RCUBE: Parallel Multi-Dimensional ROLAP Indexing

Frank Dehne, Carleton University, Canada
Todd Eavis, Concordia University, Canada
Andrew Rau-Chaplin, Dalhousie University, Canada

ABSTRACT

This article addresses the query performance issue for Relational OLAP (ROLAP) datacubes. We present RCUBE, a distributed multidimensional ROLAP indexing scheme which is practical to implement, requires only a small communication volume, and is fully adapted to distributed disks. Our solution is efficient for spatial searches in high dimensions and scalable in terms of data sizes, dimensions, and number of processors. Our method is also incrementally maintainable. Using "surrogate" group-bys, it allows for the efficient processing of arbitrary OLAP queries on partial cubes, where not all of the group-bys have been materialized. Our experiments with RCUBE show that the ROLAP advantage of better scalability, in comparison to MOLAP, can be maintained while providing a fast and flexible index for OLAP queries.

Keywords: algorithmic complexity; efficiency; data warehousing and OLAP; datacube; high performance computing; parallel ROLAP indexing; scalability issues; scalability to large databases

INTRODUCTION

Online Analytical Processing (OLAP) has become a fundamental component of contemporary decision support systems. Gray, Bosworth, Layman, and Pirahesh (1995) introduced the datacube, a relational operator used to compute summary views of data that can, in turn, significantly enhance the response time of core OLAP operations such as roll-up, drill down, and slice and dice. Typically constructed on top of relational data warehouses, these summary views (called group-bys) are formed by aggregating values across attribute combinations. For a d-dimensional input set R, there are \(2^d\) possible group-bys. Figure 1 illustrates a datacube as well as a lattice, which is often used to represent the inherent relationships between group-bys (Harinarayan, Rajaraman, & Ullman, 1996).

There are two standard datacube representations: MOLAP (multidimensional array) and ROLAP (set of relational tables). The array-based model, MOLAP (Multidimensional OLAP), has the advantage that native arrays provide an immediate form of indexing for cube queries. Interesting MOLAP based systems have been described and implemented in both the sequential and parallel settings (e.g.,
Goil & Choudhary, 1997). However, there is some evidence that MOLAP based systems may encounter significant scalability problems (Pendse & Creeth, 2002). For example, high-dimension datacubes represent extremely sparse spaces that are not easily adapted to the MOLAP paradigm. Hybrid indexing schemes are often used, significantly diminishing the power of the model. Moreover, since MOLAP needs to be integrated with standard relational databases, middleware of some form must be employed to handle the conversion between relational and array-based data representations. The relational model, ROLAP (Relational OLAP), does not suffer from such restrictions. Its summary records are stored directly in standard relational tables without any need for data conversion. Its table-based data representation does not pose scalability problems. Yet, many current commercial systems use the MOLAP approach (Pendse & Creeth, 2002). The main reason, as outlined in Pendse and Creeth, is the indexing problem for the fast execution of OLAP queries. The problem for ROLAP is that it does not provide an immediate and fast index for OLAP queries. Many vendors have chosen to sacrifice scalability for performance. However, this path is becoming increasingly unsustainable. As outlined in the 2005 Winter Report (Winter Corporation, 2005), the size of data warehouses grew exponentially during recent years. More precisely, between 2001 and 2005, the average size of data warehouses grew by 243%, and the size of the largest data warehouses grew by an astounding 578% (Winter Corporation, 2005). Hence, there is an urgent need for scalable (i.e., ROLAP) and high-performance datacube indexing methods.

This article addresses the query performance issue for ROLAP and proposes a novel distributed multidimensional ROLAP indexing scheme. We show that the ROLAP advantage of high scalability can be maintained, while at the same time providing a fast index for OLAP queries. We present RCUBE, a distributed indexing scheme which is a combination of packed R-trees with distributed disk striping and Hilbert curve based data ordering. Our method requires only very low communication volume between processors and works in
A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents
www.igi-global.com/article/data-mining-based-olap-aggregation/1772?camid=4v1a

A TOPSIS Data Mining Demonstration and Application to Credit Scoring
www.igi-global.com/article/topsis-data-mining-demonstration-application/1768?camid=4v1a