XML Data Integration in OGSA Grids

Carmela Comito and Domenico Talia

DEIS, University of Calabria,
Via P. Bucci 41 c,
87036 Rende, Italy
{ccomito, talia}@deis.unical.it
http://www.deis.unical.it/

Abstract. Data integration is the flexible and managed federation, analysis, and processing of data from different distributed sources. Data integration is becoming as important as data mining for exploiting the value of large and distributed data sets that are available today. Distributed processing infrastructures such as Grids can be used for data integration on geographically distributed sites. This paper presents a framework for integrating heterogeneous XML data sources distributed among the nodes of a Grid. We propose a query reformulation algorithm to combine and query XML documents through a decentralized point-to-point mediation process among the different data sources based on schema mappings. The above cited XML integration formalism is exposed as a Grid Service within the GDIS architecture. GDIS is a service-based architecture for providing data integration in Grids using a decentralized approach. The underlying model of such architecture is discussed and we show how it fits the XMAP formalism/algorithm.

1 Introduction

The Grid offers new opportunities and raises new challenges in data management arising from large scale, dynamic, autonomous, and distributed nature of data sources. A Grid can include related data resources maintained in different syntaxes, managed by different software systems, and accessible through different protocols and interfaces. Due to this diversity in data resources, one of the most demanding issue in managing data on Grids is reconciliation of data heterogeneity. Therefore, in order to provide facilities for addressing requests over multiple heterogeneous data sources, it is necessary to provide data integration models and mechanisms.

Data integration is the flexible and managed federation, analysis, and processing of data from different distributed sources. In particular, the rise in availability of web-based data sources has led new challenges in data integration systems for obtaining decentralized, wide-scale sharing of data, preserving semantics. These new needs in data integration systems are also felt in Grid settings. In a Grid it is not suitable to refer to a centralized structure for coordinating all the nodes because it can become a bottleneck and, most of all, it doesn't benefit from the dynamic and distributed nature of Grid resources.

The Grid community is devoting great attention toward the management of structured and semi-structured data such as databases and XML data. The most

J.-M. Pierson (Ed.): VLDB DMG 2005, LNCS 3836, pp. 4–15, 2005.

significant examples of such efforts are the *OGSA Data Access and Integration* (OGSA-DAI) [1] and the *OGSA Distributed Query Processor* (OGSA-DQP) [2] projects. However, till today only few of those projects [3,4] actually meet schema-integration issues necessary for establishing semantic connections among heterogeneous data sources.

For these reasons, we propose a framework for integrating heterogeneous XML data sources distributed over a Grid. By designing this framework, we aim at developing a decentralized network of semantically related schemas that enables the formulation of distributed queries over heterogeneous data sources. We designed a method to combine and query XML documents through a decentralized point-to-point mediation process among the different data sources based on schema mappings. We offer a decentralized service-based architecture that exposes this XML integration formalism as a Grid Service [5]. We refer to this architecture as the *Grid Data Integration System* (GDIS). The GDIS infrastructure exploits the middleware provided by OGSA-DQP, OGSA-DAI, and Globus Toolkit 3 [6], building on top of them schema-integration services.

The remainder of the paper is organized as follows. Section 2 presents a short analysis of data integration systems focusing on specific issues related to Grids. Section 3 presents the XMAP integration framework; the underlying integration model and the XMAP query reformulation algorithm are described. Section 4 illustrates the deployment of the XMAP framework on a service-based Grid architecture. Finally, Section 5 outlines future work and draws some conclusions.

2 Data Integration and Grids

The goal of a data integration system is to combine heterogeneous data residing at different sites by providing a unified view of this data. The two main approaches to data integration are federated database management systems (FDBMSs) and traditional mediator/wrapper-based integration systems.

A federated database management system (FDBMS) [7] is a collection of cooperating but autonomous component database systems (DBSs). The DBMS of a component DBS, or component DBMS, can be a centralized or distributed DBMS or another FDBMS. The component DBMSs can differ in different aspects such as data models, query languages, and transaction management capabilities.

Traditional data integration systems [8] are characterized by an architecture based on one or more mediated schemas and a set of sources. The sources contain the real data, while every mediated schema provides a reconciled, integrated, and virtual view of the underlying sources. Moreover, the system includes a set of source descriptions that provide semantic mappings between the relations in the source schemas and the relations in the mediated schemas [9].

Data integration on Grids presents a twofold characterization:

1. data integration is a key issue for exploiting the availability of large, heterogeneous, distributed and highly dynamic data volumes on Grids;
2. integration formalisms can benefit from an OGSA-based Grid infrastructure, since it facilitates dynamic discovery, allocation, access, and use of both data