Chapter V
Placement and Scheduling over Grid Warehouses

Rogério Luís de Carvalho Costa
University of Coimbra, Portugal

Pedro Furtado
University of Coimbra, Portugal

ABSTRACT

The computational grid offers services for efficiently scheduling jobs on the grid, but for grid-enabled applications where data handling is a most relevant part, the data grid kicks in. It typically builds on the concept of files, sites and file transfers between sites. These use a data transfer service, plus a replica manager to keep track of where replicas are located. The authors consider a multi-site, grid-aware data warehouse, which is a large distributed repository sharing a schema and data concerning scientific or business domains. Differently from typical grid scenarios, the data warehouse is not simply a set of files and accesses to individual files. It is a single distributed schema and both localized and distributed computations must be managed over that schema. Given this difference, it is important to study approaches for placement and computation over the grid data warehouse and this is our contribution in this book chapter.

INTRODUCTION

The Grid is an infrastructure that coordinates and provides transparent access to distributed and heterogeneous resources for geographically distributed users. One of its main usages is to enable the implementation of Virtual Organizations (Foster, 2001).

Virtual Organizations (VO) are composed by (members of) distinct real organizations or institutions, which cooperatively participate in the Grid, sharing resources and using the shared resources in order to solve common problems or to achieve common objectives. In such multi-institutional grid it is somewhat frequent that individual domain administrators impose some kind
of usage constraints on local domains’ resources to remote users (Foster, 2001; Schopf & Nitzberg, 2002). Business historical information, molecular simulations and measurements of sensor data are examples of some current and future applications that run on VO and which can produce large amounts of distributed data.

In Data Grids (Chervenak et al., 2001; Krauter et al., 2002), the Grid software is used to manage data storage, transparent data access or the execution of data intensive jobs. Flat files are frequently used to store data. Grid-based services are used to execute efficient data movement (e.g. GridFTP (Allcock et al., 2005)) and to locate database replicas (e.g. Globus’ Replica Location Service – RLS (Chervenak et al., 2004)).

On the other hand, Data Warehouses (DWs) are huge mostly read-only databases used primarily for decision support purposes. They are target of periodic data loads as part of ETL (Extraction Transform and Load) processes: data from several OLTP (Online Transaction Processing) systems’ is transformed and loaded into the DW, which traditionally resides at a single site (Chaudhuri & Dayal, 1997) (parallel hardware is commonly used to improve performance).

Therefore, the Grid-based Data Warehouse can be used as a tool to support several processes in global Virtual Organizations, including decision making and knowledge discovery. It enables users from different geographical locations (and possibly from different real organizations), which are participating in the VO, to share and query shared data transparently and efficiently. Hence, given a data warehouse schema, grid-enabled DWs have queries, data mining analysis and computations of all kinds happening over large quantities of data. For instance, groups of scientific organizations may provide public multi-site data repositories and computing platforms for simulations and analysis, as it is already beginning to happen with biological data such as protein conformation studies.

Implementing efficient data placement and query scheduling strategies is one of the challenges in such environment. For instance, placing all the data at a single site may not be the most interesting choice. In Grids, distributed resources from multiple sites are available to be cooperatively used to perform tasks. Using such distributed resources to implement a Grid-based DW may lead to a high performance environment at a reduced monetary cost, when compared to placing all the DW at a single grid site (especially due to reduced requirements in hardware and network links at individual sites). Hence, data and query distribution across several sites should be considered in order to obtain a high performance grid-based DW.

Besides that, it is important to notice that warehouse users may query the database according to some kind of access pattern (Deshpande et al., 1998). For instance, users from distinct sites may be more interested in doing drill-down and drill-up operations in distinct geographically-related data (each user accessing data that is somewhat related to the region where the user is placed). As doing data movement across different grid sites may be time consuming, the existent users’ interests must also be considered when doing placement across distinct sites. Furthermore, as workloads and resource characteristics may change over time, the system should also have some kind of dynamic data placement (or replication) strategy, which is also useful when users’ access patterns are not known in advance (or change over time).

But the specialized data placement strategy and the abovementioned aspects lead to the need of a specialized query scheduling strategy, which may consider the existence of partitioned and replicated tables when scheduling query execution among distinct sites.

In this chapter we study data placement and query scheduling in the Grid-based DW. We present a distributed static placement strategy and propose an efficient dynamic placement policy which does replica management considering existent resource constraints. We also propose a query scheduling strategy that considers both lo-