The Denodo Data Integration Platform

Alberto Pan, Juan Raposo, Manuel Álvarez, Paula Montoto, Vicente Orjales, Justo Hidalgo
Lucía Ardao, Anastasio Molano and Ángel Viña

Denodo Technologies Inc.
Real 22, 3º
A Coruña
Spain
{apan,jraposo,malvarezd,pmontoto,vorjales,jhidalgo,lardao,amolano,avina}@denodo.com

Abstract

The world today is characterised by the proliferation of information sources available through media such as the WWW, databases, semi-structured files (e.g. XML documents), etc. Nevertheless, this information is usually scattered, heterogeneous and weakly structured, so it is difficult to process it automatically. DENODO Corporation has developed a mediator system for the construction of semi-structured and structured data integration applications. This system has already been used in the construction of several applications on the Internet and in corporate environments, which are currently deployed at several important Internet audience sites and large sized business corporations. In this extended abstract, we present an overview of the system and we put forward some conclusions arising from our experience in building real-world data integration applications, focusing in some challenges we believe require more attention from the research community.

1. Introduction

Mediator architectures[1] have received considerable attention from the research community during the last years since they potentially allow for a much faster, cheaper and far less intrusive approach for data integration than traditional techniques which relied on some kind of expensive and hard to maintain big central repository.

The Denodo Platform, constructed by following such architecture, allows for a fast yet powerful extraction and combination of information from various heterogeneous, structured or semi-structured sources, to create an unified global schema for such information.

This system has already been used for the construction of various industrial data integration applications, both in the Internet and Intranet environments, successfully integrating information from more than 700 sources such as Web sites, databases, spreadsheets, semi-structured files (e.g. XML documents), etc.

This extended abstract is organized as follows: section 2 provides an overview of the system, section 3 reports our experience in using the system in real applications, focusing in some aspects in mediator systems we believe require more attention from the research community. Finally section 4 exposes related work.

2. The Denodo Data Integration Platform

The Denodo Platform follows a mediator architecture whose components and their interrelationships are shown in Figure 1. The physical layer is comprised of wrappers for different data sources: relational databases, web sites, flat files, spread sheets, etc. Wrappers are semi-automatically generated by a tool. The logical layer, named Denodo Aggregation engine, is the core of the platform containing the Query/Plan Generator, the Optimizer and the Execution Engine. An administration tool is used to manage the data dictionary. A cache module stores materialized views, avoiding querying the sources when queries can be solved using the cache. The Denodo Planning tool permits periodical data pre-fetches. The Denodo Security SDK allows for ciphering of data when required.

The following sub-sections provide an overview of some key issues of the system: source model, wrapper generation and definition of the global schema relations.

