A Framework for Heterogeneous Data Integration Using Grid Technology

Thandar Su Nge Htwe
University of Computer Studies, Mandalay
Thandar.tdsnh@gmail.com

Abstract

Grid has emerged recently as an integration infrastructure for sharing and coordinated use of diverse resources in dynamic, distributed environment. OGSA-DAI (Open Grid Services Architecture Data Access and integration) is a framework for building distributed data access and integration systems. In this paper, we present a system for integration of heterogeneous databases based on Grid technology, which can provide a uniform access interface and efficient query mechanism to different databases. Mediator-wrapper architecture is used to implement the system in which existing services of Open Grid service Architecture-Data Access and Integration (OGSA-DAI) middleware is also used as wrappers to provide a standard interface of heterogeneous data resources. The query processor has many extensibility points, making it easy to customize. In addition, we tend to introduce a new OGSA-DAI views resource that provides a flexible method for defining views over XML data.

1. Introduction

Nowadays, data integration is becoming important in many commercial applications and scientific research. It integrates multiple databases and heterogeneous resources on the internet so that a unified views of these databases and resources can be provided to users. While many modern medical applications involve accessing heterogeneous databases in a geographically distributed environment, there is an arising across different information systems or administrative domains to provide a uniform access interface for users. Most effort has been made in the research on integration of heterogeneous data resources, including the early multi-database systems [1].

Grid technology is an emerging technology for seamless and loose integration of diverse resources distributed on the Internet. In order to achieve federation of heterogeneous databases, we have developed a system for supporting a drug discovery process using Globus Toolkit3/OGSA-DAI [2]. As the grid is evolving from providing raw computing power to a infrastructure integrating huge amount of data from various organizations and origins, middleware which integrates and homogenizes this data is urgently needed for the successful development of higher level services, e.g. for data mining. The prototype implementation which is based on XML schemes to provide the information needed to build a virtual data source, already shows the feasibility of the developed concepts [3].

OGSA-DAI services provide metadata about the DBMS, e.g. whether it is an Oracle, DB2 or MySQL, etc., DBMS system that is being exposed to the Grid. For relational databases the database schema may be extracted from the service, which may be helpful to higher level services such as distributed query processing. The metadata may be provided statically, that is when the service is configured, or dynamically which may require additional coding. On the whole the static metadata model is extensible so that communities that employ OGSA-DAI to access databases within a Grid context can provide community specific metadata for the databases they expose to the Grid [4].

In this paper, we put emphasis on database integration by employing Grid technology. Grid addresses the issue of heterogeneity by developing common interfaces for access and integration of diverse data sources. These interfaces can be used to represent an abstract view of data sources, which can permit homogeneous access to heterogeneous databases.

2. Related Work

In report [5], they discussed the various applications and the results of the use of OGSA-DAI on various databases. The main purpose of OGSA-DAI is to provide the required data resource sharing not only to support the access to disparate resources but also to transform, integrate and deliver the data.

ISPIDER project[6]- aim is to provide an environment for constructing and executing analyses over proteomic data and a library of proteomics-aware components that can act as building blocks for such analyses.

AutoMed wrappers[7] are used to extract sources' metadata and accomplish schema mapping and integration described view generation and view optimization in the AutoMed heterogeneous data integration framework. In AutoMed, schema integration is based on the use of reversible schema transformation sequences. And then, they presented techniques for optimizing these generated views, firstly by optimizing the transformation sequences, and secondly by optimizing the view definitions generated from them.

In report [8], OGSA-DQP is a service-based distributed query processor on the Grid, which aims to exploit the service-oriented middleware provided by OGSA-DAI and OGSA reference implementation, Globus Toolkit (GT), by plugging into the port types defined by the constituent services of those frameworks.

In our work, we tend to introduce a new OGSA-DAI views resource that provides a flexible method for defining views over XML data.

3. Background Theory

3.1. OGSA-DAI

Open Grid Service Architecture –Data Access and Integration (OGSA-DAI) is a project that develops middleware to assist with access and integration of data from separate sources via the Grid [9]. The goal of OGSA-DAI is to provide a uniform service interface for data access and integration to databases exposed to the Grid, hiding differences such as database driver technology, data formatting techniques and delivery mechanisms. Therefore, OGSA-DAI services can provide the basic operations that can be used by higher level services to offer greater functionality, such data federation and distributed queries. The OGSA-DAI architecture also encourages the design of efficient applications by supporting for grouping multiple requests on an OGSA-DAI service into a single message sent to a service. This reduces latency by increasing both the number of messages exchanged and the quantity of data transferred.

