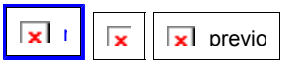


Next: Introduction

High Performance Multidimensional Analysis and Data Mining

Sanjay Goil and Alok Choudhary
Center for Parallel and Distributed Computing,
Department of Electrical & Computer Engineering,
Northwestern University,
Technological Institute, 2145 Sheridan Road, Evanston, IL-60208
{sgoil,choudhar}@ece.nwu.edu

Abstract:

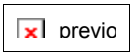
Summary information from data in large databases is used to answer queries in On-Line Analytical Processing (OLAP) systems and to build decision support systems over them. The *Data Cube* is used to calculate and store summary information on a variety of dimensions, which is computed only partially if the number of dimensions is large. Queries posed on such systems are quite complex and require different views of data. These may either be answered from a materialized cube in the data cube or calculated on the fly. Further, data mining for associations can be performed on the data cube. Analytical models need to capture the multidimensionality of the underlying data, a task for which multidimensional databases are well suited. Also, they are amenable to parallelism, which is necessary to deal with large (and still growing) data sets. Multidimensional databases store data in multidimensional structure on which analytical operations are performed. A challenge for these systems is how to handle large data sets in a large number of dimensions. These techniques are also applicable to scientific and statistical databases (SSDB) which employ large multidimensional databases and dimensional operations over them.

In this paper we present *(1)* A parallel infrastructure for OLAP multidimensional databases integrated with association rule mining. *(2)* Introduce Bit-Encoded Sparse Structure (BESS) for sparse data storage in *chunks*. *(3)* Scheduling optimizations for parallel computation of *complete* and *partial* data cubes. *(4)* Implementation of a large scale multidimensional database engine suitable for dimensional analysis used in OLAP and SSDB for (a) large number of dimensions (20-30) (b) large data sets (10s of Gigabyte)

Our implementation on the IBM SP-2 can handle large data sets and a large number of dimensions by using disk I/O. Results are presented showing its performance and scalability.

Proceedings of the 1998 ACM/IEEE SC98 Conference (SC'98)
0-8186-8707-X/98 $17.00 © 1998 IEEE





Next: Introduction

Sanjay Goil
Fri Aug 7 14:58:04 CDT 1998