HEAVEN
A Hierarchical Storage and Archive Environment for Multidimensional Array Database Management Systems

Bernd Reiner, Karl Hahn

FORWISS (Bavarian Research Center for Knowledge Based Systems)
Technical University Munich
Boltzmannstr. 3, D-85747 Garching b. München, Germany
{reiner, hahnk}@forwiss.tu-muenchen.de

Abstract. The intention of this paper is to present HEAVEN, a solution of intelligent management of large-scale datasets held on tertiary storage systems. We introduce the common state of the art technique storage and retrieval of large spatio-temporal array data in the *High Performance Computing* (HPC) area. An identified major bottleneck today is fast and efficient access to and evaluation of high performance computing results. We address the necessity of developing techniques for efficient retrieval of requested subsets of large datasets from mass storage devices. Furthermore, we show the benefit of managing large spatio-temporal data sets, e.g. generated by simulations of climate models, with *Database Management Systems* (DMBS). This means DBMS need a smart connection to tertiary storage systems with optimized access strategies. HEAVEN is based on the multidimensional array DBMS RasDaMan.

1 Introduction

Large-scale scientific experiments often generate large amounts of multidimensional data sets. Data volume may reach hundreds of terabytes (up to petabytes). Typically, these data sets are stored as files permanently in an archival mass storage system on up to thousands of magnetic tapes. The access times and/or transfer times of these kinds of tertiary storage devices, even if robotically controlled, are relatively slow. Nevertheless, tertiary storage systems are currently the common state of the art storing such large volumes of data. Concerning data access in HPC area the main disadvantages are high access latency compared to hard disk devices and to have no direct access. A major bottleneck for scientific application is the missing possibility of accessing specific subsets of data. If only a subset of such a large data set is required, the whole file must be transferred from tertiary storage media. Taking into account the time required to load, search, read, rewind and unload several cartridges, it can take many hours/days to retrieve a subset of interest from a large data set. Entire files must be loaded from the magnetic tape, even if only a subset of the file is needed for a further processing. The processing with data across a multitude of data sets, for example, time slices is hard to support. Evaluation of search criteria requires network transfer of each required data set, implying sometimes a prohibitively immense amount of data to be shipped. Hence, many interesting and important evaluations currently are impossible. Another disadvantage is that access to data

sets is done on an inadequate semantic level. Applications accessing HPC data have to deal with directories, file names, and data formats instead of accessing multidimensional data in terms of simulation space or time interval. Examples of large-scale HPC data are climate-modeling simulations, cosmological experiments and atmospheric data transmitted by satellites. Such natural phenomena can be modeled as spatio-temporal array data of some specific dimensionality. Their common characteristic is that a huge amount of *Multidimensional Discrete Data* (MDD) has to be stored. For overcoming the above mentioned shortcomings and for providing flexible data management of spatio-temporal data we implemented HEAVEN (***H****ierarchical Storag****e*** *and* ***A****rchive En****v****ironment for Multidim****e****nsional Array Database Ma****n****agement Systems*).

2 HEAVEN System Architecture

HEAVEN combines the advantages of storing big amounts of data and the realization of efficient data access and management with DBMS. This means the DBMS must be extended with easy to use functionalities to automatically store and retrieve data to/from tertiary storage systems without user interaction. We implemented such intelligent concepts and integrated it into the kernel of the first commercial multidimensional array DBMS RasDaMan (*Raster Data Management*). RasDaMan is specially designed for generic multidimensional array data of arbitrary size and dimensionality.

