preserving Data Mining using UML.

3. ARCHITECTURE OF DATA EXTRACTION EMBEDDING PRIVACY USING UML

Deployment diagram for architecture of data extraction embedding privacy is shown in figure 1. The lover level deployment consists of different source location where in the data is distributed, the data may be present as a flat file or DBMS. Each of this is represented as a separate component since each of these contains data in different format. The next level contains server component which retrieves information from different Data sources and also contains additional components such as Retrieval component, source identifier and Join components which has not been indicated in the architecture diagram. The retrieval component does the actual retrieval of data from source to destination. Source identifier identifies different sources that is required by the DSA and also informs join component if there is any change in the content of the source. Join component does an integration of different sources into the appropriate format suitable for Privacy Preserving Area (PPA).

PPA contains privacy algorithm component which converts specific attributes to privacy representation. It also does cleaning and transformation which is a major task. The order of processing would be cleaning and transformation first followed by privacy representation. The resultant is passed on to the Data Staging Area(DSA) which contains secure data. DSA does the major task of loading data on to logical model of Data Warehouse which is used by data mining algorithm ensuring privacy.

4. UML MODELING FOR DATA EXTRACTION EMBEDDING PRIVACY

The class diagram for data extraction embedding privacy is shown in figure 2. The main components of this diagram are Data Source, retrieval component, source identifier, join, PPA and DSA. The description of these components is as follows

4.1 Data source

Contains functionalities of setting server-id, setting source-type and setting source client for retrieval.

4.2 Source identifier

Identifies new sources by sid and informs to the join. It also identifies if there is a change in the source content and those sid will be informed to the join.

4.3 Retrieval component

Task of retrieval component is to connect data source and update the contents to the join.

4.4 Join class

Does the job of integrating data that is received from multiple clients in to a single logical schema representation.

4.5 PPA

is the most important and major component of this architecture, its task include cleaning and transformation of data followed by privacy preserving of some of the attributes. Privacy preservation may be done by means of data distortion or data manipulation which depends on the content of data base.

Use case diagram and sequence diagram is given in figure 3 and 4 respectively. UML modelling has been implemented using star UML.

5. ALGORITHM FOR DATA EXTRACTION EMBEDDING PRIVACY

The main procedure for data extraction embedding privacy is as follows.

// identifying the sources and creating the source list.
This is done by the methods of Source Identifier class
1. Identify the list of clients connected to the server
2. Set the properties for the source
3. If it is a new source add to the data source list
//source identifier maintains a copy of the new source list which is used later on to differentiate between the new source and the updated source
4. Source identifier informs join component
5.Join component establishes the connection and extracts data by means of retrieval component
6.Retrieved data is Loaded on to the join component which does the merger of multiple attributes present in different locations by means of a equi join
7. Join components informs PPA.
8. PPA does the transformation of the loaded data
// it also removes Identifiers and retains quasi identifiers & sensitive data
9. The transformed data is privacy preserved
10. Privacy preserved data is loaded on to the data staging area which is used for mining
11. Source identifier also identifies the changes in data source, if there is a change then the steps are repeated from step 4.
// Modification / updating of Data Staging Area (DSA).

6. EXPERIMENTAL RESULTS

The method of extraction, transform and load embedding privacy preservation was done on medical data set. The medical data contains sensitive information which may be revealed during mining. After merging multiple tables and transformation the resultant table is shown in Table 1. Identifiers patient_name & patient_address are removed before loaded on to the Data Staging Area. The resultant table contains the following fields Race, DOB, Sex, Zip , Marital Status and Health problem.. This information is passed on to secure data function which does k-anonymity. Anonymization technique [10, 11, 12] has been considered for privacy preservation. The experiment was conducted using Matlab 7. K=2 and Maxsup=1 has been considered for the resultant table is shown in Table 2.

