An Object Oriented Modeling and Implementation of Web Based ETL Process

Radha Krishna Author1[†] and Sreekanth Author2[††],

Information Systems & Management Consultants Pvt. Ltd.
Phase1, Tower 2, Electronic City SW Park,
Electronic City Post, Bangalore - 560 100.

Summary
The data warehousing environment includes components that are inherently technical in nature. These cleansing components function to extract, clean, model, transform, transfer and load data from multiple operational systems into a single coherent data model hosted within the data warehouse. The analytical environment is the domain of the business users who use application to query, report, analyze and act upon data in the data warehouse. The conventional process of developing custom code or scripts for this is always a costly, error prone and time-consuming. In this paper, we propose a web based frame work model for representing extraction of data from one or more data sources and use transformation business logic, load the data within the data warehouse. All the above mentioned have been modeled using UML because the structural and dynamic properties of an information system at the conceptual level are more natural than other classical approaches. New feature of entire loading process of data movement between source and target system is also made visible to the users. In addition a reporting capability to log all successful transformations is provided.

Keywords :
data warehouse, transformation, loading, log, extraction,UML

1. Introduction

In the world of distributed, heterogeneous database systems there is always a need for data migration process where there is a need to consolidate data from one or more data sources and move from one system to another. This process involves extracting the data from known sources which are of business interest. Once the decision is made to perform data migration and before migration can begin the following analysis must be performed:

- Analyze structure of data in the legacy system
- Analyze and define target structure of data in the new system
- Perform field mapping between source and target structure with data cleansing
- Define the migration process

To analyze and define source and target structures, analysis must be performed on the existing system as well as the new system to understand how it works, who uses it, and what they use it for. A good starting point for gathering this information is in the existing documentation for each system. This documentation could take the form of the original specifications for the application, as well as the systems design and documentation produced once the application was completed. Often this information will be missing or incomplete with legacy applications, because there may be some time between when the application was first developed and now.

Consider a banking scenario, where there may exist a tons of volume of complex data available at different sources in different formats distributed across the network which are of business interest hence there is always a need to perform migration of data by extract ,transform and load process to get data in unique and generalized format. The data warehousing is a challenging area where the data is critical for business decision. The ETL process is the key component area which needs to be addressed for correct business decision results and to improve data quality.

Much of research is not yet been done for ETL phase in web based scenario Hence there is a need for modeling the ETL process in web based distributed environment to provide effective business decision results for organizations.

2. Related Works

The design, development, and deployment of ETL processes currently being performed are adhoc and inhouse in nature and need design, modeling and methodological foundations.

Electronic data available on the web is exploding at an ever increasing pace. Most of this data is unstructured and may also available in different format with redundant data hidden thus increasing complexity to analyze on the data volume available and hence a time consuming and

Manuscript received February 5, 2010
Manuscript revised February 20, 2010

challenging task, thus making search difficult and hence restricts traditional database querying.
A conceptual model based on ontology to extract and structure the data automatically is given by Embley[1].
Further the conceptual and logical modeling of ETL process has been discussed by vassilidis. [2][3]
Tomasic discussed the distributed mediator architecture model and modeling of data source connections, the interface to underlying data sources and query processing semantics.[5]
In[4] Skiadopoulos, Spiros discussed on modelling formal logical model for ETL processes and how this model is reduced to an architectural graph.
In [8], authors study the lifecycle of a DW and propose a method for the design, development and deployment of a DW.
In this paper we model the framework for entire ETL process using UML because the structural and dynamic properties of an information system at the conceptual level are more natural than the naive approaches such as the ER-model. A software implementation for the same has been provided.

3. Proposed Solution

About 80% of the existing solutions currently available in the market are thick clients, maintenance of these involves typical hardware, storage, security requirements and also a need of backup and storage mechanisms.
One of the important consideration particularly for database or data warehousing projects is the performance requirements. For example, no user will accept the software if it does not return results needed in few seconds.

3.1 Architecture For ETL Process

In our proposed approach the whole system is decomposed into three components, viz. extraction, transformation,
loading are treated as ETL software as shown in Figure. 1.
There will be only one server system with any number of clients present in the network or one or more servers if needed. Both the client and the server can communicate through the network. The client can only view the information as an end-user.
Many subsystems run on more than one machine depending on access to internet. Hence we need to carefully examine the allocation of subsystems to machine and the design infrastructure for supporting communication between subsystems.
These machine are modeled as nodes in the deployment diagram. Extraction and Loading subsystem run on the web Server while Transformation subsystem runs on the On board machine.

