DESIDOC Bulletin of Information Technology, Vol. 26, No. 5, September 2006, pp. 3-9

Modelling of Data Extraction in ETL Processes Using UML 2.0

M. Mrunalini, T.V. Suresh Kumar, D. Evangelin Geetha and K. Rajanikanth

M.S. Ramaiah Institute of Technology, Bangalore-560 054
E-mail: mruna_nalini@yahoo.co.in

ABSTRACT

The topic of data warehousing encompasses architectures, algorithms, and tools for bringing together the selected data from multiple databases or other information sources into a single repository called data warehouse. Extraction-transformation-loading tools are pieces of software responsible for the extraction of data from several sources, their cleaning, customisation, and insertion into a data warehouse. The paper proposes an object-oriented approach to accomplish the data extraction modelling of extraction-transformation-loading process. The data extraction scenario consists of data staging area, heterogeneous information sources, wrappers, monitors, integrator, and source identifier. All the afore mentioned entities have been modelled using Unified Modelling Language 2.0. Banking system has been used as an application to illustrate the modelling.

Keywords: ETL, data extraction, distributed systems, heterogeneous data, UML, banking system, extraction-transformation-loading

1. INTRODUCTION

Providing integrated access to multiple, distributed, heterogeneous databases and other information sources has become one of the leading issues in database research and industry[1]. Most approaches to the data integration problem are based on the following two steps:

- Accept a query, determine the appropriate set of information sources to answer the query, and generate the appropriate sub-queries or commands for each information source
- Obtain results from the information sources, perform appropriate translation, filtering, and merging of the information, and return the final answer to the user or application.

The process is referred as a lazy or on-demand approach to data integration, since information is extracted from the sources only when queries are posed. The process may also be referred to as a mediated approach, since the module that decomposes queries and combines results is often referred to as a mediator[2]. The natural alternative to a lazy approach is an eager or an in advance approach to data integration. In an eager approach

Received on 2 May 2006, revised on 22 July 2006

information from each source, which may be of interest, is extracted in advance, translated and filtered as appropriate, merged with relevant information from other sources, and is stored in a (logically) centralised repository. When a query is posed, it is evaluated directly at the repository without accessing the original information sources. The lazy and warehousing approaches are viable solutions to the data integration problem, and each is appropriate for certain scenarios.

Data warehouses (DWs) are complex systems. The main goal of the DWs is to facilitate the decision making process of knowledge workers. Extraction-transformation-loading (ETL) processes are responsible for the extraction of data from heterogeneous operational data sources, their transformation (conversion, cleaning, normalisation, etc.), and their loading into DWs. ETL processes are key component of DWs, because incorrect or misleading data will produce wrong business decisions. Therefore, a correct design of these processes at early stages of a DW project is absolutely necessary to improve data quality. However, not much research has dealt with the modelling of ETL processes.

In this paper, approach based on the Unified Modelling Language (UML 2.0), which allows to accomplish the modelling of ETL and data extraction processes, has been presented. UML 2.0 was used for designing DWs, because structural and dynamic properties of an information system at the conceptual level are more natural than classic approaches such as the entity-relationship model.

2. RELATED WORK

The design, development, and deployment of ETL processes, which are currently being performed in an ad-hoc and in-house fashion, need modelling, design, and methodological foundations.

The concept of DWs is motivated. The general architecture and a number of technical issues arising from the architecture have been discussed by Jennifer Widom[3]. A conceptual modelling of ETL processes in DWs using UML 2.0 has been discussed by Vassiliadis[4], *et al*. The necessary mechanisms for an easy and quick specification of the common operations defined in the ETL processes have also been provided.

Electronica data available on the web is exploding at an ever-increasing pace. Much of this data is unstructured, which makes searching difficult and traditional database querying impossible. A conceptual modelling approach based on ontology to extract and structure data automatically has been given by Embley[5], *et al*.

The conceptual and logical modelling of ETL process have been discussed by Vassiliadis[6,7], *et al*. A framework for the conceptual, the logical and the physical design of ETL processes has been discussed by Simitsis[8].

Distributed mediator architecture, the data model and modelling of data source connections, the interface to underlying data sources and the query rewriting process, and query processing semantics have been discussed by Tomasic[9], *et al*.

