



Toward XML-Based Data Warehouse Architecture

Rami Rifaieh

R&D-TESSI INFORMATIQUE

72 bis Rue Bergson, 42000 Saint-Etienne, France

Email: rrifaieh@tess2i.fr

Nabila Aïcha Benkat

Department of Informatics, LIRIS-INSA de Lyon

69621 Villeurbanne, France

Email: nabila.benharkat@if.insa-lyon.fr

ABSTRACT

XML enables data to migrate from relational databases and other sources into future applications. It integrates structured and unstructured data to present new application and knowledge management opportunities. By another way, data warehousing is an essential element of decision support, which has increasingly become a focus of the database industry. Meta-data contains data dictionary and repository; it describes warehousing process, data storage and information delivery. This paper defines how XML affects and changes the concept of the data warehouse. It describes an XML based tool for a structured, reusable and more efficient data warehouse (DW). Otherwise, it shows the ability to specify the warehousing tools with rules defined in the meta-data dictionary.

1 INTRODUCTION

Data warehouse system is a collection of data used for decision-making. The success of data warehouses implementation for business intelligence activities implies an increasing demand for new concepts and solutions [2]. It includes growth across platforms, tools, and applications. In all cases, the integration of data from heterogeneous sources is the essential stage of warehousing process. These sources could arise from existing legacy systems, which continue to generate data from mainframes computers through client server architecture. Moreover, disparate Web applications such as data sharing, real time data interchange, e-business, B2B activities and systems as ERP provide information that may be used to populate the company data warehouse. Then, data warehouse should be improved to support new applications such as real-time data warehouse and data web-house.

Furthermore, meta-data is an essential information that defines the what, where, how, and why about the used data. It can range from the conceptual overview of the real world to detailed physical specifications for the particular database management system. Meta-data can be used by automated tools (e.g. indexing robots) to improve the data interpretation / exploitation. Specifically, meta-data of business warehouse contains management rules [2]. They define which elements are supposed to be decisional and how they are calculated.

XML enables data to migrate from relational databases and other sources into new applications. It includes the ability to exchange data between application programs and browsers, between application programs and other application programs etc.

The purpose of this paper is to show how XML and its standards support data warehousing process. Therefore, XML could be used in all the construction process, especially in legacy data extraction, input transaction capture, cleansing procedures, direct storage of XML, and front-end information delivery. As we also deal with the evolution of meta-data, the XSLT component associated with XML documents helps us to support progressive meta-data for management rules. XML schema can provide another interesting dimension to XML text by defining datatypes. Moreover, an automatic generation of XSLT transformation could be a solution for personalization process based on meta-data description.

We will focus on showing where and how the extensible markup language can reduce process complexity. We try to show how using XSLT makes it easier to acquire and transform data. We discuss also the

automatic generation of XSLT from progressive meta-data, and show the usefulness of XML schema in the loading process.

The paper is organized as follows. In the next section, we will glance over DW and its construction process. The section 3 will detail the advantages of using XML for DWs. The section 4 will describe the design of XETL (XML-based Extraction, Transformation and Load tool). We will conclude our work in the section 5 with the future perspectives.

2 THE DATA WAREHOUSE SYSTEM

Data warehousing constitutes the background to enable business intelligence solution, which lets organizations access, analyze, and share information internally with employees and externally with customers, suppliers, and partners.

According to standard data warehouse architecture, the data warehouses systems include:

- ETL or warehousing tools: a set of tools which are responsible of preparing the data constituting the Warehouse database;
- Restitution tools: the diverse tools, which help the analysts to make their business decisions, and;
- Meta-data: it brings together the data about all the components inside the DW.

2.1 ETL Tools

ETL (Extraction, Transformation, and Load) represents the warehousing tools, or population tools. Warehousing tools have challenge to provide maintenance capability, availability, task management, and evolution support. Data integration and reuse possibilities are wide open but not yet very well realized. Although, some tools provide reused functions, these solutions still limited. Indeed, existing functions do not allow users to utilize an existing transformation plan and specify it with parameters to create a new data warehouse.

Maintenance process is not better; for example, if a user wants to change a SKU (stock keeping unit) number definition from five digits to seven. How many programs need to be changed to affect this enhancement? For the most of existing tools, in order to enable this operation a query has to be formulated into the data dictionary. Then, the user has to update all the concerning programs. In the population process, our suggestion is to combine the ETL tools with XML, since our concern is first of all the diversity of data sources, data targets.