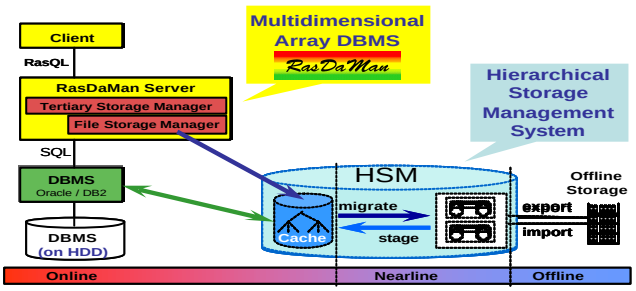

Fig. 1: HEAVEN system architecture

Fig. 1 depicts the architecture of the extended RasDaMan system (client/server architecture with the conventional DBMS) with tertiary storage connection. We realize the tertiary storage support by connecting a *Hierarchical Storage Management* (HSM) System with RasDaMan. Such HSM-Systems have been developed to manage tertiary storage archive systems and can handle thousands of tertiary storage media. The virtual file system of HSM-Systems is separated into a limited cache on which the user works (load or store his data) and a tertiary storage system with robot controlled tape libraries. The HSM-System automatically migrates or stages data to or from the tertiary storage media, if necessary. For realizing the retrieval of subsets of large data sets (MDDs) RasDaMan stores MDDs subdivided into sub-arrays (called tiles). Detailed information about tiling can be found in [1, 3, 5]. Tiles are in RasDaMan the smallest unit of data access. The size of tiles (32 KByte to 640 KByte) is optimized for main memory and hard disk access. Those tile sizes are much too small for data sets held on tertiary storage media [2]. It is necessary to choose different granularities for hard disk access and tape

access, because they differ significantly in their access characteristics (random vs. sequential access). A promising idea is to introduce additional data granularity as provided by the new developed Super-Tile concept. The main goal of the Super-Tile concept is a smart combination of several small MDD tiles to one Super-Tile for minimizing tertiary storage access costs. Smart means, exploiting the good transfer rate of tertiary storage devices, and to take advantage of other concepts like clustering of data. Super-Tiles are the access (import/export) granularity of MDD on tertiary storage media. Extensive tests have shown that a Super-Tile size of about 150 MByte shows good performance characteristics in most cases. The retrieval of data stored on hard disk or on tertiary storage media is transparent for the user. Only the access time is higher if data stored on tertiary storage media. Three further strategies for reducing tertiary storage access time are clustering, query scheduling and caching. Please find more information in [4].

3 Demonstration

We will demonstrate HEAVEN using the visual front-end RView, to interactively submit RasQL (*RasDaMan query language*) queries and display result sets containing 1-D to 4-D data. For demonstrating tertiary storage access we will use our own developed HSM-System with a connected SCSI DDS-4 tape drive. Demonstration will start by showing sample retrieval, thereby introducing basic RasQL concepts. Queries will encompass both, search and array manipulation operations. We will show typical access cases like retrieval of subsets and data access across a multitude of objects. Furthermore, performance comparison of DBMS cache area access, HSM cache access, HSM tape access and traditional access will be presented. Next, selected queries will serve to demonstrate the effect of several optimization concepts, like inter and intra Super-Tile clustering. Also performance improvements regarding query scheduling and caching algorithms will be shown. This allows discussing, how they contribute to overall performance. Finally, implications of physical data organization within RasDaMan (tiling strategies) and on tertiary storage media (Super-Tile concept) will be presented. Queries such as sub-cube extraction and cuts along different space axes indicate strengths and weaknesses of particular tiling and Super-Tile schemata. Various tiling strategies are offered as a database tuning possibility similar to indexes for optimal query performance.

References

1. Chen L. T., Drach R., Keating M., Louis S., Rotem D., Shoshani A.: Efficient organization and access of multi-dimensional datasets on tertiary storage, Information Systems, vol. 20, no. 2, p. 155-183, 1995
2. Chen L. T., Rotem D., Shoshani A., Drach R.: Optimizing Tertiary Storage Organization and Access for Spatio-Temporal Datasets, NASA Goddard Conf. on MSS, 1995
3. Furtado P. A., Baumann P.: Storage of Multidimensional Arrays Based on Arbitrary Tiling, Proc. of the ICDE, p. 480-489, 1999
4. Reiner B., Hahn K., Höfling G.: Hierarchical Storage Support and Management for Large-Scale Multidimensional Array Database Management Systems, DEXA conference, 2002
5. Sarawagi S., Stonebraker M.: Efficient Organization of Large Multidimensional Arrays, Proc. of Int. Conf. On Data Engineering, volume 10, p. 328-336, 1994