A formal logical model for ETL processes is proposed and how this model is reduced to an architectural graph has been discussed by Vassiliadis[10], *et al*. This paper used a general graphical structure to represent ETL activities.

In all the above mentioned studies, the UML 2.0 for modelling ETL and data extraction has not been used.

3. ARCHITECTURE OF DATA EXTRACTION USING UML 2.0

In Fig. 1, the bottom layer shows the information sources such as DBMS, RDBMS, ORDBMS, flat files, news wires, HTML pages, and so on. The information source may be the operational database or an external source. In general, it is called information or data sources (terms information and data have been used interchangeably). Each of these

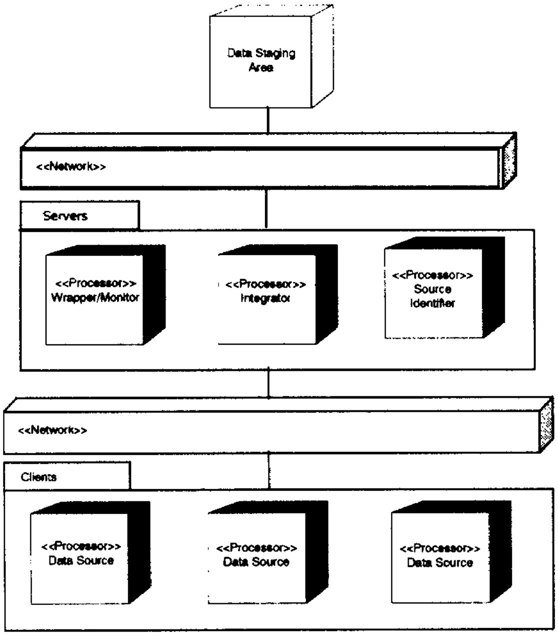
Data Staging Area
<<Network>>
Servers
<<Processor>> Wrapper/Monitor
<<Processor>> Integrator
<<Processor>> Source Identifier
<<Network>>
Clients
<<Processor>> Data Source
<<Processor>> Data Source
<<Processor>> Data Source

Figure 1. Deployment diagram for architecture of data extraction.

information sources is connected to respective wrapper/monitor. The wrapper component of this module is responsible for connecting to the source and extracting data from that source. The monitor is responsible for automatically detecting changes of interest in the source data and reporting them to the integrator.

When a new information source is attached to the ETL system, or when relevant information at a source changes, the new or modified data is propagated to the integrator by source identifier. The integrator is responsible for installing the information in the data staging area (DSA). To properly integrate new information into the DSA, it may be necessary for the integrator to obtain further information from the same or different information sources. The DSA is an intermediate storage in the ETL process.

The extracted data from the distributed heterogeneous data sources is stored in DSA and is used in the next level of ETL process, i.e., cleaning and transformation. The cleaned and transformed data is then loaded into the data warehouse/data marts.

4. UML MODELLING FOR DATA EXTRACTION AND ETL

The class diagram for data extraction is given in Fig. 2. The main elements in this diagram are the class wrapper, monitor integrator, source identifier, and the DSA. Functionalities and information of these classes have been described in the diagram. For example, the data source class contains source type, source ID and client name. The methods in this class are used to set the attributes of this class.

The source identifier class identifies any new additions in the network, assigns an instance of wrapper/monitor component to that source and informs it to the integrator. The wrapper class sets connection with the data source and extracts data from the source, and gives it to the integrator. The integrator installs this data in the DSA. If there is any change in the existing data, that will be identified by the monitor class. The information is then passed on to the integrator.

Use case and sequence diagrams for data extraction and ETL are given in Figs. 3, 4, 5, and 6, respectively.