**Proceedings of the 28th VLDB Conference,
Hong Kong, China, 2002**

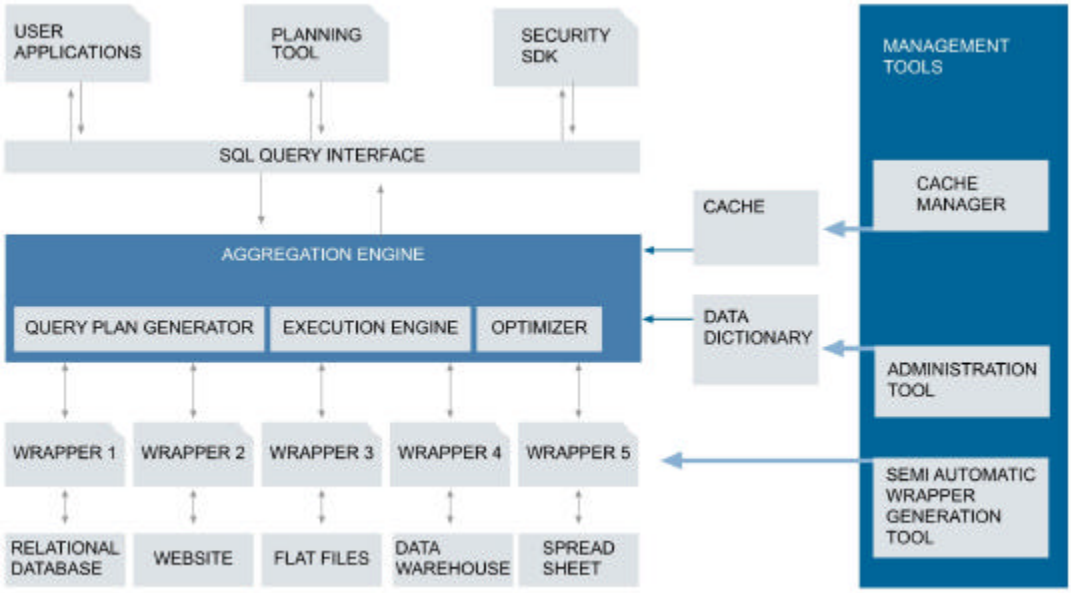

Figure 1: Denodo Platform Mediator System Architecture

2.1 Source modelling

In the Denodo Data Integration Platform each source exports a combination of relations, called base relations, following a relational-like model where record and array-type attributes are also supported.

In the context in which the mediator systems operate, base relations generally have limitations in the way in which they can be queried. For example, in Web sources query possibilities are often limited to those provided by some type of HTML form.

Each relation in a source will represent its query capability through what we term as *search methods*. Our language for query capability definition relies on real sources capabilities and has the following features:

- allows for specifying mandatory and optional fields.
- supports query languages using arbitrary operators.
- gives support to the concept of multiplicity (which indicates how many query conditions for a given attribute and operator the source can perform at the same time) and "infinite" query conditions.
- supports attributes that can only be queried by a limited set of values.

2.2 Wrapper Generation

The following stage in the creation phase is the construction of wrappers. Each wrapper must provide access to the base relations of a source in such a way that, when faced by the mediator, it behaves much alike as a table within a relational database. The wrapper generation process for Web sources, JDBC, XML databases and structured or semi-structured text files is performed with the assistance of a semi-automatic generation tool which enables wrappers to be created and maintained in a fast and simple way.

In the case of Web sources (see [15] for detail), access to and navigation of Web sources are carried out by using the Microsoft Internet Explorer (MSIE) navigator. We have created NSEQL, a navigation sequence specification language which works at a "browser-level" instead of at a "http-level" (as most previous approaches do).

NSEQL let users define "macros" directly over a web browser interface, thus making an access to a source in our system identical to the process carried out by a user who connects to this source using MSIE. This enables the wrapper to be completely independent of the source session maintenance mechanisms (which can be highly complex in commercial sources), Javascript code, dynamic HTML, etc. It also allows for wrapper navigation sequences to be generated "by means of examples", i.e. simply by navigating.

For extracting the required information from HTML pages (or other markup languages), our tool uses the approach of providing a specification language to generate specialized grammars. Our language (which is called DEXTL) makes use of various heuristics in the presentation of data to provide a simpler language than other similar ones but without a concomitant loss of power. In addition, it provides graphic tools to construct the wrapper specifications visually, through an iterative process. This enables the wrappers to be generated by staff with no programming skills.

2.3 Global Schema Relations Definition

Once the base relations have been defined and their wrappers constructed, each relation of the global schema

is defined by a query involving the base relations, in a similar way to the definition of views in a conventional database. This approach is known in mediator literature as Global As View [2]. The query is expressed in a language very similar to SQL.

It should also be pointed out that a view, like base relations, can also be defined by previously defined intermediate views, allowing therefore for a hierarchical mediator structure.

As mentioned above, the base relations can be limited in terms of their query capability. When the global schema relations are defined, the mediator is capable of automatically computing their query capabilities through the view tree. This allows for the mediator to know in advance if a certain global query is going to be answered and also makes it possible for a mediator to be a source for other mediators. See [14] for detail and examples.

