Information integration among Heterogeneous and Autonomous Applications

Benabdelkader, A.

Publication date
2002

Chapter 7

Conclusions and Future Work

7.1 Overview

A wide variety of distributed applications are nowadays emerging in diverse domains of science, business, engineering, education, e-commerce, tourism, etc. These applications deploy various database systems for the management of their information, in which the diversity stems from reasons related to the specific information management requirements and the objectives targeted by these applications. Other reasons may also concern suitability, efficiency, and security. In today's organizations, new and existing applications such as design, manufacturing, or decision making environments, require access to data stored in several of pre-existing databases detained at several local and remote sites. To satisfy the new information management requirements of these organizations, a strong information integration system must be designed and developed, serving the need for information integration and interoperation among these organizations.

This dissertation describes the design and development of an information integration approach to support the integration of heterogeneous information sources while preserving their local autonomy and distribution. The first step in this direction consists of a global survey focusing on the related research and approaches for information integration and interoperation among autonomous and distributed systems. The survey of existing approaches emerging in this domain forms the ***state-of-the-art*** and the related research work for the dissertation. Considering the main emphasis of the thesis, this survey is conducted by a *classification of existing approaches and methodologies* for information integration. Classification of these approaches, as addressed so far by other researchers, is mostly based on three concepts of database architectures, data access and storage mechanisms, and systems interoperation. The taxonomy for information integration approaches, proposed in chapter 2, divides them into two main categories: *Distributed Systems* and *Integrated Systems*. Distributed systems typically share common database control software at both DBMS servers and applications. Integrated systems however, support database applications that address decentralized/autonomous database control, using different representations and data modeling systems. Within each of these two categories several approaches are identified, studied, and evaluated based on the applications' requirements.

Research on integrated systems distinguishes between *Physical Integration* and *Virtual Integration*. In a *Physical Integration* the data originating from local and remote sources are integrated into one single database on which all queries can operate. In *Virtual Integration*,

data remains on the local/remote sources, queries operate directly on them and data integration takes place on the fly during the query processing. At deeper levels of the taxonomy, when the required level of integration becomes more complex and when the requirements are higher, the variety of the proposed approaches becomes more and more specific and complex. On one hand, the physical integration expands into *centralized databases* and *data warehouses*. In a centralized database, information is migrated from various sources into a universal DBMS, while in data warehousing information may be imported in different format and volume than it exists in its originating sources. On the other hand, the virtual integration derives into *federated* and *non-federated* systems. Furthermore, each of these systems can be either *loosely* or *tightly coupled*.

Most approaches presented and discussed in chapter 2, do not properly support the extensibility and the evolution of applications, rather they address specific domain-dependent cases. Adding a new site to the federation, or applying a given approach to a different application domain, requires considerable expertise and effort in order to interface it with all systems participating within the federation. Still, these approaches bring considerable advantages, which can be adapted and deployed. The information integration approach, we presented in chapter 6, benefits from these approaches and follows a strategy, supporting standard languages, generic tools, and middleware solutions. The major benefits from which, the integration approach takes advantages are discussed in details in section 7.2.

The development of several systems and tools to support the management and exchange of information and the data integration purposes, during the preparation of the thesis, has provided the means to better understand the complexity of such processes, and to better deploy standard tools and middleware solutions.

1. The ***Waternet*** system, presented in chapter 3, aims at an evolutionary knowledge capture and management system supporting the control, optimal operation, and decision making for the management of water distribution in a network of expert systems, provided the proper environment for developing an open and flexible architecture for integration of different Waternet modules. From the development of the Waternet integrated/federated environment, we learned how to design and develop flexible, open, and reliable environments for information management systems supporting the following characteristics:

   - System openness, so that different sites can be added to/removed from the federation community, in order to support the specificities of different application domains.
   - Data distribution, so there is no need to develop a single global schema and a common glossary of concepts among the networked sites.
   - Complete data location transparency to the user, of logical/physical distribution of information among the sites in the network.

   The developed aspects and the lessons learned, during the design and development of the *Waternet* system, contributed to **GFI$_2$S** by tackling the fundamental schema management challenges at the federation layer, and by serving the system openness through the adoption of the data adapters at the node layer.