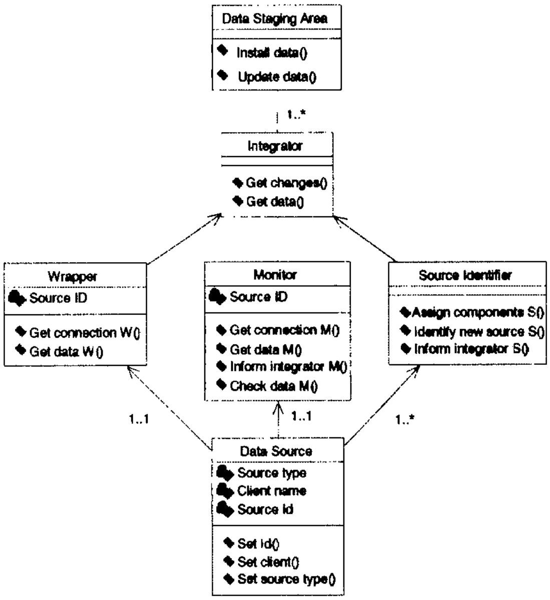

Figure 2. Class diagram for data extraction

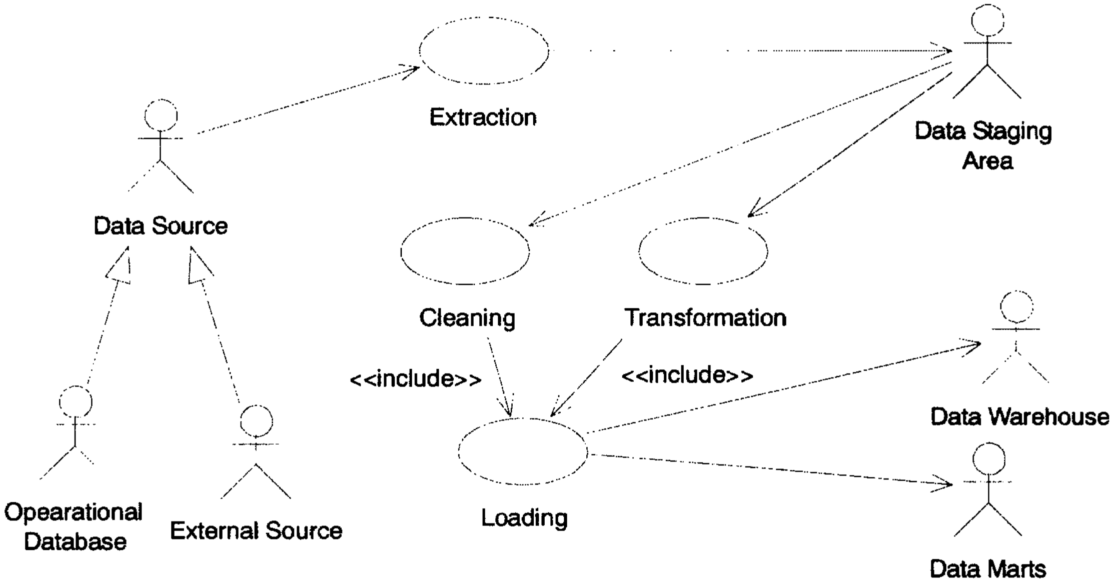

Figure 3. Use case diagram for data extraction

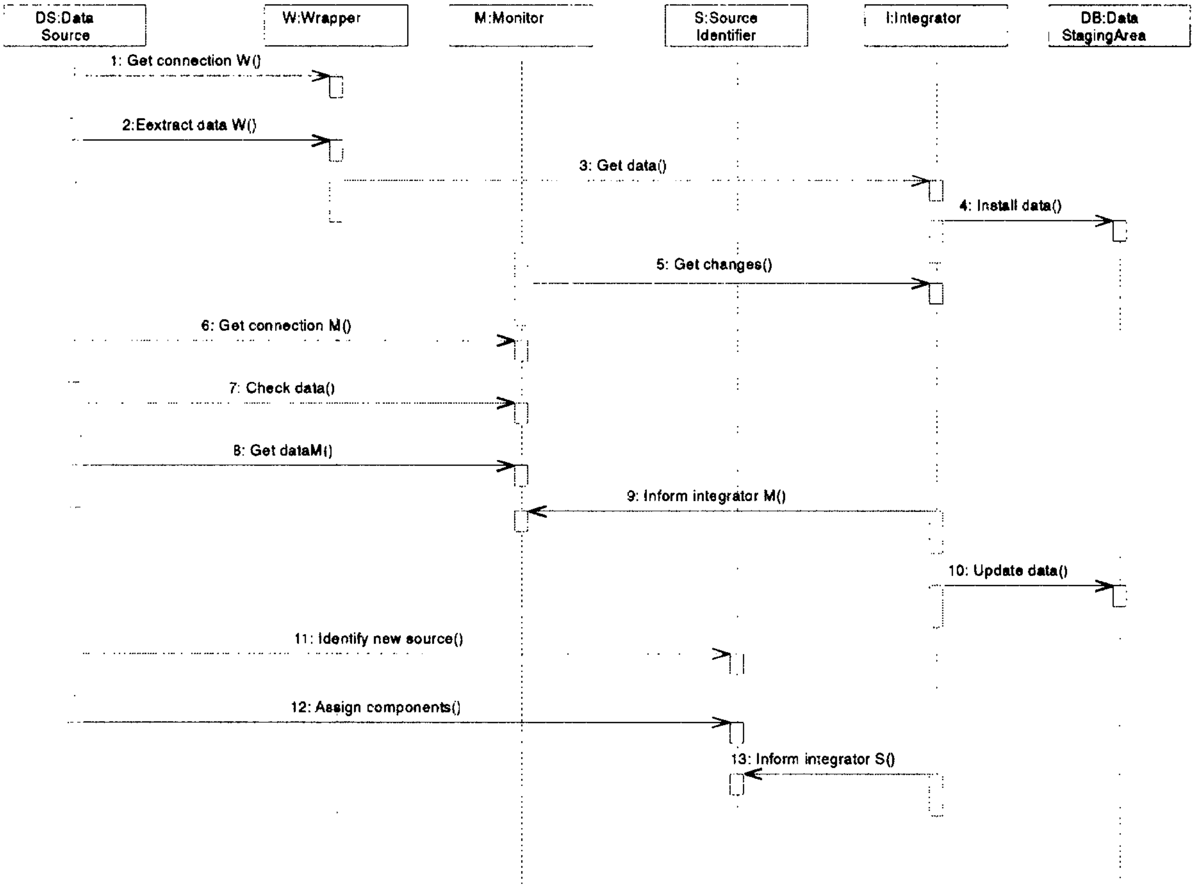

Figure 4. Sequence diagram for data extraction

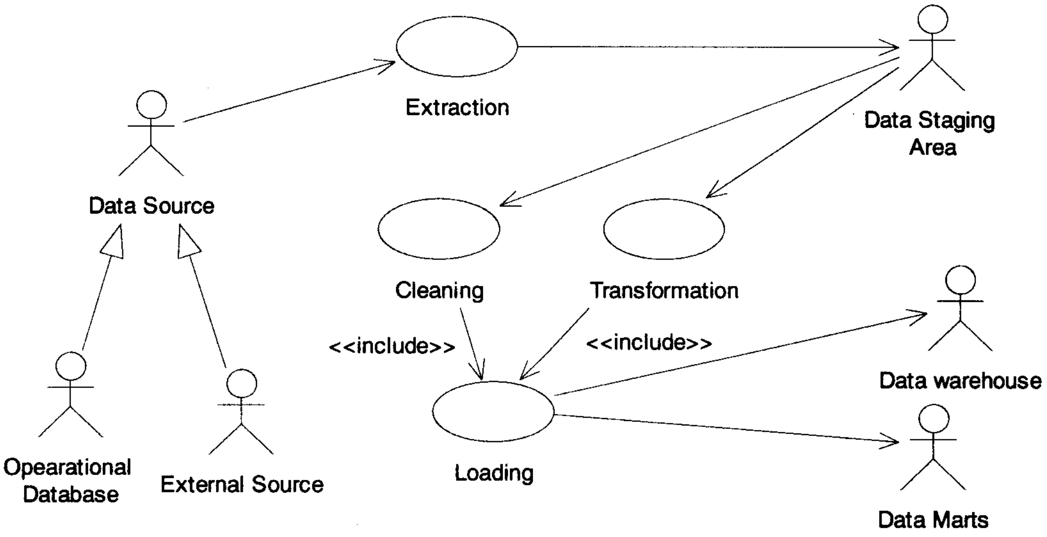

Figure 5. Use case diagram for ETL process

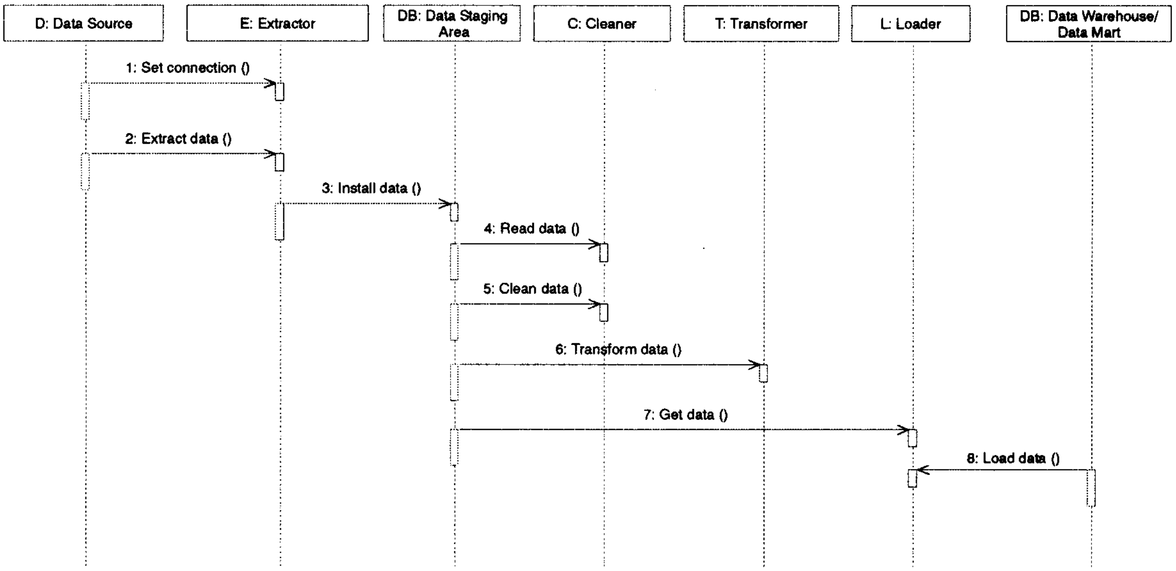

Figure 6. Sequence diagram for ETL process

5. CASE STUDY FOR BANKING SYSTEM

The banking system has heterogeneous distributed database system. It has multiple servers, and multiple client architecture. Clients are heterogeneous and distributed. Each client may have different data models, which cause heterogeneity. The information sources may be like DBMS, RDBMS, ORDBMS, flat files, magnetic tapes, web pages, etc. From these heterogeneous data sources, data/information have to be extracted and stored in a common storage area.

5.1 Architecture of Banking System

The architecture of a multinational banking system may be like the head office (main server) located in one country with its various branches in different countries. In each country, there are different regions and branches. Among these, one branch acts as head office for that country. Every branch has a client/ server architecture (Fig. 7).

