Tertiary Storage Support for Large-Scale Multidimensional Array Database Management Systems

Bernd Reiner, Karl Hahn

FORWISS (Bavarian Research Center for Knowledge Based Systems)
Technical University Munich
Orleansstrasse 34, 81667 Munich, Germany
{reiner, hahnk}@forwiss.tu-muenchen.de

Abstract

Many large-scale scientific domains often generate huge amounts (hundreds of terabytes) of multidimensional data. The only practicable way for storing such large volumes of multidimensional data is a tertiary storage system. Unfortunately in commercial multidimensional *Database Management Systems* (DBMS) the access is optimized for performance with primary and secondary memory. Tertiary storage memory is not or only in an insufficient way supported for storing or retrieval of multidimensional array data. The intention of this paper is, to combine the advantage of both techniques, storing large amounts of data on tertiary storage media and realizing efficient data access for retrieval with the commercial multidimensional array DBMS RasDaMan.

1. Introduction

Many natural phenomena like atmospheric data transmitted by satellites, computational fluid dynamics, climate-modeling simulations, flow modeling of chemical reactors or dynamics of gene expressions can be modeled as spatio-temporal array data of some specific dimensionality. Their common characteristic is that large volumes (hundreds of terabytes) of *Multidimensional Discrete Data* (MDD) have to be stored. The common state of the arte of storing such large volumes of data is actually *Tertiary Storage* (TS) systems, where data are stored as file on robot controlled tape libraries or jukeboxes which provide automated access to thousands of media (e.g. magnetic tapes). Main disadvantages, concerning data access, are high access latency compared to hard disk devices and to have no direct access to specific subsets of data. If only a subset (black area) of such a large data set is required, the whole file must be transferred from tertiary storage media (left side of Figure 3), which can take many hours (load, search and rewind several cartridges).

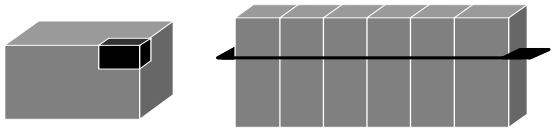

Figure 1: Typical example of data access

The further processing with data across a multitude of data sets, for example, time slices (black area), is hard to support (right side of Figure 1). 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 [2].

In high performance computing applications DBMS are typically used for meta-data management only, where meta-data contains the information about the location of the data sets (stored on which medium). Since few years commercial multidimensional array DBMS like RasDaMan offer efficient storage, retrieval or manipulation of MDDs.

**Proceedings of the 28th VLDB Conference,
Hong Kong, China, 2002**

RasDaMan has a extended query language (RasQL) with special multidimensional operations like geometric, induced and aggregation operations [1, 5]. RasDaMan is also supporting the retrieval of subsets of large data sets by using tiling strategies, which is a main advantage. Unfortunately the possibility of TS access is lacking, so mass data can't be managed automatically with RasDaMan. Within the ESTEDI project the multidimensional DBMS RasDaMan will be enhanced with intelligent mass storage handling and optimized for high performance computing [2].

2. Tertiary Storage Support for DBMS

Within the ESTEDI project we have integrated the tertiary storage support into the first commercial multidimensional array DBMS RasDaMan, which is distributed by Active Knowledge GmbH. The kernel of RasDaMan was extended with easy to use functionality to automatically store and retrieve data to/from TS-Systems. Figure 2 depicts the architecture of the extended RasDaMan system.

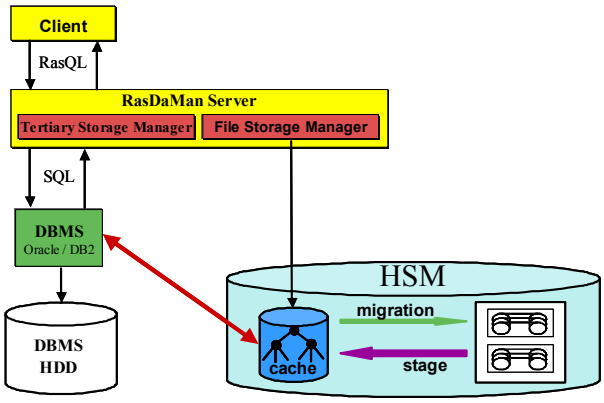

Figure 2: Extended RasDaMan architecture with tertiary storage interface

The left side of the figure shows the original RasDaMan client/server architecture with the conventional DBMS (e.g. Oracle, which is used by RasDaMan as storage and transaction manager). We realize the TS support by connecting a *Hierarchical Storage Management* (HSM) System like SAM (Storage Archiving System) from LSC Incorporation or UniTree with RasDaMan (to the bottom right in Figure 2). Such HSM-Systems have been developed to manage TS archive systems and can handle thousands of TS media. Another important reason for using HSM-Systems was, that such systems already in use by the ESTEDI project partners.