2.4 Query Execution

The Denodo Platform Execution engine has the following features:

- It works asynchronously
- It is able of dealing with partial results when not all the sources are available
- It stores query cost statistics in order to predict cost information for future queries

3. Experience obtained in using the system in Real Applications

This system has been used to construct various commercial applications which are currently deployed in several large-sized operational environments. Applications constructed up to now can be classified into two different groups: 1) Search, comparison and aggregation applications on the Internet, and 2) Corporate environment data integration applications.

3.1 Internet Applications

In the first group the following applications have been constructed (among others): comparison shopping applications, job offer searches, metasearch of Web sites, news searches in on-line press, financial aggregation , etc. In these cases, all the sources on which the mediator operates are Web sites. This combination of applications are now accessing more than 500 different data sources, and they are in operation in various Internet portals aimed at the Spanish market.

In this group of applications, the methods by which source data are combined are normally quite simple (i.e. mainly unions, projections and selections).

About the materialization scheme, we usually use the virtual approach along with a cache system. The cache stores results from recent queries and it is able of determine if a new query can be answered using any subset of the current cache content.

This simple schema turns out to be very effective in this kind of applications which present a high uniformity in queries. For instance, a real music-comparison shopping application built with our system shows a cache effectiveness rate higher than 80%.

The greatest difficulties in Internet applications are encountered in the construction and maintenance of wrappers. The reason for this is that we come across a high number of sources and that they also present a degree of complete autonomy.

In our experience, the most difficult problem involved in commercial web wrapper generation is not parsing (which has been the main issue addressed in literature), but creating the navigation sequences required to access the data. Dealing with Dynamic HTML, HTTPS, frames, Javascript code or complex non-standard session maintenance mechanisms based on randomly generated session-ids, make this task extremely difficult and time-consuming when we work at “http-level” (as most research systems do). It is also a task reserved for expert programmers.

At the present time, average wrapper generation in our system takes few minutes and is performed entirely by non-programmer staff. This was only possible since we extended our previous wrapper generation techniques to generate navigation sequences at a “browser-level” instead of at a “http-level” (see [15] for detail).

3.2 Corporate Applications

In the second group, the platform has been used to developed various corporate solutions whose aim is to integrate multiple scattered and heterogeneous information into a single unified schema. Typical application scenarios include CRM (Customer Relation Management), where the customer data are distributed across many heterogeneous data repositories, EIPs (Enterprise Information Portals), where our system is used to provide an unified view over the heterogeneous content to be delivered through the portal and Business Intelligence applications, where our platform is used to extend the enterprise decisional information with relevant internet sources. For instance, our system is often used to provide an unified view which permits an enterprise to compare its own products with the equivalent from its competitors (using competitors online catalogs as source), and alert product managers when detecting special conditions such as strong price variations.

In these cases, the data are normally combined in a much more complex way than in the applications from the previous group. The problem of wrapper maintenance is minimized in these scenarios as there are a smaller number of sources, they change less frequently, and because when the corporate ones do change, the mediator administrators can usually be notified beforehand.

Nevertheless, other issues arise. Some of them follows.

A first issue derives from the fact that limitations on sources query capabilities are very common in practice. Some techniques (see [12]) have been proposed in order to address that kind of problems. Nevertheless, most research works do not compute the query capabilities of the global relations from the sources capabilities.

Our experience says that this is a must-have feature: it makes possible to use mediators as source for new ones, easily enabling incremental data integration processes. It is also very important that users know in advance the queries supported by the mediator. In our experience, industrial users do not find "trial&error" query processes acceptable.

As far as we are aware, this important issue has been dealt with only in [7]. But the query capability description framework used is too restrictive to properly model many real sources we had to aggregate. That is why we have created our own algorithm for computing mediator query capabilities [14], which is able of supporting more advanced source query features.

