Paper # 184-29: SAS® Intelligence Storage Overview

Keith Woolner, Storage Product Manager, SAS Institute, Cary NC

ABSTRACT

Relational Database Management Systems evolved out of the need for reliable transactional processing. However, the emergence of business intelligence and analytical applications places different demands on the RDBMS. SAS Intelligence Storage is designed from the ground up to serve the needs of business intelligence and analytical applications. This paper details some of the characteristics of SAS Intelligence Storage that make it uniquely suited for these types of applications.

INTRODUCTION

Relational Database Management Systems (RDBMS) have been successfully used in operational business systems for years. This is because they have been designed and tuned for operational applications. However, the emergence of business intelligence and advanced analytics applications has created different types of demands on data storage, and traditional RDBMS vendors have been slow to react and accommodate the demands of these applications. SAS Intelligence Storage, in contrast, has been designed from the ground up top be optimized for business intelligence and analytic intelligence applications.

DESIGNING FOR TRANSACTIONAL VS INTELLIGENCE APPLICATIONS

To understand the critical differences between traditional RDBMSes and SAS Intelligence Storage, we need to understand the history of the operational systems that RDBMSes were originally designed for. Typically, operational systems are transactional systems, collecting large numbers of small transactions, or amounts of data, such as orders, financial entries, or inventory changes. In order to efficiently store these small transactions, RDBMSes typically use a row-based storage mechanism that it optimized for reading and writing individual rows of data within a table.

In addition, these operational systems were designed to handle large numbers of concurrent users, all of whom may be entering new data or altering existing data. In order to accommodate the activities of all of these different users, RDBMSes must manage concurrent access to all of the data tables, and respond robustly when transactions fail, or are cancelled by the user. The necessary criteria for correctly managing transactional systems are captured by the so-called "ACID" transaction processing guarantees of Atomicity, Consistency, Isolation, and Durability.

However, RDBMSes that ensure ACID incur a lot of extra work and overhead in providing those capabilities. Coordinating the actions of potentially thousands of users against the same data table consumes CPU time and I/O throughput. If the system will not have a transactional kind of usage pattern, the overhead of ACID has little value, and simply degrades performance.

DIFFERENT ACCESS PATTERNS OF INTELLIGENCE APPLICATIONS

Business intelligence and analytic intelligence applications don't share the same usage profile as transactional applications. Rather than creating small amounts of data repeatedly, these applications typically read large amounts of existing data at once, processing and analyzing the data without altering it. When business and analytic intelligence applications create data table structures, such as during an Extract-Transform-Load (ETL) operation, the system is reading a large amount of data at once, processing it as a single operation, rather than independent bits of activity. The overhead associated with ACID doesn't buy you much in this scenario.

SAS INTELLIGENCE STORAGE

The components of SAS Intelligence Storage, including SAS Datasets, SAS Scalable Performance Data Server, and SAS OLAP Server are all designed for the usage patterns of business intelligence and analytical intelligence applications. They are optimized to read large amounts of data at once, quickly and efficiently, without the overhead of unnecessary ACID overhead of transaction management. We'll focus on the two server components on SAS Intelligence Storage, and discuss some of the ways they are optimized for the data access patterns of business intelligence and analytic intelligence applications.

SCALABILITY THROUGH PARALLELISM

Both SAS Scalable Performance Data Server and SAS OLAP Server are able to partition tables across multiple storage devices and I/O channels. This enables tables and cubes to be stored across multiple devices and read in parallel, increasing total I/O throughput. As reading large amounts of data is common in business intelligence and analytic intelligence applications, increasing the total effective I/O throughput in the storage component increases the performance of these applications, and increases the number of concurrent users these servers can handle.

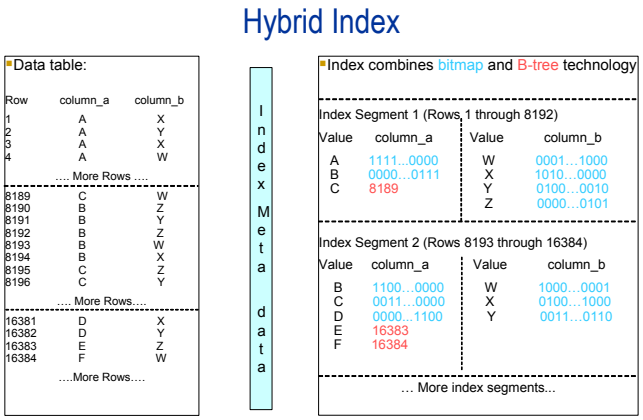
Similarly, both SAS Scalable Performance Data Server and SAS OLAP Server are able to leverage multiple CPU's and run multiple threads of processing. This

enables SAS to more effectively leverage all available computing resources, and increase speed and scalability of data intensive applications that depend on accessing the data.