Table 1. Resultant Data after Transformation

Race	DateOf Birth	Sex	ZIP	Marital Status	Health Problem
asian	9/27/1964	female	94139	divorced	Hypertension
asian	9/30/1964	female	94139	divorced	Obesity
asian	4/18/1964	male	94139	married	chest pain
asian	4/15/1964	male	94139	married	Obesity
black	3/13/1963	male	94138	married	Hypertension
black	3/18/1963	male	94138	married	shortness of breath
black	9/13/1964	female	94141	married	shortness of breath
black	9/7/1964	female	94141	married	Obesity
white	5/14/1961	male	94138	single	chest pain
white	5/8/1961	male	94138	single	obesity
white	9/15/1961	female	94142	widow	shortness of breath

Table 2. Resultant Data after Secure Data

Race	DateOf Birth	Sex	ZIP	Marital Status	Health Problem
asian	9/*/1964	female	94139	divorced	Hypertension
asian	9/*/1964	female	94139	divorced	Obesity
asian	4/*/1964	male	94139	Married	chest pain
asian	4/*/1964	male	94139	Married	Obesity
black	3/*/1963	male	94138	Married	Hypertension
black	3/*/1963	male	94138	Married	shortness of breath
black	9/*/1964	female	94141	Married	shortness of breath
black	9/*/1964	female	94141	Married	Obesity
white	5/*/1961	male	94138	Single	chest pain
white	5/*/1961	male	94138	Single	Obesity

7. CONCLUSION AND FUTURE WORK

This paper discusses embedding privacy preserving in Extraction Transformation Load using UML modelling. We have shown an UML modelling from initial stages of software which helps in better design representation. This representation gives an architectural change and functions that must be incorporated for embedding privacy in ETL and has been validated in the motivation example of medical data set. This architecture also gives us flexibility of adding various transformations which ultimately helps in preserving data.

Future work may include a combination embedding secure communication and privacy preserving. In this approach we have considered conceptual modelling of ETL future work may also concentrate on reliability, performance etc.

8. REFERENCES

[1] Fang Ying-lan and Han Bing, Design and Implementation of ETL Management Tool, In proc. of International Symposium on Knowledge Acquisition and Modeling, pp 446-449, 2009.

[2] Xudong Song, Xiaolan Yan and Liguo Yang,Design ETL Metamodel based on UML Profile, In proc. of Second International Symposium on Knowledge Acquisition and Modeling, pp 69-72,2009.

[3] Lunan Li, A Framework Study of ETL Processes Optimization Based on Metadata Repository, In proc of Computer Engineering and Technology (ICCET), pp 125-129, 2010.

[4] Darshan M. Tank, Amit Ganatra and Y P Kosta, Speeding ETL Processing in Data Warehouses Using High-Performance Joins For Changed Data Capture (CDC), In proc. of International Conference on Advances in Recent Technologies in Communication and Computing, 2010

[5] L. Muñoz, J. N. Mazón and J. Trujillo, ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study, IEEE Latin America Transactions, vol. 9, no. 3, june 2011.

[6] Alkis Simitsis, Panos Vassiliadis, and Timos Sellis, State-Space Optimization of ETL Workflows, IEEE Transactions On Knowledge and Data Engineering, vol. 17, no. 10, october 2005.

[7] Alkis Simitsis, Chetan Gupta, Song Wang and Umeshwar Daya, Partitioning Real-Time ETL Workflows, In Proceedings of ICDE Workshops, pp 159-162, 2010.

[8] M. Mrunalini, T.V. Suresh Kumar, D. Evangelin Geetha and K. Rajanikanth, Modelling of Data Extraction in ETL Processes Using UML 2.0, DESIDOC Bulletin of Information Technology, Vol. 2, pp. 3-9, 2006.

[9] M Mrunalini, T V Suresh Kumar and K Rajani Kanth, Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes, In proc. Of Third UKSim European Symposium on Computer Modeling and Simulation,2009.

[10] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In Proc. of the 10th Int'l Conference on Database Theory, January 2005.

[11] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymity. In Proc. of the 21st Int'l Conference on Data Engineering, April 2005.

[12] K. Lefevre, D.J. Dewitt, R.Ramakrishna, Incognito: Efficient fulldomain k-anonymity, In Proc. of the 24th ACM SIGMOD International Conference on Managemant of Data, pp. 4960, Baltimore, Maryland, USA, 2005

[13] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 188, 1998.