Managing Inconsistencies in Data Grid Environments: A Practical Approach

Ejaz Ahmed, King Fahd University of Petroleum and Minerals, Saudi Arabia and University of Bedfordshire, UK
Nik Bessis, University of Bedfordshire, UK
Peter Norrington, University of Bedfordshire, UK
Yong Yue, University of Bedfordshire, UK

ABSTRACT

Much work has been done in the area of data access and integration using various data mapping, matching, and loading techniques. One of the main concerns when integrating data from heterogeneous data sources is data redundancy. The concern is mainly due to the different business contexts and purposes from which the data systems were originally built. A common process for accessing data from integrated databases involves the use of each data source’s own catalogue or metadata schema. In this article, the authors take the view that there is a greater chance of data inconsistencies, such as data redundancies when integrating them within a grid environment as compared to traditional distributed paradigms. The importance of improving the data search and matching process is briefly discussed, and a partial service oriented generic strategy is adopted to consolidate distinct catalogue schemas of federated databases to access information seamlessly. To this end, a proposed matching strategy between structure objects and data values across federated databases in a grid environment is presented.

Keywords: Data Integration, Grid, Metadata, Pattern Matching, Plug-In Relations, Staging DBMS

INTRODUCTION

The role of traditional data integration (Bessis et al., 2007; Austin et al., 2006) and loading techniques and methods requires more attention when functioning within the grid environment. One reason is the higher chance for conflict between metadata and structures when integrating data (multi-vendor relational, object-relational DBMSs) within a grid environment since data sources have been originally produced for a purpose other than their integration (Bessis et al., 2009). This is our main motivation for undertaking this research.

The grid is an emerging infrastructure that supports the discovery, access and use of distributed computational resources (Alpdemir et al., 2003), including data integration in the de-facto
OGSA-DAI (Open Grid Service Architecture – Data Access Integration) specification framework. Significant effort has gone into defining requirements, protocols and implementing the OGSA-DAI specification framework as the means for users to develop relevant data grids to conveniently control the sharing, accessing and management of large amounts of distributed data in Grid environments (Antonioletti et al., 2005; Atkinson et al., 2003). Ideally, OGSA-DAI as a data integration specification aims to allow users to specify ‘what’ information is needed without having to provide detailed instructions on ‘how’ or ‘from where’ to obtain the information (Reinoso Castillo et al., 2004). However, mapping multiple physical replicas to one single logical file increases data redundancy (Jacob et al., 2005). Yin et al. (2009) also explain that running join queries over a data grid environment requires appropriate strategies as decomposing and disseminating the query to as many as possible sources, as processing the user’s query in parallel will also bring in overhead in repetitive computing, redundant data transmission, and result merging.

In this article, our main contribution is the discussion and proposal of a schema matching exercise for identifying two objects that semantically relate them, while we refer to mapping as the transformations between the objects concerned. That is to say when data is accessed it will attempt to virtualize and/or transfer them within the target source. It is anticipated that if there is more than one accessing object in the same or dispersed data source(s), like DBMSs, then it is necessary to define matching patterns to access the correct objects concerned with their properties. In our present discussion, a schema is treated as a set of elements connected by some structure. A target database schema requires access to certain data from many objects (or elements of a schema) of various distinct schemas of databases with the help of common sources’ metadata.

With this in mind, the main intention of our article is multi-fold: firstly, to offer a brief review of data matching exercises; secondly, to present our proposed strategy by discussing our data matching and mapping framework including a linguistic matching approach; thirdly, to discuss in full our methodological approach for a grid based matching process using a number of developed metadata algorithms. We finally conclude by presenting some pilot experimental results and further work.

LITERATURE REVIEW

The section presents an overview of and a comparison between matching approaches. Rahm and Bernstein (2001) developed taxonomy of schema matching approaches which for some remains the significant contribution in this field. The taxonomy consists of two major branches (Figure 1): (a) individual matcher approaches and (b) combining matchers. Regarding combining matchers, hybrid matchers integrate matching criteria prior to mapping, whereas composite matchers integrate the results of individual matchers post-mapping. The description of individual matchers is from (Rahm & Bernstein, ibid).

- **Instance vs. schema**: matching approaches can consider instance data (i.e., data contents) or only schema-level information.
- **Element vs. structure matching**: match can be performed for individual schema elements, such as attributes, or for combinations of elements, such as complex schema structures.
- **Language vs. constraint**: a matcher can use a linguistic based approach (e.g., based on names and textual descriptions of schema elements) or a constraint-based approach (e.g., based on keys and relationships).
- **Matching cardinality**: the overall match result may relate one or more elements of one schema to one or more elements of the other, yielding four cases: 1:1, 1:n, n:1, n:m. In addition, each mapping element may interrelated one or more elements of the two schemas. Furthermore, there may be different match cardinalities at the instance level.
Unified Data Access/Query over Integrated Data-views for Decision Making in Geographic Information Systems
www.igi-global.com/chapter/unified-data-access-query-over/19349?camid=4v1a

Li Mao, De Yu Qi, Wei Wei Lin, Bo Liu and Ye Da Li (2016). International Journal of Grid and High Performance Computing (pp. 43-57).
www.igi-global.com/article/an-energy-efficient-resource-scheduling-algorithm-for-cloud-computing-based-on-resource-equivalence-optimization/153969?camid=4v1a