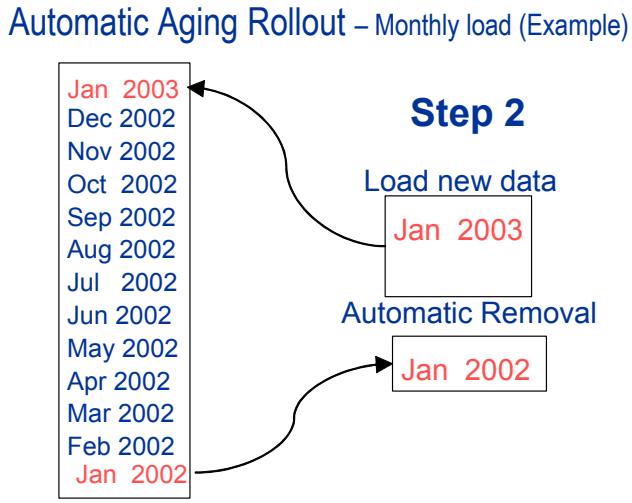
SAS SCALABLE PERFORMANCE DATA SERVER

SAS Scalable Performance Data Server also has additional capabilities tuned to the needs of business intelligence and analytic intelligence applications. The SAS Scalable Performance Data Server uses fixed length records to store rows in a table, which enables a very fast calculation of the location of any row within the table, ultimately delivering faster lookups and retrievals of subsets of data.

To further enhance locating subsets of data, the SAS Scalable Performance Data Server uses a unique hybrid indexing technology. Indexing is a way to speed up locating specific rows in a table based on the values of a particular column without having to do a full read of the entire table. Two types of indexes are common in traditional RDBMSes: B-tree and bitmap. A B-Tree (or balanced tree) is a specific kind of index algorithm that divides the values of the index into subgroups to help find them quickly. Bitmap indexes use a different strategy, and work best when there are relatively few distinct values in a column. A bitmap is created for each possible value of a column and for each row, a bit is set to 1 if it matches the bitmap and 0 if it isn't. Each approach has different strengths. Depending on the uniqueness of the data values, the B-Tree index straetegy is most efficient when most or all values are unique (or occur only a few times). The Scalable Performance Data Server is adaptable in that it combines both approaches in a single index strategy that leverages the strength of each. The end results are indexes that are very fast, easy to maintain, and consume fewer disk resources than traditional indexes use.



Automatic Aging Rollout is a new feature in the latest release of the Scalable Performance Data Server. This feature is for loading and removing time based units of data from a data warehouse table. For example, a DBA may want to store only the most recent 36 months worth of sales data in the warehouse. As each new month of data is added, the oldest month should be deleted from the warehouse. The Automatic Aging Rollout feature makes this easy and automatic. It uses a virtual table called a cluster that contains slots – 36 slots in this case. The sales data is partitioned into time-based slots -- each slot contains one month's worth of data. To applications talking to the SAS Scalable Performance Data Server, the virtual table appears no different from any other table. When the slots are loaded with the SAS Scalable Performance Data Server tables and a new one is added, the oldest one will be automatically deleted to accommodate the newest, thus making sure that only the amount of data desired is retained. This reduces the cost of maintaining the warehouse, and ensures that the right amount of data is available to the intelligence applications relying upon it.



SAS OLAP SERVER

SAS OLAP Server is another component of SAS Intelligence Storage. It allows users to explore, slice, and dice large amounts of data very quickly, by pre-computing the aggregate results along the most common ways of breaking down the data. The categories used to classify and aggregate are referred to as dimensions of the data, and within each dimension, there can be many different levels of granularity that a user may want to see (e.g. users looks at sales revenue by geography may want to see revenue for a region, country, state, or city, which represent different levels of the geography dimension). The combination of dimensions and measures (such as total revenue) collectively are called a cube. Automatically handling the computation of totals along different levels and dimensions rather than running a query to determine the sum each time it is requested enables SAS OLAP Server to rapidly serve portions of the cube to users of business intelligence and analytic intelligence applications.

SAS OLAP Server, like SAS Scalable Performance Data Server, can partition and distribute cube data across

multiple I/O devices to maximize parallel I/O throughput. Similarly, SAS OLAP Server uses a multithreaded query engine to leverage multiple CPU's to efficiently use all available computing resources to service large numbers of users running complex queries.

SAS OLAP Server also allows administrators to compress seldom used data to make more efficient use of storage resources, and tune the in-memory query cache to maximize the speed at which the most frequently requested results are delivered back to users.

CONCLUSION

The separation of SAS Intelligence Storage from the transactional operational database enables a system that can be tuned for the unique demands of business intelligence and advanced analytics applications. The benefits include

- transparent access to data,
- fast response times for large queries,
- the ability to perform complex calculations,

Knowledge workers can effectively perform data exploration, what-if simulations, sensitivity analysis, modeling, and prediction, in addition to accessing summary and detailed data. As the volume of data within the enterprise explodes, having a storage solution designed to handle data intensive applications and complex usage patterns becomes even more critical. By providing storage options designed from the outset to provide relational, OLAP and parallel storage options to suit all types of application profile, SAS Intelligence Storage ensures that maximum flexibility is available to meet different application usage scenarios in the most appropriate manner while lowering total cost of ownership and without sacrificing performance and scalability.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Keith Woolner
SAS Institute
SAS Campus Drive, R-1150
Cary, NC 27513
Work Phone: 919-531-9271
Fax:919-677-4444
Email:keith.woolner@sas.com
Web:www.sas.com