Grid Data Service (GDS) is the primary OGSA-DAI service that supports the interaction with a single database [10].It manages an authenticated collection of requests to a resource and stores the status and results of requests. An instance of a GDS is created by the Grid Data Service Factory (GDSF). GDSF can be located by a DAI service Group Registry (DAISGR), which is used to publish data resource metadata and capabilities. At present, data resources that can be exposed via OGSA-DAI include: relational databases such as MySQL, SQL Server, IBM DB2, Oracle; XML databases such as eXist, Xindice; files and directories in format such as OMIM,EMBL.

The extensibility and scope of OGSA-DAI made it an ideal tool when researchers began looking at better ways to manage the massive amounts of information created by some scientific research programs. Some projects have employed OGSA-DAI successfully as their core functional component of data in integration and management.

Database systems are well-known for consistent storage, retrieval and manipulation of

data. At the same time the Extensible Markup language (XML) is generally accepted as data description language for both web-based information systems and electronic data interchange between different organizations. XML has become the first choice for data exchange between different organization.XML document fall into two broad categories: data-centric and document-centric. Data centric documents are those where XML is used as a data transport. XML documents are self-describing, it., XML tags allow to describe the meaning of the content itself. New tags and attributes can be defined, document structures can be nested to any level of complexity and documents can be associated with and validated against a schema specification in terms of document type definition [11].

4. Framework for the data Integration

As Grid technology deals with sharing distributed heterogeneous resources without compromising local resources administration and OGSA-DAI provides functions to federate heterogeneous data resources, we designed here a system for integrating heterogeneous databases based on Grid technology with OGSA-DAI. The goal of the system is to provide a uniform query and access mechanism for heterogeneous databases. OGSA-DAI is a powerful workflow engine geared towards integration of heterogeneous data resources. The most commonly exploited feature of the framework is its ability to provide Web service wrappers to data sources. However, this feature alone is only a small subset of the OGSA-DAI feature set. The OGSA-DAI workflow engine allows data processing to take place close to the data. Users can integrate multiple data resources exposed by a single OGSA-DAI server, but more importantly it is possible to build complex data integration systems by coordinating workflow execution and data transfer between several distributed OGSA-DAI web services.

4.1. Architecture of the Framework

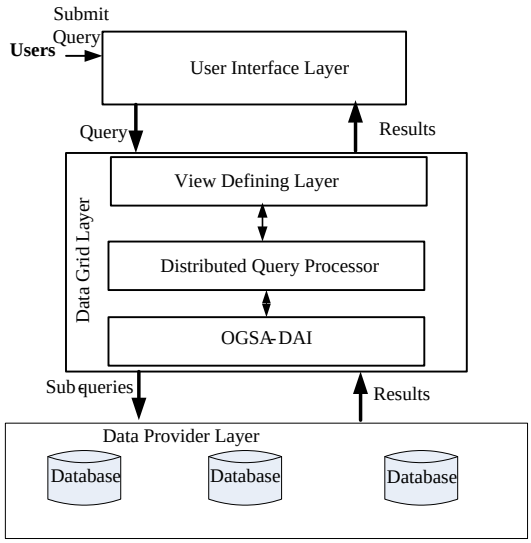

Figure1. Overview of the Framework

Overview of the proposed system is depicted in figure1. It has three main layers :(i) Data provider Layer (ii) Data Grid Layer (iii) User Interface Layer.

4.1.1 Data Provider Layer

This includes distributed database management systems which use database servers to handle large amount of records which are stored in relations, where each row (called tuple) represents a data item and each column is an attribute describing those items. Different databases have been created without following a predetermined common schema. Each system is a complete DBMS in itself with access control catalogs and query processing, etc.

4.1.2 Data Grid Layer

The data grid layer plays the vital role in the proposed system. It aggregates all dispersed data resources into a single and uniform view. It also address authentication, secure transfers and mechanisms for searching desired information provided by users. There are three main component in Data Grid layer (1) OGSA-DAI (2) distributed query processor (3) View defining Layer. As depicted in figure2. Mediator –wrapper approach is used in the system to build the Data Grid infrastructure in which the process of wrapping databases into the grid environment is executed by OGSA-DAI.A strong point of using OGSA-DAI as a wrapper to build the Data Grid infrastructure is that a wide range of relational, object-relational, and XML database products is supported, along with flat files and its ability to provide web-service wrappers to data sources in a uniform way as services. In our framework, OGSA-DAI functionalities are used to make disparate, heterogeneous data sources be available as services that are fully integrated with other OGSA-DAI, while hiding differences such as database driver technology, data formatting techniques and delivery mechanisms. Distributed query processor is used on top of OGSA-DAI as a mediator to federate all OGSA-DAI wrapped data resources. View defining Layer is built on top of them to simplify writing complex queries and to solve schema mismatch problems of different databases.