In Figure.1 shown below, ETL web interface application API library files and other necessary software are available forming a system at client tier while the application and database server together are available as a different subsystem at server tier level.

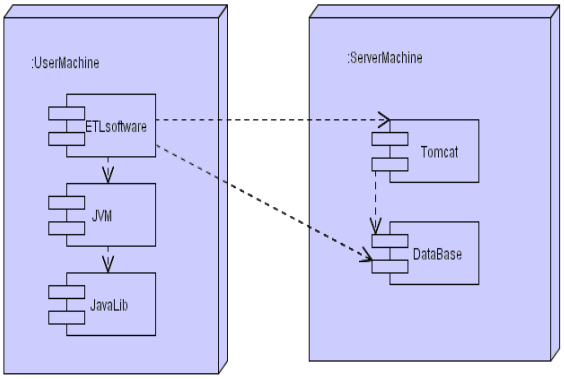

Fig.1 Deployment diagram for modeling web based ETL process

3.2 UML Modeling For Web-Based ETL Process

In Figure.2, The class diagram for ETL mechanism is given. The main elements in this diagram are the class OLTP, OLAP, Transformation, Loading etc. The data is extracted as shown in class diagram from various data sources like flat files, databases as needed after data cleansing and then it undergoes transformation based on business transformation logic by mapping fields like ‘m’ for male, ‘f’ for female represented as a transformation class and loaded finally to OLAP. A unique transaction identifier is used to identify a particular transaction performed as represented with trans class.

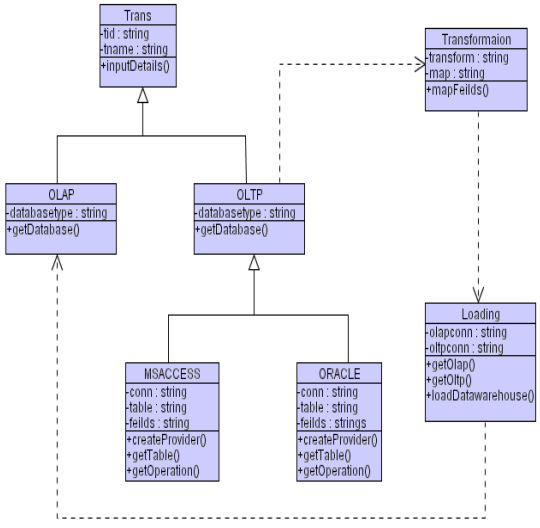

Fig. 2 Sample Class diagram for ETL process

The sequence diagram for actors user and administrator are given in Figures. 3, 4 respectively

