A Survey of Parallel and Distributed Data Warehouses

Pedro Furtado, Universidade Coimbra, Portugal

ABSTRACT

Data Warehouses are a crucial technology for current competitive organizations in the globalized world. Size, speed and distributed operation are major challenges concerning those systems. Many data warehouses have huge sizes and the requirement that queries be processed quickly and efficiently, so parallel solutions are deployed to render the necessary efficiency. Distributed operation, on the other hand, concerns global commercial and scientific organizations that need to share their data in a coherent distributed data warehouse. In this article we review the major concepts, systems and research results behind parallel and distributed data warehouses. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: Data Warehousing; Database; Distributed Databases; Parallel Processing Systems; SQL

INTRODUCTION

Decision support systems are important tools in the hands of today’s competitive and knowledgeable organizations, and data warehouses (DW) are at the core of such systems. They store huge detailed and summarized historical data for decision makers to generate queries, make reports and perform analysis and mining that are the basis for their decisions and deeper knowledge. Users also need fast response times on complex queries in data warehousing, OLAP and data mining operations. Two major forces have contributed to the importance of parallel and distributed data warehousing: On one hand, the fact that data warehouses can be extremely large and highly resource demanding, while queries and analyses must be answered within acceptable time limits has led to a series of specialized techniques that were developed specifically for them, including view and cube materialization (Rousopoulos 1998), specialized indexing structures (O’Neil and Graefe 1995) and implementations on parallel systems, which we review along this article. While all these specialized techniques and structures play an important role in the performing data warehouse, we focus on parallel systems in particular, which can provide top performance and scalability. Parallel processing answers satisfactorily the need to handle huge data sets efficiently, in both query processing and other
PARALLEL ARCHITECTURES FOR DATA WAREHOUSING

Due to their high-demand on storage and performance, large DWs frequently reside within some sort of parallel system. In this section we review different base architectures that can be used to store and process the parallel data.

There is a whole range of architectures for parallelization, from shared-nothing to shared-disk and hybrid ones, as current state-of-the-art servers come with multiple processors. There are different nomenclatures for the basic models by which a parallel system can be designed, and the details of each model vary as well. Consider three basic elements in a parallel system: the processing unit (PU), the storage device (S) and memory (M). The simplest taxonomy defines three models, as described in (DeWitt and Gray 1992):

- **Shared Memory (SM):** the shared memory or shared everything architecture, illustrated in Figure 1, is a system where all existing processors share a global memory address space as well as peripheral devices. Only one DBMS is present, which can be executed in multiple processes or threads, in order to utilize all processors;

- **Shared Nothing (SN):** the shared nothing architecture, illustrated in Figure 2, is composed of multiple autonomous Processing Nodes (PN), each owning its own persistent storage devices and running separate copies of the Database Management System (DBMS). Communication between the PNs is done by message passing through the network. A PN can be composed of one or more processors and/or storage devices.

- **Shared Disk (SD):** the shared disk architecture, illustrated in Figure 3, is characterized by possessing multiple loosely coupled PNs, similar to SN. However, in this case, the architecture possesses a global disk subsystem that is accessible to the DBMS of any PN.
Maintaining Dimension’s History in Data Warehouses Effectively
[www.igi-global.com/article/maintaining-dimensions-history-in-data-warehouses-effectively/228937?camid=4v1a](www.igi-global.com/article/maintaining-dimensions-history-in-data-warehouses-effectively/228937?camid=4v1a)

U.S. Federal Data Mining Programs in the Context of the War on Terror: The Congress, Courts, and Concerns for Privacy Protections
[www.igi-global.com/chapter/federal-data-mining-programs-context/29149?camid=4v1a](www.igi-global.com/chapter/federal-data-mining-programs-context/29149?camid=4v1a)

[www.igi-global.com/article/a-framework-for-evaluating-design-methodologies-for-big-data-warehouses/198972?camid=4v1a](www.igi-global.com/article/a-framework-for-evaluating-design-methodologies-for-big-data-warehouses/198972?camid=4v1a)