Grid data management based on dataspace

Ni Tongqian[1,2] Wu Kaigui[1] Liu Peng[2] Liu Yongjin[3]

([1]College of Computer Science, Chongqing University, Chongqing 400044, China)
([2]MILGRID, PLA University of Science and Technology, Nanjing 210007, China)
([3]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: To manipulate the heterogeneous and distributed data better in the data grid, a dataspace management framework for grid data is proposed based on in-depth research on grid technology. Combining technologies in dataspace management, such as data model iDM and query language iTrails, with the grid data access middleware OGSA-DAI, a grid dataspace management prototype system is built, in which tasks like data accessing, abstraction, indexing, services management and answer-query are implemented by the OGSA-DAI workflows. Experimental results show that it is feasible to apply a dataspace management mechanism to the grid environment. Dataspace meets the grid data management needs in that it hides the heterogeneity and distribution of grid data and can adapt to the dynamic characteristics of the grid. The proposed grid dataspace management provides a new method for grid data management.

Key words: grid; dataspace; data model; OGSA-DAI; workflow

The distribution and heterogeneity of the grid data brings great challenges for the sharing of the data. In order to provide facilities for addressing requests over multiple heterogeneous and distributed data sources, it is necessary to provide a uniform data access model and mechanism. Some notable researches have been realized based on grid middleware such as OGSA-DAI[1]. For the schema-heterogeneity problem, a common solution is to expose to users a uniform interface for posting queries on data resources by building semantic mappings.

The shortcoming of the semantic mappings method is that building schema mappings requires the participating of the domain experts and takes a lot of time; also, we cannot share the data until the accurate schema mappings have been created[2]. Franklin et al. [3] introduced a new abstraction method for information management by describing a platform supporting dataspace, the main idea of which is to manage various types of data from different management systems by abstracting data using a uniform model and providing a series of services for these data. Dataspace emphasizes the content of the data and ignores their formats, and it has a pay-as-you-go characteristic for which one can integrate one's data as needed[4]. The same as data in the dataspace, data on grid accord with the heterogeneity of the container, the heterogeneity of format and semantic heterogeneity.

This paper proposes an effective solution to grid data management by introducing dataspace. We provide the architecture of the dataspace management system of grid data, and then analyze its feasibility and key techniques related to its realization.

1 Related Work

At present, technologies relevant to dataspace are becoming flourishing more and more. Dittrich et al. [5] applied dataspace theory to personal information management. They realized a prototype system of personal information management on VLDB2005. Then with deeper research, they expatiated on the framework of personal dataspace and created a personal dataspace system based on a new data model named iDM[6]. Recently, they explored a pay-as-you-go information integration method in dataspace by use of a query language named iTrails[7]. Also, Halevy, one of the innovators of the concept of dataspace, investigated further with his student Dong reference reconciliation, indexing, answering queries and the heterogeneity of dataspace and realized their dataspace management system SEMEX[8]. Fig. 1 is the initial framework of dataspace proposed by Halevy et al. [2].

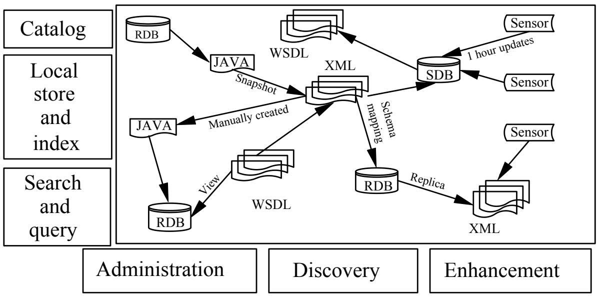

Fig. 1 Components of dataspace management system[2]

Elsayed et al. [9] depicted the idea of combining dataspace and the grid, in which the unfinished system named GridateX aimed at achieving the dataspace management system on the grid. But the relevant achievement has not appeared up to now.

Halevy provided the principles of the dataspace system. For abstraction of data, Dong used a model similar to an ontology while Dittrich used iDM to translate data in dataspace. The two models above have their own methods of data indexing and strategies for answer-query. And the dataspace based on a grid are also in explorative stages in theory.

2 Framework of Grid Dataspace Management System (GDSMS)

2.1 Logical layers

The dataspace management system manages all the data within it and the relationships among them. In order to make the system adapt to the grid environment, we introduce architecture which consists of the following five layers:

Received 2008-04-15.

Biographies: Ni Tongqian (1985—), male, graduate; Wu Kaigui (corresponding author), male, doctor, associate professor, kaiguiwu@cqu.edu.cn.

Citation: Ni Tongqian, Wu Kaigui, Liu Peng, et al. Grid data management based on dataspace[J]. Journal of Southeast University (English Edition), 2008, 24(3): 257 − 259.

1) Grid data source layer: This layer represents all the data managed by the subsystems, such as relational databases, XML databases, file systems, etc. These data can be accessed through specific grid services.

