DMS: A Parallel Data Mining Server

Felicity George
High Performance Research Center
Tandem Business Unit, Compaq
Wallace View, Hillfoots Road,
Stirling Fk9 5PY
Scotland
email: Felicity.George@compaq.com

Tandem's Data Mining Server (DMS) is a parallel data engine designed to enable data mining tools to store, access and analyse high volumes of data very efficiently. In contrast to traditional database management systems, data structures are optimized for analysis and pattern recognition, rather than for accessing individual rows. The data stored in DMS tables are encoded automatically, so the required disk space is typically 3 to 5 times less than the raw data size. This approach not only minimizes disk space and disk access time, it enables the majority of processing to be done in memory.

Data can be imported into DMS's internal format from ORACLE tables or ASCII files or through the user's own import modules. Functionality available in DMS includes:

Aggregations: including operations such as *min, max, average, variance, sum* and *count* on *n* dimensions of data.

Transformations: new data columns are created through operations on existing columns.

Zoom-in functions: Subsets of data can be identified using simple or complex predicates, then can be used to zoom in on interesting areas of data.

Hierarchies: Data in a star or snowflake schema in a traditional database can be loaded into DMS with its structure maintained.

Continuous Variables: Functions are provided to 'bin' continuous or high cardinality variables into discrete groups.

DMS has been designed primarily for speed and scalability; in order to process large amounts of data rapidly, simplicity of design has also been crucial. DMS can currently perform operations like building a histogram of an attribute at several million rows per second per CPU. Other features of the system are that it is easy to extend to add further primitives as required, and that, whilst the server is targeted at data mining, the primitives are general enough that other more traditional decision support tools could also make use of the server with relatively little effort. It is also highly scalable.

The high data processing speeds attained by DMS are due to several factors; primarily

- Effective parallelisation of the data
- Efficient encoding of the data.
- Simple and optimised algorithms are used to manipulate the data.

More information on DMS is available from the author, and also from Martin.Hahn@compaq.com, Wouter.Senf@compaq.com or Iain.Robertson@compaq.com

Proceedings of the 24th VLDB Conference New York, USA, 1998