2.2 Restitution Process

Often the information in data warehouses is published to a company's intranet web site. HTML is actually used to build these sites. Thus, restitution tools should provide the ability to generate HTML pages from warehouse database.

Nevertheless, delivering operation results from restitution model to mobile portals is going to be a new feature for restitution tools. Wireless devices (Pocket and mobile phone) will be able to capture information and use it in a real time with geographically distant warehouse. Thus, new standard output format should be used such as WML (Wireless Markup Language).

3.3 How Meta-data Improve Data Transformation

For actual data warehouse architecture, few of tools communicate directly with meta-data repository. Meta-data are used to answer administrator queries. These queries offer information concerning the structure, models, and the warehousing process. Meanwhile, the set of meta-data is passive and query limited.

The existing tools do not deal with the evolution of management rules and reusability. Indeed, by one way, meta-data of business warehouse contains management rules. A passive use of meta-data implies two sorts of updates: meta-data repository and programs that handle data into warehouse. By another way, ERP systems [3] store data of different applications by a similar way on mainframe computers. In this case, data dictionary should provide more than just data attribute descriptions, range value, valid value, etc. It should supply the personalization process by applying the needed mapping to achieve the construction of different data warehouse. Thus, the same data element might be used by different entities of different applications to mean different things. Different data element names could also be used to represent the same things, potentially creating hundreds of instances of the same data all inconsistently named. In this case, data dictionary should provide more than just data attribute descriptions, range value, valid value, etc. It should supply the personalization process by applying the needed mapping to achieve the construction of different data warehouse. A generic plan is needed to establish mapping between source and target data. If meta-data contains all the generic plans for a target application, a personalization process can be used to extract data from source and apply the needed mapping.

The idea of making meta-data being *active* can improve warehouse systems. The solution can be performed with traditional transformation query with an automatic query generator or it can be a specific transformation language as used in XETL tool.

4 PRESENTATION OF XETL TOOL

In this section, we present XETL, an XML based Extraction, Transformation and Load tool (Fig.1). The basic idea of our tool consists of using the XML format to create new generation of data warehouse, where XML is used as a pivot format to perform the warehousing process. A full case study of using XETL with real scale commercial data is described in [12]. This study shows the effectiveness of using XETL as a warehousing tool to populate a relational database.

4.1 XETL and Existing Tools

XETL takes advantage of XML ability to realize interoperability, scalability, maintenance, efficiency, and data integration opportunity. We consider that data transformation process is generated directly from meta-data. Thus, the designer will not be called to verify the consistency of the process, since this process uses parameters extracted from meta-data repository. At the same time, traditional request over meta-data is accessible for system's administrator. As we are dealing with XML format, we have tried to perform this functionality by generating XSLT transformation from meta-data.

4.2 XETL Architecture

XETL's architecture (Fig.1) integrates data from different and heterogeneous sources. Moreover, XETL is an active ETL; it interacts with meta-data to give parameters for extraction programs, to generate the transformation, and to specify the loading with target schema. Hence, it optimizes the flow of data, reduces the update of warehousing process by making automatic the creation of valid transformation, and generates schema from meta-data repository. The different components of XETL tool are described below:

4.2.1 The Meta-Data Components

The Meta-data repository is the essential element in the system. It includes a set of information of which:

- The mapping model (MM): this model is used to describe the mapping expressions. By mapping expressions, we mean the needed information to identify how a target field could be mapped from a set of source

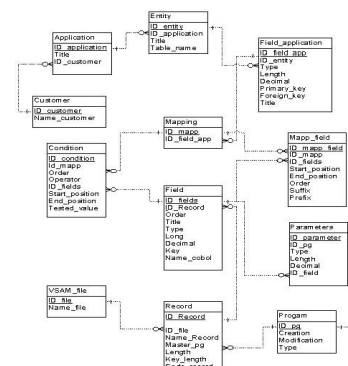


Fig 2. A mapping Meta-data model

fields. Fig.2 represents an example of MM used in [12].

- The source model (SM) contains the model of the source data. It covers relational, object oriented and semi-structured data modelling [11].
- The target model (TM) is alike to (SM); it describes the target data model. Moreover, it could cover the multidimensional model used by OLAP processing.
- Management rules (MR) is the set of rules defined by the administrator in order to fulfil business requirements.

4.2.2 The Extraction Process

The role of these programs consists in converting to a common XML format, the different sources. These sources could be traditional databases or inner digital documents that are produced by applications in enterprise or even web documents published by other partners. This first stage supplies the extraction program with the needed parameters from meta-data repository.