Finally, another big challenge is finding more automated ways for mapping format and semantic heterogeneities between sources. These applications often require to map taxonomies composed by hundreds of items. The approaches based on information retrieval techniques (such as [13]), are valuable contributions, but they are not enough to solve the problems encountered in real scenarios. Information semantic heterogeneities are much more complex and very often involve differences in granularity. Most industrial applications approach this problem by providing users with graphical tools which let them visually define the mapping rules. But, in our experience, this process is often long, tedious and error-prone. We believe that this is a major research area.

4. Related Work and Conclusions

In recent years, a significant amount of academic mediator systems has been developed, such as TSIMMIS [3] or Hermes [4]. Various specific aspects in the construction of mediator systems have also been studied by the research community: wrapper generation for Web sources [5][[6], query optimization [4] or reformulation mechanisms [8]. See [2] for a survey.

These research systems do not deal with all the complexity one can encounter in real scenarios: performance, need of a flexible schema concerning materialization of views (virtual vs. warehouse) or wrapper maintenance for existing highly complex commercial web sites

In industry, much attention has been paid to mediator systems in recent times. Application specific systems have been developed in domains such as financial aggregation or comparison shopping (Yodlee [10], MySimon [11]).

General-purpose systems such as Nimble [9] also have appeared in the market. This system follows a similar mediator architecture, however it does not deal with the important issue of wrapper generation (relying on XML data sources). They also use a semi-structured data model, providing a XML-QL query interface. Our approach uses a relational-like model and relies on providing advanced tools to create wrappers which include functionalities to handle common situations with semi-structured data (such as optional fields or converting metadata into data). In our experience, this approach simplifies the training of a database administrator to act as a mediator administrator, thus facilitating its integration into industry.

7. References

[1] G. Wiederhold. Mediators in the architecture of future information systems. *Computer*, 25(3), March 1992.

[2] Daniela Florescu, Alon Levy, Albert Mendelzon. Database Techniques for the World-Wide Web: A Survey. In *SIGMOD Record*. 27(3). 1998

[3] H. García Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman and J. Widow. The TSIMMIS Approach to Mediation: data Models and Languages. In *Proceedings of NGITS*. 1995

[4] S. Adali, K. Candan, Y. Papakonstantinou and V.S. Subrahmanian. Query Caching and Optimization in Distributed Mediator Systems. In *Proceedings of ACM SIGMOD Conference on Management of Data*, Montreal, 1996

[5] C.A. Knoblock, K. Lerman, S. Minton and I. Muslea. Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. In *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. 1999

[6] Ling Liu, Carlton Pu and Wei Han. XWRAP: An XML-enabled wrapper construction system for web information sources. In *Proceedings of IEEE ICDE*. 2000

[7] Ramana Yerneni, Chen Li, H. García-Molina and Jeffrey Ullman. Computing Capabilities of Mediators. In *Proceedings of ACM SIGMOD Conference*. 1999

[8] Alon Y. Halevy. Theory of Answering Queries Using Views. In *ACM SIGMOD Record vol. 29, nº 4*. December 2000.

[9] Denise Draper, Alon Halevy, Dennis Weld. The Nimble Integration Engine. In *Proceedings of ACM SIGMOD 2001*.

[10] Yodlee Corporation. *http://www.yodlee.com*

[11] Mysimon. *http://www.mysimon.com*

[12] Yannis Papakonstantinou, Ashish Gupta and Laura Hass. Capabilities-Based Query Rewriting in Mediator Systems. In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems*. 1996

[13] William Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. En *Proceedings of ACM SIGMOD*. 1998

[14] Alberto Pan, Manuel Álvarez, Juan Raposo, Paula Montoto, Anastasio Molano and Ángel Viña. A Model For Advanced Query Capability Description in Mediator Systems. In *Proceedings of the ICEIS Conference.* 2002.

[15] Alberto Pan, Juan Raposo, Manuel Álvarez, Justo Hidalgo and Ángel Viña. Semi Automatic Wrapper-Generation for Commercial Web Sources. *To appear in Proceedings of IFIP WG8.1 EISIC*. 2002