2. The ***MegaStore*** framework, presented in chapter 4, aims at the design and set-up of the necessary database structure and platform architecture for advanced e-commerce applications, and in specific, addressing the CD and music industry. It provides a

good example for the deployment of database standards and middleware solutions. Its development is supported through the coupling of Web standards and middleware with advanced database technologies. The main idea behind the developed framework for MegaStore is to design a comprehensive system to support applications with the following characteristics:

- Facilitate the storage and manipulation of multimedia large data sets.
- Provide a flexible information classification and clear separation between public and proprietary data.
- Extend Web services in E-Business applications with the functionalities for flexible navigation through complex Web objects, scalability as necessary for multimedia large objects, high performance as required by multi-users applications, and so on.

The design and development of MegaStore framework contributed to **GFI$_2$S** through (1) the deployment of database standard and Internet Middleware supporting system reusability, (2) the development of a parallel/distributed database server assuring system efficiency, and (3) the development of user friendly interfaces assisting advanced/ordinary users in accessing the underlying information sources.

3. The ***Virtual Laboratory (VL)*** Information Management, presented in chapter 5, aims at the design and development of a software layer and an enhanced architecture, supporting scientists in their experimentations, and providing the basic information management requirements for the emerging multimedia applications in e-science. Our contribution within VL Information Management focuses on specific advanced features, functionalities, and facilities introduced and developed for management of scientific data for VL applications. Specific subjects addressed within VL include:

   - *Strategies* for storage and retrieval of large scientific data sets.
   - *Use of standards* for scientific data modeling and archiving.
   - *Universal and schema free access* to scientific data.
   - Access to scientific data based on the *predefined visibility restricted schemas*.
   - *Scientific Results Publishing, performance issues, Benchmarking tests*, etc.

   The concepts, addressed within the VL Information Management framework, addressed a number of important issues that can be applied at every site to better enable it as a node in the cooperation network. On one hand, individual sites can benefit from the use of these concepts to efficiently build their applications independently of other sites. On the other hand, the use of standards at local sites helped the development of information integration mechanisms for **GFI$_2$S**.

The approaches for *Waternet* and *MegaStore* systems provide specific mechanisms for the design and development of information integration systems, and illustrate the main benefits of using standard solutions to support the information sharing and the data integration. While, the VL Information Management, particularly employs these standard concepts to the information integration mechanism and provides a forward benefit to it. VL information management framework applies the Web and database standards at every site, making its components stronger and more suitable for integration and cooperation, while preserving their full autonomy and specific characterization. The various concepts and lessons learned

during the development of the application cases, described above, have contributed to the design and partial development of a generic and flexible information integration system.

The ***Generic and Flexible Information Integration System (GFI_2S)*** is designed and partially developed to give its users and applications, access to heterogeneous information sources through generic and flexible interfaces. The distinctive features of the **GFI_2S** integration approach reside in (a) *the specific combination* of database standards and Internet middleware with the fundamental research approaches, and (b) *the way in which they are deployed and inter-linked* within the components of GFI2S architecture. Its architecture smoothes the transition from relational and object-relational database systems to a system that unifies most DBMSs capabilities. The **GFI_2S** approach, which follows the ODMG standards, is considered to support different data sources and provides application developers with a single, seamless application view, and unified access to all information in those underlying data sources. The **GFI_2S** integration architecture is constituted of two main components of: (1) Local Adaptation Layer (LAL), that facilitates the access to the underlying databases in the node, and (2) the Node Federation Layer (NFL), that provides links to the information and applications outside the nodes and supports the information sharing and interoperation. This two-component architecture of **GFI_2S** provides existing systems with efficient means for their interconnection and interoperation, while preserving their heterogeneity, distribution, and full autonomy. The **GFI_2S** architecture benefits from existing approaches and applies emerging information technology to support the new requirements of scientific and advanced applications. More details regarding the benefits of **GFI_2S** comparing to other approaches are described within the next sections.

