Towards Dynamic Information Integration

Jürgen Göres

University of Kaiserslautern, Heterogeneous Information Systems Group
goeres@informatik.uni-kl.de

Abstract. To utilize the full potential of structured or semi-structured data stored across different information systems, users and applications must not be confronted directly with the individual, heterogeneous data sources, but instead be supplied with a customized integrated view on the data. Traditional information integration is relying on a human-driven process to accomplish this task. While feasible in static, closed-world scenarios, this approach fails in settings like the nascent data grids, which are characterized by a large, permanently changing set of autonomous data sources. We describe the end-to-end integration approach underlying our PALADIN project which aims to reduce and ultimately eliminate the dependency on human experts in the integration process in order to provide fast and cost-effective integration services for these dynamic environments.

1 Introduction

Grid technology aims to provide distributed, potentially global access to computing resources by providing a layer of middleware that abstracts not only from location but ultimately also from the inevitable heterogeneity of these resources. Originating in the area of high performance computing, grids initially focused on the use of CPU and memory resources. In these *computing grids*, the transfer of data is inherently file-based, as it was initially understood only as part of the general infrastructure to pass job executables, input data and results between grid nodes. The good reliability and performance characteristics of this infrastructure led to the adoption of grid technology as a means to provide a large distributed storage space for data-intensive research projects like particle physics. While they are often called *data grids* by their users, we will refer to them as *storage grids*, as their very nature is still just that of a large container for bulk data stored in files, which – from the perspective of the grid – have no discernable structure. Given the massive amounts of raw data these distributed file systems have to deal with, the focus of research and development is on performance and replication aspects.

Today, the understanding of the term data grid is moving away from this file-centric virtual hard disk, towards a grid where preexisting structured or semi-structured data stored in databases is the resource of interest. Thousands of publicly available databases already exist, for example, in areas like life sciences and space observation. However, access to these sources is far from being uniform,

J.-M. Pierson (Ed.): VLDB DMG 2005, LNCS 3836, pp. 16–29, 2005.
© Springer-Verlag Berlin Heidelberg 2005

based on proprietary interfaces. Beyond these technical barriers, users are faced with different data models, each offering many degrees of design freedom that result in schemas with largely differing structures even when modelling the same real-world concepts. To make things worse, the vocabulary used to describe these concepts is inconsistent even in comparably narrow fields of interest.

Current development in standardization bodies like the Data Access and Integration working group of the GGF focuses on defining interfaces that abstract from technical heterogeneity by building on the well-established web service technology. However, plain access is not sufficient to effectively use data coming from different data sources available on a future data grid, but only represents a necessary first step in the development of data*base* grids. To truly benefit from publicly available data sources, users and applications require a transparent, integrated view on the distributed data sources that abstracts not only from different hardware, operating systems, protocols and interfaces, but also from the logical and semantical heterogeneity among the schemas of these sources.

Providing such an integrated view requires a mapping from the data source schemas to the integrated schema that adapts the different data models and structures as well as the diverging uses and understandings of terms in these schemas. In conventional integration scenarios, which deal with fixed requirements and a relatively stable set of data sources under a single administrative control, this mapping is developed in a human-driven process, involving experts both of information integration and of the respective application domain. We argue that while existing approaches to information integration are appropriate under these static circumstances, they are unsuitable for the ad hoc nature of cooperation and the high degree of autonomy of the data sources in a grid environment, due to their heavy reliance on human expertise and the resulting cost and delay.

In this paper, we describe an end-to-end process to automate information integration that is customized to the dynamic nature of data grids. We start with the analysis of the user requirements for the desired integrated database, followed by the discovery of data sources that can contribute. Next, a mapping is created between these sources' schemas and the integrated schema, which is then deployed in a runtime environment.

This process serves as blueprint for our current research project PALADIN (Pattern-based Approach to LArge-scale Dynamic INformation integration). We describe the basic elements of PALADIN's infrastructure as well as two central concepts we use to augment and ultimately replace human expertise in the integration process: *Domain schemas*, essentially data-centric ontologies, provide machine-processable domain knowledge by describing the concepts of a given application domain and their relationships. *Integration patterns* are used to declaratively describe common problem constellations encountered when mapping between heterogeneous schemas and to specify an approach to their solution, which can then be adapted and applied to the concrete problem.

The remainder of this paper is structured as follows: Section 2 discusses related work. Section 3 gives a short overview of the five conceptual phases of information