The country server will have a database called DSA, while the regional server will have the integrator, source identifier, and wrapper/monitor components. The branch server is the data source. Whenever a new data source is added in the network, the source identifier identifies that new source and assigns an instance of wrapper/monitor component to that source and informs to the integrator. The wrapper component extracts data from that source and gives it to the integrator. Any changes in the existing data will be identified by the monitor. Monitor then informs it to the integrator. Integrator is responsible to insert this data in the DSA. The country server has the other components of the ETL tool, i.e., cleaner, transformer, and loader. These components access the data from DSA, clean it, transform and load it in the data warehouse or data mart.

6. CONCLUSION & FUTURE WORK

It is observed that the scenario of data extraction gives more flexibility, if represented using UML 2.0. The heterogeneous database system used to support a banking application model found that this architecture gives flexibility of adding various types of information sources, which ultimately helps in storing the data into the DWs. At present, the authors are trying to develop an object-oriented approach to support/popularise DW where ETL is the key process. Future directions may include analysing issues of multimedia information sources and performance modelling using UML 2.0.

REFERENCES

1. Ram, Sudha (Ed). Special Issue on Heterogeneous Distributed Database Systems. *IEEE Computer*, 1991, **24**(12).

2. Wiederhold, G. Mediators in the architecture of future information systems. *IEEE Computer*, 1992, **25**(3), 38-49.

3. Widom, Jennifer. Research problems in data warehousing. *In* Proceedings of 4th International Conference on Information and Knowledge Management, November 1995.

4. Vassiliadis, Panos; Simitsis, Alkis; Georgantas, P. & Terrovitis, M. A framework for the design of ETL Scenarios. *In* Proceedings of the 15th International Conference on Advanced Information Systems Engineering, Velden, Austria, 16 June 2003.