7.2 GFI_2S Compared to Other Approaches

The design approach of **GFI_2S** benefits from the careful choice of its specific components. The components are decided based on the state-of-the-art in the area of information integration for systems interoperation. Additionally, in order to support the new requirements from emerging scientific and advanced applications, the **GFI_2S** approach utilizes database standards, Internet technology, and Middleware solutions. The main aspects which are taken into account in the design of **GFI_2S** system are listed below:

- Standard languages for data modeling and information access are adopted at the federated layer of **GFI_2S**. In the area of information exchange and data integration, several initiatives are emerging in the direction of standardization (e.g. STEP, and NetCDF). Most initiatives consider their specific terminologies for the data representation and manipulation. Thus, new standards are appearing rapidly, while, similar solutions already exists: the database standards. The **GFI_2S** approach benefits from these initiatives and extends the usage of their architectures via the use of database standards in term of data modeling, querying language, and information exchange. In the **GFI_2S** information integration approach, data is not bi-translated, rather, queries are sent from one site and data is received from the other site.
- The federated schema constitution within **GFI_2S** is based on and extends the PEER approach. PEER uses its specific language for schema definition, mapping derivation, and query formulation. While, **GFI_2S** uses the UML for data modeling, ODL for data definition, and OQL for data access and information retrieval. Such an approach for schema and data management makes **GFI_2S** an open integration facility for other systems, which are compliant to these standards.

- The usage of XML and OIF data formats within the **GFI$_2$S** system, facilitate the applicability of database concepts to the federation by enforcing the data exchanges between different organizations in a widely accepted format. The availability of data in standard formats of XML and OIF, within the collaborative environments, reduces the number of wrappers to be developed, and facilitates the data translation among heterogeneous systems when services are requested between them.
- The use of Object-Oriented database standards as common languages for data modeling and querying provides the possibility for integrating most types of applications ranging from the CODASYL network model, to relational model, to object-oriented and Object-Relational models.
- The use of extended ODMG mechanisms supports the mapping specification and derivation operations between the underlying data sources and the integrated database schema, at the **GFI$_2$S** federated layer.
- The structural representation and the semantics resolution of data from heterogeneous sources are enforced at the **GFI$_2$S** federated layer by a dictionary of terms and a dictionary of semantics. These dictionaries, which are available for each exported/integrated schema, help users in defining their own federated schemas without the need for external support from experts that devote the sharable (exported) information. The dictionary of terms serves for automatic conflicts resolution, while, the dictionary of semantics reflects in fact the experts' knowledge of the application.
- The **GFI$_2$S** federated architecture adapts the approach of defining conceptual wrappers for legacy databases and developed the Local Adaptation Layer (LAL), in order to provide interoperability between legacy systems. The LAL extends the role of wrappers to also include information about users authentication and information visibility levels.
- The specific data structure and querying language of each node within the federation are preserved. Queries formulated at the **GFI$_2$S** federated layer are translated to be conform to the local data source query language before being executed. Additionally, the local results are translated to the common format adopted at the federated layer. Therefore, the **GFI$_2$S** federated approach does not require translating the complete existing data of different databases to the federated layer, rather, it focuses on translating only the part of data that is needed to be exchanged, e.g. the result of a query into a common format.

7.3 Lessons Learned

Generic and Flexible Information Integration Systems must satisfy the requirements of Flexibility and Genericness. From the design of the **GFI$_2$S** information integration approach, we learned that flexibility of information integration systems resides in the architecture they rely on, while their genericness can be achieved via the deployment of database standards, Internet technologies, and middleware solutions. Thus, the learned aspects are to be considered in the design and development of information integration systems, in order to provide an open facility for integration/interoperation among heterogeneous, distributed, and autonomous sites. Below is a list of the main lessons learned and the expertise gained within the design and development work of the various R&D projects, during the preparation of this dissertation:

❶ The use of two components (LAL and XFL) for the information integration among networked sites makes the integration mechanism flexible. This flexibility is supported from two sides, on one hand, sites can join or quit the federation; on the other hand, the schema integration strategy followed at the node federation layer allows for a customized integration, which can be tailored to the need of each site.

❷ The use of object-oriented database standards and middleware solutions at the federated layer of the **GFI$_2$S** makes its architecture generic. Each site that wishes to join the federation only needs the knowledge about its "underlying database system" and about the "standard languages and formats" adopted at the federation layer. The local users at each site gain proper expertise about the underlying local application's characteristics and specifications. At the same time, the standard languages and formats adopted at the federation layer are mostly understood by these users.

❸ The use of standard languages for data definition and information access (ODL, OQL/SQL, XML), which are widely adopted by a large community, reduces the efforts needed when defining export and integrated schemas, and facilitates the access to data within networked applications.

