Chapter 3.22
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.

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 concerns such as loading or creation of auxiliary structures; On the other hand, the evolution of the data warehouse concept from a centralized local repository into a broader context of sharing and analyzing data in an internet-connected world has given birth to distributed approaches and systems.

In this chapter we review important research and trends on these parallel and distributed approaches. Our approach is to introduce and illustrate the major issues first, and then to review some of the most relevant systems and research results on the field. Our first discussion is on parallel architectures, the physical infrastructure over which to store and process the data, with crucial implications on performance and scalability of the solutions. With this in mind, we then discuss types of parallelism and architectural issues in parallel database management systems. Then we discuss partitioning and allocation, one of the most fundamental enablers of intra-query horizontal parallelism. After discussing architecture and partitioning, we then turn our attention to parallel processing and optimization, including an illustration on how to process in horizontal intra-query parallelism. After reviewing the architectural, partitioning and processing basics of parallel data warehousing, we devote a section to the discussion of systems and research results on the subject of parallel data warehouses and another one on distributed data warehouses. Distributed data warehouse systems are a most relevant subject, since WAN-connected geographically distributed organizations share both data and analysis, and networking technology currently enables long distance collaboration.

Parallel and distributed data warehousing is an exciting field, and research in these issues is far from being exhausted. In a few words, autonomy, scalability, ubiquity and application contexts are some of the most fundamental issues that will certainly deserve a lot of attention in the future. We end the article with conclusions and a brief discussion on these future trends.

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;