2) Data abstract layer: This layer is responsible for abstracting multifarious data in the dataspace. The problem lies in finding an appropriate data model with which to abstract the grid data. It is very important because it hides the heterogeneity of the data and provides services to the layer above to manage the dataspace easily.

3) Logical services layer: This layer is designed for creating the catalog and index for the data from the data abstract layer, and it also establishes copies of the data to improve the performance of the system. We will provide more details in section 2. 2.

4) Management layer: The main task of this layer is managing the logical services layer, providing control on the indexing and copy creation of the abstract data, providing retrieval of information in a secure environment. It also allows user interaction with the system to achieve a pay-as-you-go pattern as required by a dataspace management system.

5) Application and user interface layer: This layer represents the applications built on top of GDSMS and GUI of the system. Applications may choose either to benefit from the grid services offered by the GDSMS or to access the grid services offered by the data sources themselves.

Regarding aspects of physical deployment, the grid data source layer is located in the local storage of the grid nodes while the data abstract layer and the logical services layer are implemented at the machines where grid services are deployed. Moreover, the management layer is placed on a center grid node, managing data sources previously registered on it, and offers functions via grid services or other forms of interface.

2. 2 Architecture and realization of GDSMS

The foundation of system realization is the open grid services architecture data access and integration (OGSA-DAI) [1]. OGSA-DAI can support a uniform interface for different types of data resources, including relational database, XML and files, etc. And data-centric workflows can implement functions specifically by extending the function of workflow executable units. OGSA-DAI also supports grid-FTP and OGSA-DQP. The former will be of benefit for data transfers in the system. OGSA-DQP is an extension of OGSA-DAI that provides a service-based distributed query processor and supports the evaluation of queries over collections of potentially remote relational data services.

The proposed framework of GDSMS in Fig. 2 consists of two levels: 1) Sub-dataspace; 2) Dataspace management center. They are described as follows.

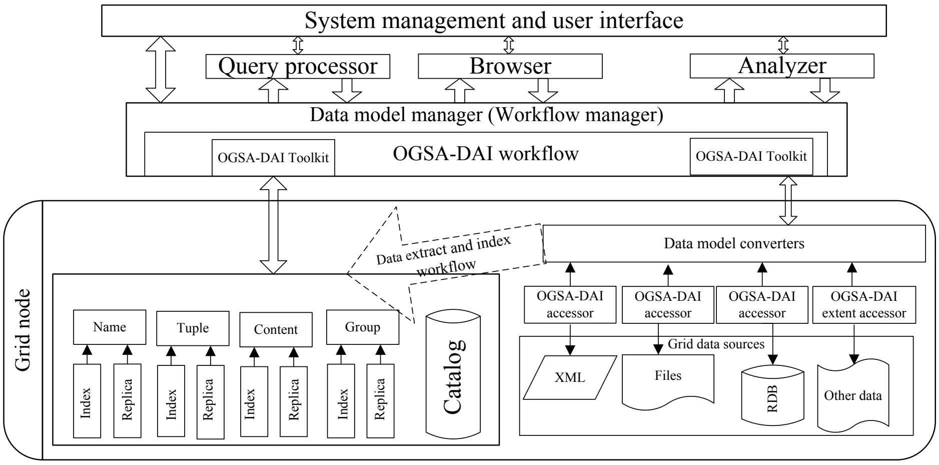

Fig. 2 Architecture and realization of GDSMS

2. 2. 1 Sub-dataspace

Participants of the grid deploy their data sources on their own grid nodes. Once a participant has been registered in the dataspace, the grid node it is located at will become a subset of the dataspace, namely sub-dataspace, which has several components as below:

1) Data sources: Data sources include relational database, XML, files and other types of data. We use data services provided by OGSA-DAI to hide heterogeneities and distribution of data. All the data sources can be exposed by the OGSA-DAI container to ensure consistent access to data and metadata from any data source of a sub-dataspace.

2) Data model converters: Currently there are two different data abstract methods. Dong finished it by extracting information from data sources, which is then represented using a domain model (ontology) , whereas Dittrich extracted data through a data model named iDM[6]. The method proposed by Dong is not adequate enough for the schema-later required by the dataspace principle and makes it hard to integrate information in a pay-as-you-go fashion. So we adopt the iDM model to extract data. An iDM resource view is a 4-tuple (η, τ, χ, γ), where η is a name component, τ is a tuple component, χ is a content component, and γ is a group component. More details of this can be found in Ref. [6]. Data model converters are created on the OGSA-DAI services. We extend the function of the OGSA-DAI service to enable the data abstraction.

3) Catalog and index module: The dataspace management layer will create a catalog and an index automatically once data sources are exposed. The index is produced in the form of an inverted list which is built through indexing for the iDM resource view. The catalog and index modules are deployed on the grid nodes, so that indices can be easily built free of mass transfer loads on the grid.