❹ The use of middleware and standards mechanisms for data access (e.g ODBC and JDBC), information exchange (e.g. XML), and communication protocols(e.g. CORBA) play an important role in reducing the number of intermediate interfacing tools, unifies the access to shared information, and facilitates the data integration among heterogeneous databases and applications.

❺ The Consideration of data aspects such as, scientific information, large data sets, and complex inter-linked objects supports the development of complex applications. In addition, the provision of generic mechanisms and tools for scientific data publishing, based on tailored views on the sharable data, preserves systems' autonomy and hides private data from outside users.

Various concepts, enumerated above, provide the base information integration aspects for systems interoperation. The next section will identify some of the remaining issues, in the area of information integration, that need to be further addressed and described.

7.4 Future Work

In order to facilitate the information integration process among heterogeneous applications, attempts to integrate autonomous, distributed, and heterogeneous applications must be strongly based on the use of database standards and middleware solutions. Middleware solutions unify the communication process among interconnected applications, while database standards unify their exchange of data. Thus, the use of database and middleware standards constitute the base for flexible, open, and generic integration among networked applications. Use of standards for data modeling and information retrieval also eases the interoperation/collaboration process with pre-existing application and legacy systems, and reduces the need for construction of individual data translation wrappers.

However, in order to apply standard concepts to the environment of networked applications, certain extensions to database and middleware standards must be addressed to better support the information exchange and data integration among interoperable systems. For

instance. In the database area (similar to many others), standards lag behind in supporting new features and the extensions provided by certain commercial and research database management systems.

Following areas require to be further addressed by the researchers in the field of information management, by the standardization community, and by the DBMS developers:

- The development of advanced standard constructs to better support the specific requirements of complex scientific applications, and to address their data types, object-orientation concepts, interoperation/integrated facilities, distributed computing, etc.
- The extension of database definition language to properly support the mapping constructs and the derivation operations is required, to better support the federation of several heterogeneous databases. As such the object definition language (ODL), for instance, needs to be extended to support the mechanism of schema integration in the area of federated databases. The extensions to the ODL are expected to address the definition of export and integrated schemas, and to provide a set of operations supporting the needs for their derivation mappings.
- The extension of database query language with object-oriented features is required. Currently, different DBMS developers use their specific SQL/OQL extension mechanisms, which differ from one DBMS to another. To overcome the issue of specific extensions, if not by standardization, there must be a consensus among DBMS developers about these extensions; at least a common agreement regarding the main required features such as object identifier, inheritance, path expression, and cross-relationship references must be achieved.
- The consideration of standard data exchange formats, e.g. XML for databases and in particular for information integration, raises several challenges for database research. Having XML focused only on the syntax for data representation partially increases the prospect of its integration. To support information integration at the semantic level, however, there must be a further standardization or agreements upon DTDs (schemas) in XML. Furthermore, at present, the XML data does not conform to a fixed schema: names and meanings of the used tags are arbitrary and the data is self-describing in XML documents. Therefore, several XML-issues need to be addressed to enable the information integration (e.g. languages for describing the contents and capabilities of XML sources, query reformulation algorithms, translation among DTD's, and obtaining source descriptions).

In addition, we must also admit that the issue of information integration is of a very high complexity, especially when the information sources are heterogeneous, distributed, and their local autonomy is preserved. This thesis work described a high level architecture for information integration, in order to give a global overview of a generic and flexible information integration system. Therefore, complete descriptions and full coverage of all the components of **GFI$_2$S** are out of the scope of this dissertation. Several components of **GFI$_2$S** are addressed to the required level of details, while others are globally described, leaving a number of issues and problems to be further addressed by other researchers. Among the remaining issues that require further research, we enumerate: updates in federated schemas, derivation mappings to cover the relationship concepts, and better addressing technology-independent issues (e.g. theories, and formal specifications). Some other issues in the domain of information integration were already addressed by the research community, for that reason these issues are addressed but not fully described in **GFI$_2$S** (e.g. federated query processing, wrappers, and semantics resolution).

However, the **GFI$_2$S** architecture is flexible enough to be augmented with software components, which are designed/developed by other research/software institutions. For instance, the **GFI$_2$S** architecture allows an application to use an existing tool for conflicts resolution when defining export and integrated schemas. Such a tool can be based on, and enforced by, the dictionary of terms and dictionary of semantics defined at the federation layer of **GFI$_2$S**.