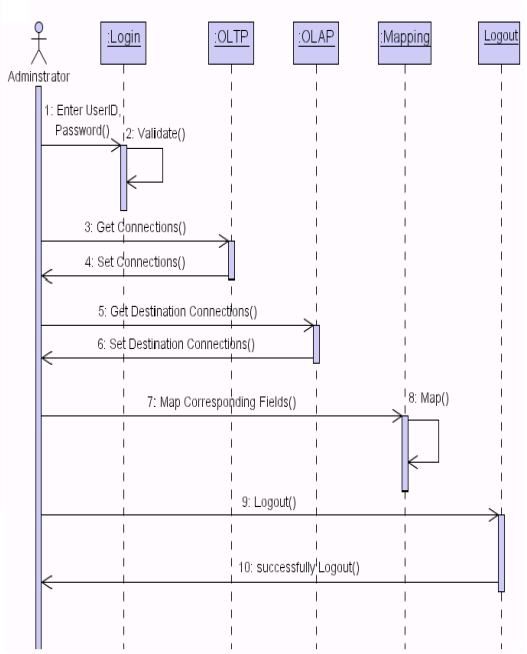

Fig. 3 Sequence diagram for administrator

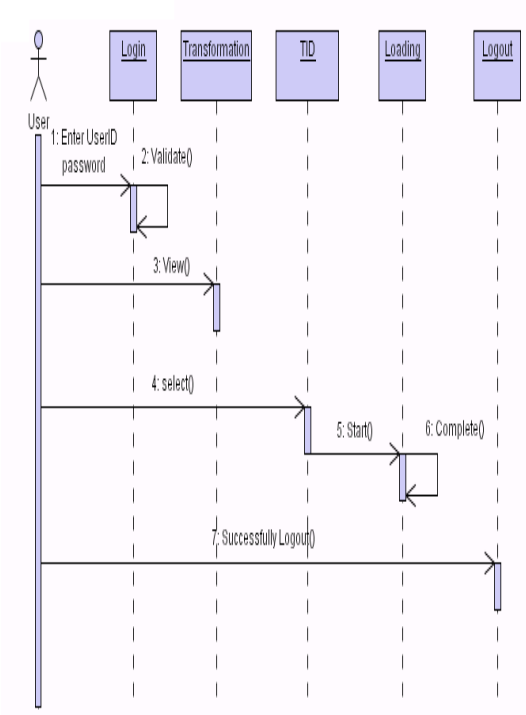

Fig 4. Sequence diagram for user

4. Case Study and Results

We have implemented a system that handles the ETL process in web based scenario. The software is implemented using an object oriented approach based on the architecture and the analysis modeling provided in above sections 3.1 and 3.2 respectively. This system uses J2EE technologies and metadata framework has been developed that provides information for identifying the user transactions being performed and handling ETL related metadata i.e. describes source data, target data and transformation carried out is stored in the repository. User can generate all types of transformations on all types of data bases. A new feature is provided in our system that allows the entire loading process of data being migrated and loaded to target environment along with validation of source and target field mapping to be visible to authenticated users as shown in Figure 8. The implementation process involves.

- Login to system and generating unique Transaction id
- Selecting OLAP or OLTP source system
- Get the tables, generate tables, map fields
- Load data in to the database

Consider a Tire Corporation with branches across Asia, America and Europe as shown below in figure 5. First tire was produced in 1983 and it wasn't until two years later that they began exporting their tires outside Japan and entered the US market only in 1990.

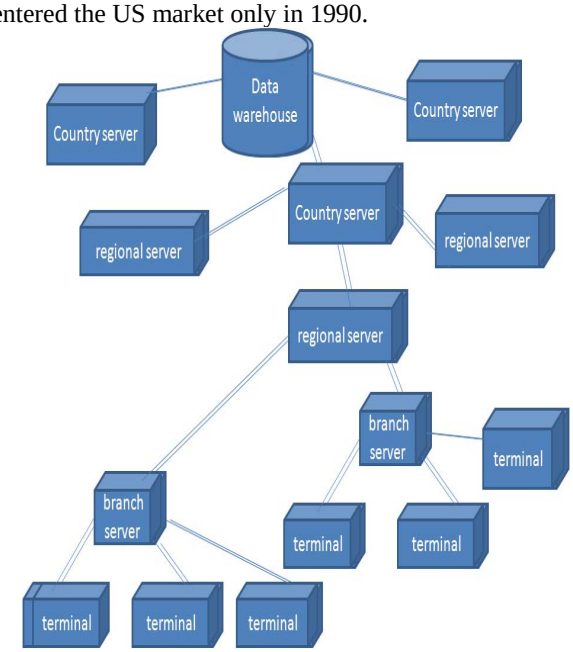

Fig. 5 Deployment diagram for example scenario

As the organization grew, so did the need for accurate and immediate information. The company lacked an effective way to consolidate, manage and distribute business data. Data resided in many different sources and in different formats, limiting its ability to analyze and deliver it within a single platform

The fundamental problem is that their Sales data was in the Operation System, Objective was in a flat file and Market data was in a SQL database. It was not only time consuming to create a report but also not flexible to do further analysis

Traditionally, Tire Corporation could not track or forecast actual sales. Therefore, it was essential to access information in the data mart with ease. They needed to get a deeper insight into their key sales performance to better forecast the market.

After studying the data from different sources, we can come up with a model of how the information fit together. To enable this, there is a need to understand the existing information and find the correlated pieces, then create a common format for the data mart in a database.

In order to achieve this we can use our proposed system implemented, available at different client stations and convert the data as needed to a single format as need and stored at some warehouse.

This can be achieved with minimal over head. Reliability is provided in transformation and loading process as entire loading process with data being migrated and other results are made visible to the user with the feature provide in our system. It gives an insight clear picture of what data is being transformed and migrated to which target system etc.