4.2.3 The XSLT Generator

The XSLT generator is a module, which can read useful parameters, rules, and mapping specification from meta-data repository to create a style sheet transformation. Concretely, a query is performed on the mapping model (MM). The result file is used by the Generator to produce the style sheet file. Then, an XSLT processor permits to execute the transformation on XML source documents. It generates a new collection of clean XML documents.

Indeed, the selection and filters get rid of superfluous data. During this process, all control and check are applied to data. Thus, the XSLT is more than cosmetic change for XML data; it tackles the content, structure and valid values. The architecture is not limited by the using of XSLT language; any other XML query language can be used to establish such process. For example, X-Query can be useful because it is more adapted for documents databases query and it is more optimized. On the other hand, the stability of XSLT (1.0) and the multi implementation is an advantage over the X-Query.

ModelTitle	TYPE	Table	Source	CONDITION	TITLE	POSITION
ID_Customer	char(10)	Header_After_job_customer	element dt	TOP3 = 'F' AND DONNEES='CLAY' AND CODENITE='CLIENT'	CLEMENT	
ID_HASC	char(5)	Header_After_job_customer	element dt	TOP3 = 'F' AND DONNEES='CLAY' AND CODENITE='CLIENT'	CLEENITE	
Date_HASC	YYYYMM DD	Header_After_job_customer	element dt	TOP3 = 'F' AND DONNEES='CLAY' AND CODENITE='CLIENT'	DATE	
ID_ASC	num(10)	Line_After_job_customer	element dt	TOP3 = 'L'	CLEMENT	1-10
ID_HASC	num(6)	Line_After_job_customer	element dt	TOP3 = 'L'	CLEMENT	11-14
Type	char(10)	Line_After_job_customer	element dt	TOP3 = 'L'	DONNEE2	
Description	char(14)	Line_After_job_customer	element dt	TOP3 = 'L'	DONNEE1	1-14
Characteristic	char(30)	Line_After_job_customer	element dt	TOP3 = 'L'	LEBLEM	1-30
Line_status	char(2)	Line_After_job_customer	element dt	TOP3 = 'L'	DONNEE5	1-2
Weight	num(5,2)	Line_After_job_customer	element dt	TOP3 = 'L'	DONNEE5	3-7
Location	char(10)	Line_After_job_customer	element dt	TOP3 = 'L'	DONNEE3	

Fig 3. The result of query into mapping meta-data (MM)

0 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/proceeding-paper/toward-xml-based-data-warehouse/32072

Related Content

Technology Assessment of Information and Communication Technologies

Armin Grunwald and Carsten Orwat (2018). *Encyclopedia of Information Science and Technology, Fourth Edition* (pp. 4267-4277).

www.irma-international.org/chapter/technology-assessment-of-information-and-communication-technologies/184133

Entrepreneurship Concept, Theories, and New Approaches

José Manuel Saiz-Alvarez and Martín García-Vaquero (2018). *Encyclopedia of Information Science and Technology, Fourth Edition* (pp. 3020-3031).

www.irma-international.org/chapter/entrepreneurship-concept-theories-and-new-approaches/184014

Intelligent Extraction of Intangible Cultural Heritage Visual Elements and Application in Graphic Design Based on Generative Adversarial Network

Ziwei Zhang, Zhenzhen Wan and Haotian Zhu (2025). *International Journal of Information Technologies and Systems Approach* (pp. 1-16).

www.irma-international.org/article/intelligent-extraction-of-intangible-cultural-heritage-visual-elements-and-application-in-graphic-design-based-on-generative-adversarial-network/384915

Self-Efficacy in Software Developers: A Framework for the Study of the Dynamics of Human Cognitive Empowerment

Ruben Mancha, Cory Hallam and Glenn Dietrich (2009). *International Journal of Information Technologies and Systems Approach* (pp. 34-49).

www.irma-international.org/article/self-efficacy-software-developers/4025

OPGW State Evaluation Method Based on MSIF and QPSO-DQN in Icing Scenarios

Zhigang Yan, Min Cui, Xiao Ma, Jinrui Wang, Zhihui Zhang and Lidong Yang (2024). *International Journal of Information Technologies and Systems Approach* (pp. 1-27).

www.irma-international.org/article/opgw-state-evaluation-method-based-on-msif-and-qpso-dqn-in-icing-scenarios/343318