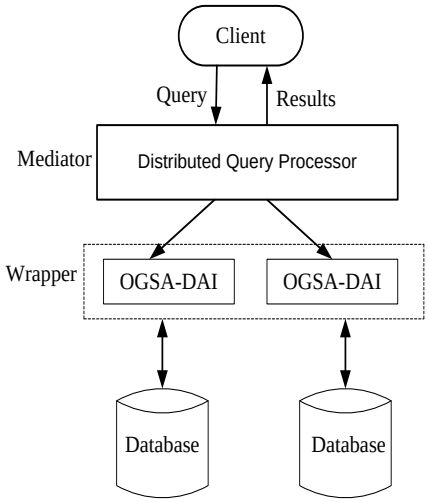

Figure2. Mediator- wrapper vision of Data Grid Layer

4.1.3 User Interfacing Layer

The main role of this layer is to act as a gateway that connects end-users with the data grid while hiding the grid complexities and the knowledge of query language. A web-based user interface is used to implement this layer from which user can submit queries using web browsers. It offers a single and easy-to-use access point to all distributed data and computing resources included in the system. Web –based user interface provides an interface for querying information locating on distributed databases. Since it aggregates contents from multiple sources into a single view, it provides a uniform access interface and efficient query mechanism to different databases.

4.2 Distributed Query Processor

In grid environment, the distributed query processor resource follows the usual query processing stages, it depicted in figure3. Distributed query processor is build by extending

OGSA-DAI's data resource services, activities. It performs two main functionalities:

(1) Federating OGSA-DAI wrapped data resources to provide virtual, single database view of all heterogeneous data resources which are geographically distributed.

(2) Planning, scheduling and executing distributed queries in parallel by evaluating queries over distributed data resources wrapped by OGSA-DAI.

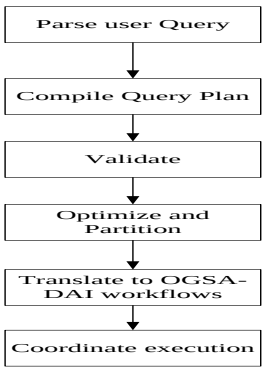

Figure3.The work flow of distributed Query Processor.

5. Conclusion

Effective integration of heterogeneous data sources is the most challenges problem in many domains. In this paper, we present a system for integration of heterogeneous data resources based on Grid technology with OGSA-DAI that provides a uniform access interface and integrate data resources in a distributed environment. Since the system utilizes advances of Grid technologies to federate geographically distributed databases as a single virtual database, users from different locations can access information from different databases servers with no necessary to connect directly to each databases and physical location of the data is transparent to them. Location transparency is supported by the system.

References

[1] T. Landers, RL.Rosenberg , "An overview of MULTIBASE", *In distributed System ,Vole. II: distributed data base systems,W.W.Chu. Artech House, Inc.391-421,1986.*

[2] Y. Tostado, T. Kosaka and S.Date , "Heterogeneous Databases Federation using Grid Technology for Drug Discovery process", *Grid Computing Life Science 3370:43-52,2005*.

[3] A. Wohrerand P. Brezany " Virtualization of Heterogeneous Data Sources for Grid Information Systems", Austria, 2005.

[4] M. Antonioletti, M. Atkinson and R. Baxter," The Design and Implementation of Grid Database Services in OGSA DAI", *Concurrency Computation Practice and Experience* 17(2-4):357-376, 2005.

[5] R.Batra, R. Kaur ," OGSA-DAI_Uses and Applications", Internaltional Journal of Computer Science & Communication,Vole.1,No.1,pp.225-227,January-June 2010.

[6] L.Zambouis, F.Hao and K.Belhajjame," Data access and Integration in the ISPIDER proteomics Grid", Data integration in the Life Sciences, Proceedings 4075:3-18,2006.

[7] E.Jasper, N.Tong, et.al," View Generation and Optimization in the Automated Data Integration Framework", Technical report, AutoMed Project.http://citeseer.ist.psu.edu/jasper03view,2003.

[8]MN. Alpdemir, A.Mukherjee and A.Gounaris,"OGSA-DQP: A service-based distributed query processor for the Grid."http;//www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/114.pdf.

[9] K.karasavvas, M.Antonioletti, "Introduction to OGSA-DAI servieces", Scienticfic applications of Grid Computing 3458:1-12,2004.

[10] M. Antonioletti, A. Krause andR.Baxter,"The design and implementation of grid database services in OGSA-DAI", Concurrentcy Computation Practice and Experience 17(2-4):357-376,2005.

[11] J. Widom,"Data management for XML-Research Directions", IEEE Data Engineering Bulletin, Special Issue on XML,Vole.22,No.3,September,1999.