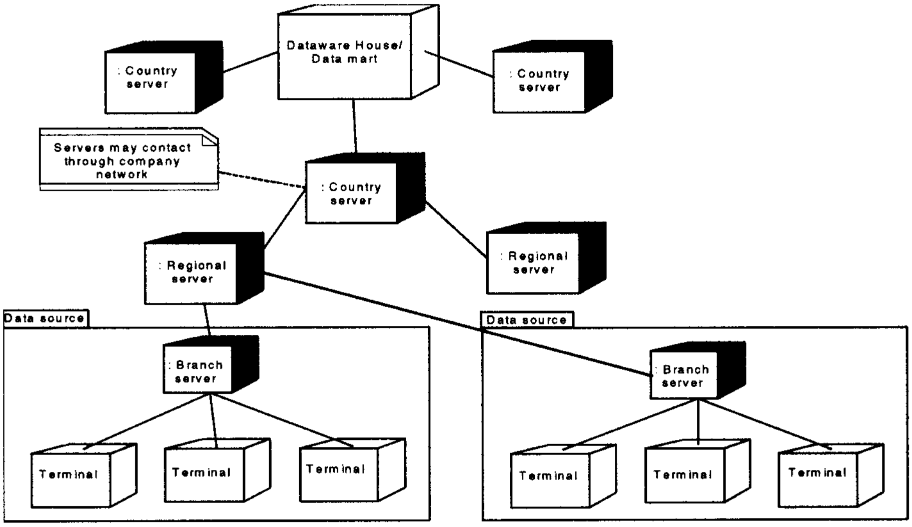

Figure 7. Deployment diagram for banking application

5. Embley, D.W.; Campbell, D.M.; Jiang, Y.S.; Liddle, S.W.; Wlonsdale, D.; Ng, Y.K. & Smith, R.D. Conceptual-model-based data extraction from multiple-record web pages. Elsevier, 22 June 1999.

6. Vassiliadis, Panos; Simitsis, Alkis & Skiadopoulos, Spiros. On the logical modelling of ETL processes. *In* Proceedings of International Conference on Advanced Information Systems Engineering, 2002. pp. 782-86

7. Vassiliadis, Panos; Simitsis, Alkis & Skiadopoulos, Spiros. Conceptual modelling for ETL processes. *In* Proceedings of DOLAP'02, 8 November 2002. McLean, Virginia, USA. ACM 1-58113-590-4/02/0011.

8. Simitsis, Alkis. Modelling and Managing ETL processes. Available at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-76/simitsis.pdf.

9. Tomasic, Anthony; Raschid, Louiqa & Valduriez, Patrick. Scaling access to heterogeneous data sources with DISCO. *In* IEEE Trans Knowledge Data Engg, 1998, **10**(5), 808-23.

10. Vassiliadis, Panos; Simitsis, Alkis; & Skiadopoulos, Spiros. Modelling ETL activities as graphs. *In* Proceedings of DMDW '2002, Toronto, Canada, 2002. pp. 52-61.

Contributors

Ms M. Mrunalini received her MCA from S.V. University, Tirupati in 2000 and MPhil (Computer Science) from M.S. University, Tirunelveli, Tamil Nadu, in 2003. Her areas of interest are ETL tools, data warehouse, software performance engineering, and object technology. She is currently working as Lecturer at the M.S. Ramaiah Institute of Technology, Bangalore.

Shri T.V. Suresh Kumar received his PhD from the S.K. University, Anantapur, in 1992. His areas of interest are software engineering, software performance engineering, object technology, and distributed systems. He is currently working as Professor at the M.S. Ramaiah Institute of Technology, Bangalore.

Ms D. Evangelin Geetha received her MCA from the Madurai Kamaraj University in 1993, and MPhil (Computer Science) from the M.S University, Tirunelveli, Tamil Nadu, in 2003. Her areas of interest are software performance engineering, object technology, and distributed systems. She is currently working as Assistant Professor at the M.S. Ramaiah Institute of Technology, Bangalore.

Shri K. Rajanikanth received his ME (Automation) and PhD from the Indian Institute of Science, Bangalore. His areas of interest are software engineering, object technology, embedded systems. He is currently working as Professor and Principal at the M.S. Ramaiah Institute of Technology, Bangalore.