The need for heavy hardware configuration is eliminated at client site, using our system with minimal effort.

The figures.6,7,8,9 show sample ETL process performed along with the additional feature of loading in web based scenario.

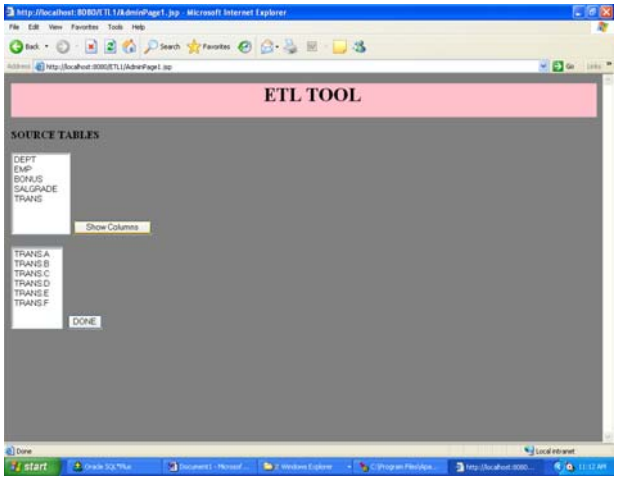

Fig. 6 Sample screen showing list of tables of source database, columns of TRANS table.

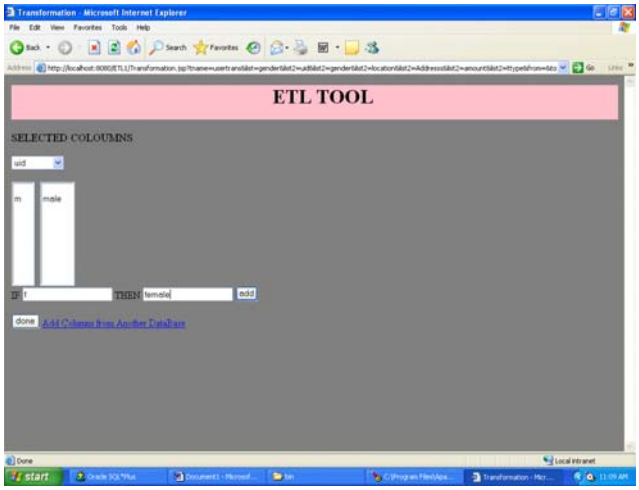

Fig. 7 Sample Transformation using web interface component

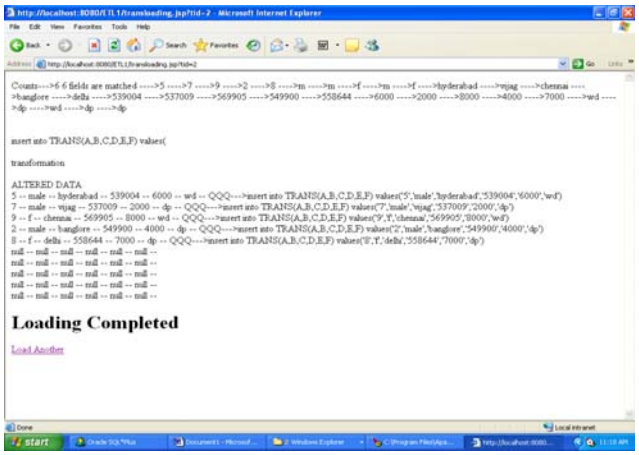

Figure 8 : Sample Loading process showing the phase of data migration to the user

