A Grid Environment for Data Integration of Scientific Databases

Hideo Matsuda
Department of Bioinformatic Engineering,
Graduate School of Information Science and Technology
Osaka University, Toyonaka, Japan.
matsuda@ist.osaka-u.ac.jp

Abstract

Effective integration of heterogeneous data sources has been studied as the most pressing challenge in various fields; such as, high energy physics, astronomy, and life sciences. In this talk, we present a data integration system by using Globus Toolkit with OGSA-DAI. For associating related data among many databases, we have introduced metadata based on their domain ontologies. Using the system one can make a database access flow for describing a set of queries as a workflow, and can query across the databases without aware of their locations and schemas.

1. Introduction

With the rapid progress of various technologies, there is a growing demand for analyzing inter-disciplinary areas with an integrated view, such as multiscale and multiphysics computations and data integrations. To tackle this problem, many Grid projects have been conducted (e.g., BioGrid [1] and NAREGI [2] in Japan).

In this talk, we focus on the data integration from a large number of scientific databases by employing data grid technology. In these databases, their schema and data description are much diverged since they are based on different data domains. In order to integrate databases across those domains, we introduced metadata for describing semantic relationships among their entities [3]. This connected network of databases makes use of the grid technology for delivering integrated searches of the databases.

2. Overview of Data Grid Environment

A Data Grid environment was developed for the data integration from a large number of databases (see Fig. 1). As a target data-source, we took life science databases since there are more than 700 databases and many of them are available on the web [4].

To cope with these issues, we construct metadata for associating all the related data among the databases and develop a dataflow engine for performing multiple queries in pipelining flows. It is composed of a database access management tool for constructing database access flows across those databases, query engines for performing simultaneous queries to the DBs, a distributed file system for storing query results in distributed storages, and a metadata construction system for constructing ontology-based metadata to manipulate heterogeneous data in various scientific fields.

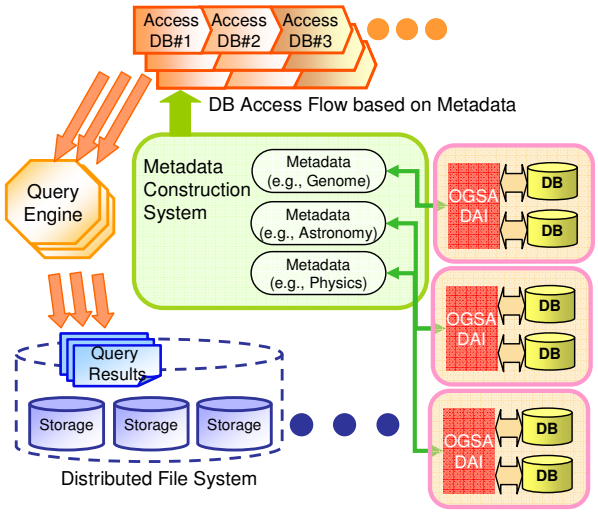

Figure 1. Overview of DB integration system.

The system was developed by using Globus Toolkit 4.0.1 with OGSA-DAI [5], and Gfarm [6] as a distributed file system. On this system, we have imported various life science DBs (e.g., UniProt, NCBI PubChem, NLM Medical Encyclopedia, etc.). Using the system one can make a database access flow for describing a set of queries as a workflow across the databases.

References

[1] http://www.biogrid.jp/

[2] http://www.naregi.org/

[3] Y. Tohsato, T. Kosaka, S. Date, S. Shimojo, and H. Matsuda, "Heterogeneous Database Federation using Grid Technology for Drug Discovery Process", *Lecture Notes in Bioinformatics*, vol.3370, pp.43-52, 2005.

[4] M.Y. Galperin, "The Molecular Biology Database Collection: 2005 Update", *Nucleic Acids Research*, vol.33, DB issue, pp.D5-D24, 2005.

[5] http://www.ogsadai.org/

[6] http://datafarm.apgrid.org/