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 [3, 5, 6, 7]. Tiles are in RasDaMan the smallest unit of data access. The size of tiles (32 KByte to 640 KByte) is optimised for main memory and hard disk access. Those tile sizes are much to small for data sets held on TS media [4, 9]. 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 TS access costs. Smart means to exploit the good transfer rate of TS devices and to take advantage of other concepts like clustering of data. Super-Tiles are the access (import/export) granularity of MDD on TS media. Extensive tests have shown that a Super-Tile size of about 200 MByte shows good performance characteristics in most cases. The retrieval of data stored on hard disk or on TS media is transparent for the user. Only the access time is higher if data stored on TS media. A more detailed description of the Super-Tile concept can be found in [9].

Two further strategies for reducing TS access time are clustering and caching. An optimised clustering of data sets reduces the positioning and exchange time of TS media significantly. Our algorithm uses the spatial neighbourhood of tiles within the data sets and supports inter and intra Super-Tile clustering. We also have realised the caching of data stored on TS media in order to reduce expensive TS media access. Please find more information about the implemented clustering and caching strategies in [9].

3. Performance Aspects

Typically in many large-scale domains whole MDDs (stored as single files) have to be loaded from TS devices, even if only a subset of the MDD is required for further analysis tasks. With our developments we realize a fast and efficient access to TS media and provide access functionality like

[1] The ESTEDI (European Spatio-Temporal Data Infrastructure for High Performance Computing) project (http://www.estedi.org) is funded by the European Commission under FP5 grant no. IST-1999-11009.

retrieval of subsets as common for DBMS since a long time. This means the request response time scale now with the size of the query box, not with the size of MDD like the traditional case in many domains. Figure 3 shows the TS access comparison between the traditional access and the access with Super-Tile granularity.

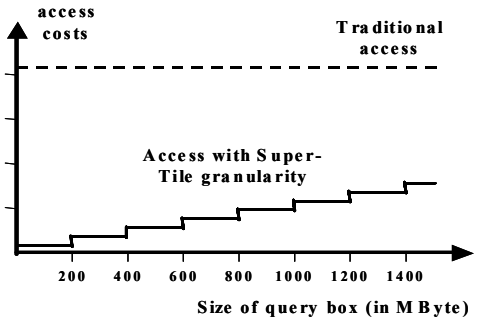

Figure 3: Tertiary storage access comparison

In this example the MDD size is 5 GByte and the query box is increasing. If we have access with Super-Tile granularity the access costs are increasing step by step. Otherwise with traditional access we have constant access costs, as the complete MDD must be loaded from TS media.

The implemented TS support for the DBMS RasDaMan is operational and shows, as we have expected, a very good performance compared to the common state of the arte of retrieving MDDs in large-scale scientific domains. Now we have the chance to make many interesting and important evaluation, which are currently impossible (see 1).

4. Future Work

The development of scheduling techniques for multidimensional array data streamlined for Super-Tiles will be the research work of the near future. For TS media scheduling means the optimization of the media read order. We want to reduce with this optimization the time expensive media seek and exchange operations. The main focus is on scheduling policies that process first all requests on a loaded medium before exchanging it in the loading station of the robotic library. Generally we can differ intra and inter query scheduling. On the one side intra query scheduling will optimize the request order within one query. On the other side inter query scheduling can be done by examining the query queue of the RasDaMan DBMS. If the processed query needs Super-Tiles from one specific MDD (stored on one tape) we determine further queries of the query queue whether they also needs several Super-Tiles from the same MDD. All needed Super-Tiles of the same MDD will be imported at the same time from the TS media. These scheduling techniques have a grate potential for performance improvement.

References

1. Baumann P.: A Database Array Algebra for Spatio-Temporal Data and Beyond, Proc. of the 4th Int. Workshop on Next Generation Information Technologies and Systems (NGITS), p. 76-93, 1999
2. Baumann P.: Array Databases Meet Super-computing Data – the ESTEDI Project, Internal ESTEDI Report, 2000
3. 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
4. Chen L. T., Rotem D., Shoshani A., Drach R.: Optimizing Tertiary Storage Organization and Access for Spatio-Temporal Datasets, NASA Goddard Conf. on Mass Storage Systems, 1995
5. Furtado P. A., Baumann P.: Storage of Multidimensional Arrays Based on Arbitrary Tiling, Proc. Of the ICDE'99, p. 480-489, 1999
6. Furtado P. A.: Storage Management of Multidimensional Arrays in Database Management Systems, PhD Thesis of Technical University Munich, 1999
7. Sarawagi S., Stonebraker M.: Efficient Organization of Large Multidimensional Arrays, Proc. of Int. Conf. On Data Engineering, volume 10, p. 328-336, 1994

Authors' Refereed Publications

8. Hahn K., Reiner B., Höfling G., Baumann P.: Parallel Query Support for Multidimensional Data: Inter-object Parallelism. To appear in the Proc. of the 13th Int. Conf. on Database and Expert Systems Applications (DEXA), 2002
9. Reiner B., Hahn K., Höfling G., Baumman P.: Hierarchical Storage Support and Management for Large-Scale Multidimensional Array Database Management Systems. To appear in the Proc. of the 13th Int. Conf. on Database and Expert Systems Applications (DEXA), 2002