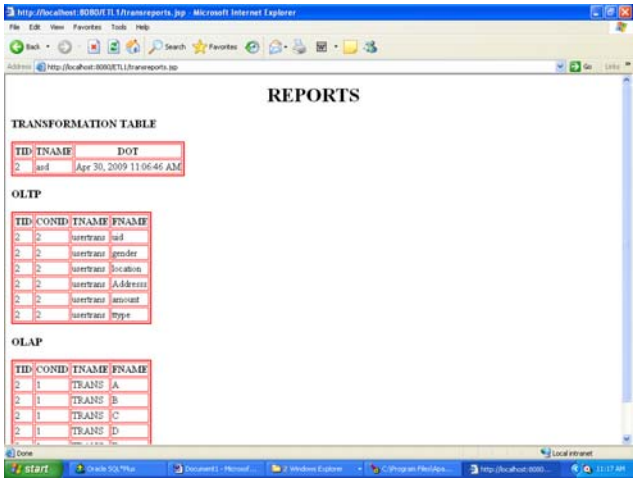

Fig. 9 Reports of successful transformation

5. Conclusion and Future Scope of Work

It is observed that the scenario of data extraction transformation and loading in the web based gives more flexibility, if represented using UML. The heterogeneous and distributed database systems that can be used to support trading corporations, banks, financial and human resource management systems of an organization at various levels when modeled with this architecture gives more flexibility and reduces maintenance cost, hardware and software infrastructure needed and storage requirements over various client stations and also loading data is made reliable with the new feature provided with implementation of visible migration of data loading process. After successful loading of data the user can view the transformation reports describing source data considered and target database to which the data is loaded. Future directions may include analyzing multimedia information sources, automating mechanisms for ETL process.

Acknowledgement

The authors would like to thank the referees for their valuable suggestions and also thank Mr.Sravankiran for his help in carrying out the work

Reference :

[1] Embley,D.W; Campbell,D.M;Jiang,Y.S; Liddle,S.W; Wionsdale,D.; Ng, Y.K & Smith,R.D. Conceptual model based data extraction from multiple record web pages,Elsevier,22 june 1999

[2] Vassiliadis, Panos; Simitsis, Akis; Georgantas, P.& Terrovitis, M. A framework for the design of ETL Scenarios . In Proceedings Of 15th International Conference On Advanced Information Systems Engineering, Velden, Austria, 16 June 2003.

[3] Vassiliadis, Panos ; Simitsis, Akis. On the logical modelling of ETL processes. In Proceedings of International Conference on Advanced Information sytems Engineering, 2002 pp.782-86

[4] Skiadopoulos, Spiros ; Vassiliadis,Panos. Modelling ETL activities as graphs. In Proceedings of DMDW '2002. Toronto, Canada, 2002. pp. 52-61

[5] Tomasic, Anthony, Raschid, Louiqa & valdurez : Scaling heterogeneous data sources with DISCO in IEEE Transactions of Knowledge and Data Engineering 1998. 10(5),808-23

[6] P. Vassiliadis et al. Arktos: Towards the modeling, design, control and execution of ETL processes. Information Systems, 26(8), pp. 537-561, December 2001, Elsevier Science Ltd.

[7] Sunita Sarawagi (editor). Special Issue on Data Cleaning. IEEE Data Engineering Bulletin, Vol. 23, No. 4, December 2000.

[8] R. Kimball, L. Reeves, M. Ross, and W. Thornth -waite. The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, 1998

[9] J. Trujillo, M. Palomar, J. Gómez, and I. Song. Designing Data Warehouses with OO Conceptual Models. IEEE Computer, special issue on Data Warehouses, 34(12):66–75, December 2001.

[10] Evolutionary Technologies International, ETI Solution. Hardware and Software Requirements, http://www.eti.com, 2004, access date:

[11] R. Fagin, P. G. Kolaitis, and L. Popa. Data exchange: getting to the core. ACM Trans. Database Syst., 30(1):174{210, 2005.

[12] S. Rizzi, J. Lichtenberger, and J.Trujillo. Research in data warehouse modeling and design: dead or alive? In DOLAP '06: Proceedings of the 9th ACM international Workshop on Data warehousing and OLAP, pages 3-10, 2006

[13] P. Vassiliadis, A. Karagiannis, V. Tziovara, P. Vassiliadis, and A. Simitsis. Towards a Bench mark for ETL Workflows. In 5th International Workshop on Quality in Databases (QDB) at VLDB, 2007

[14] Extraction Transformation loading-A road to data warehouse. 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries. September 26-28, 2008.

Radha Krishna. received his B.Tech from Bangalore University and M.Tech in Computer science and Engineering from Osmania university. He was associated with NCR Corporation limited, Hyderabad. He is a certified SQLAssociate from Cambridge intercontinental university. He has a teaching experience of 6 yrs. His areas of interest are Databases, CompilerDesign, Algorithms, Cloud computing, 4G networks, Cognitive networks.

Sreekanth. received his B.Tech and M.Tech in software Engineering from School of Information Technology, Hyderabad. Currently he is working as a Software Engineer in CGI Information Systems and having over three years of experience in data warehouse area. and oracle certified developer .His Areas of interest are Business intelligence, Databases and warehouse technologies.