2.2.2 Dataspace management center

The dataspace management center is responsible for logical control of GDSMS and for providing users as many dataspace services as possible. It implements the following:

1) Data model management: This function is implemented in the grid but it is also implemented at the client of the grid services of the sub-dataspace. We manage the participants of the dataspace by controlling the OGSA-DAI workflows which consist of a number of units called activities. An activity performs a well-defined data-related task such as running an SQL query or performing a data transformation. A workflow can be specified by data model management and supplied to a web service. The web service then performs this workflow in the scope of one web service operation.

2) Query process and integrate further: We will retrieve the data item by accessing the catalog and index modules on remote grid nodes. The query language should be structured or semi-structured as the dataspace requires[4,8]. We adopt the iTrails[7] for the iDM data model. Besides the function of searching in the index, iTrails supports the schema-integration as required. For example, the results of a query statement “// * . tuple. date” will contain the results of “// * . tuple. modified” via statement transfers on trail “// * . tuple. date→// * . tuple. modified”. Details are the same as in Ref. [7].

3) Other functions: They include browsing the dataspace, sub-dataspace register, security and authorization etc., which are all realized in the dataspace management center and some of these functions can be based on the basic functions of the Globus Toolkit.

3 Conclusion and Future Work

The contributions of this work include doing some research in grid data management based on dataspace and proposing the architecture of a GDSMS. This work can be seen as a prototype of a grid dataspace management system and is an important step towards the realization of our prospective goals. As part of future work, we plan to explore some of the key problems in more depth. There are a lot of subjects for research challenges. We will explore both grid theories and GDSMS implementation based on this work to research new methods for grid data management.

References

[1] Karasavvas Kostas, Atkinson Malcolm, Hume Ally. OGSA-DAI 3.0 user doc and related specifications [EB/OL]. (2007-07-25) [2008-04-02]. http://www. ogsadai. org. uk/documentation/ogsadai3. 0/.

[2] Halevy A, Rajaraman A, Ordille J. Data integration: the teenage years[C]//*Proc of the* 32*nd International Conference on Very Large Data Bases*. Seoul, Korea, 2006: 9 − 16.

[3] Franklin M, Halevy A, Maier D. From databases to dataspaces: a new abstraction for information management[J]. *ACM SIGMOD Record*, 2005, **34**(4): 27 − 33.

[4] Halevy A, Franklin M, Maier D. Principles of dataspace system[C]//*Proc of the Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Chicago, IL, USA, 2006: 1 − 9.

[5] Dittrich J P, Salles M A V, Kossmann D, et al. iMeMex: escapes from the personal information jungle (Demo) [C]//*Proc of the* 31*st International Conference on Very Large Data Bases*. Trondheim, Norway, 2005: 1306 − 1309.

[6] Dittrich J P, Salles M A V. iDM: a unified and versatile data model for personal dataspace management[C]//*Proc of the* 32*nd International Conference on Very Large Data Bases*. Seoul, Korea, 2006: 367 − 378.

[7] Marcos Antonio Vaz Salles, Dittrich J P, Karakashian S K. iTrails: pay-as-you-go information integration in dataspaces [C]//*Proc of the* 33*rd International Conference on Very Large Data Bases*. Vienna, Austria, 2007: 663 − 674.

[8] Dong X, Halevy Alon Y. Indexing dataspaces[C]//*Proc of the* 2007 *ACM SIGMOD International Conference on Management of Data*. Beijing, China, 2007: 43 − 54.

[9] Elsayed Ibrahim, Brezany Peter. Towards realization of dataspaces[C]//*Proc of the* 17*th International Conference on Database and Expert Systems Applications*. Washington, DC, USA, 2006: 266 − 272.

基于数据空间的网格数据管理

倪彤前[1,2]　吴开贵[1]　刘　鹏[2]　刘永金[3]

([1] 重庆大学计算机学院,重庆 400044)
([2] 解放军理工大学网格研究中心,南京 210007)
([3] 东南大学计算机科学与工程学院,南京 210096)

摘要:为了更好地管理和应用数据网格中大量分布异构的数据,在对网格技术发展现状进行深入研究基础上,提出基于数据空间概念的网格数据的管理架构.在此基础上,实现了一个网格数据空间管理原型系统,系统中将现有的一些数据空间技术如数据模型 iDM、查询语言 iTrails 等与网格数据访问中间件 OGSA-DAI 相结合,使用 OGSA-DAI 工作流来完成数据空间管理系统的数据访问、抽取、数据索引、服务管理和查询回复等一系列工作.实验表明数据空间管理机制在网格环境下是可行的,数据空间管理系统屏蔽了网格数据的分布性和异构性,且能够适应网格数据动态特性,因此满足了对网格数据的管理要求.所提出的网格数据空间架构为网格数据管理提出了新的方法.

关键词:网格;数据空间;数据模型;OGSA-DAI;工作流

中图分类号:TP393.07