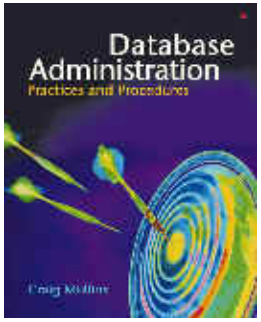
Craig S. Mullins

Return to Home Page

July 2008

The DBA Corner

by Craig S. Mullins



Consider Data Access Auditing to Classify Database Data

According to a recent study conducted by the University of California at Berkeley, each year approximately 5 exabytes *(*10^{18} bytes) of new information is produced. And ninety-two percent of that information is stored on magnetic media, mostly hard disks. Indeed, businesses today are gathering and storing more data than

ever before. Multi-terabyte databases are common, and some organizations are approaching a petabyte.

Table 1. Storage Abbreviations

Abbrev.	Term	Amount
KB	Kilobyte	1,024 bytes
MB	Megabyte	1,024 KB
GB	Gigabyte	1,024 MB
TB	Terabyte	1,024 GB
PB	Petabyte	1,024 TB
EB	Exabyte	1,024 PB
ZB	Zettabyte	1,024 EB
YB	Yottabyte	1,024 ZB

Several factors contribute to this explosion of data. For one, technology drives storage demand. Database technology has advanced to be better able to store and manage unstructured data. Whereas structured data is the data that most database professionals are familiar with (numbers, characters, and dates), unstructured data is basically anything else (e.g. text files, images, audio files, and videos). Unstructured data complicates data management because unstructured data is usually large and unwieldy as compared to structured data. Additionally, different mechanisms are required to access and modify this data, complicating program development and database administration.

Technological change is not the only force contributing to data growth – government regulations are complicit, too. Regulatory compliance places stringent rules on data retention and management (e.g. Sarbanes-Oxley, Gramm-Leach-

Bliley, HIPAA, etc.). For organizations to be in compliance with these laws usually requires retaining additional data and possibly capturing more details about how, when, and by whom the data is being used. And the amount of database access information required for regulatory compliance is so large that this additional data can create a new need in terms of storage management.

So the amount of data under management is expanding, and as it does, cost expands, too. You have to store that data somewhere and there will be an associated cost – including the initial cost of the storage device along with the additional cost to manage it.

Some organizations attempt to determine what data can be eliminated. But this can be dangerous. Enterprise data growth is going to happen, you can't stop it and you probably can't or don't want to slow it down because as organizations become more analytically mature they discover new kinds of data they would like to capture and analyze.

Another, more realistic tactic, is to categorize data based on its usage and importance, and then deploy it on the most cost-effective storage medium for its type. Such an approach is referred to as Information Lifecycle Management (ILM). But there are many issues that need to be resolved to make such an approach effective:

- A useful categorization technique needs to be defined
- A method of measuring data such that it can be categorized is required
- All of the available storage options need to be aligned with the data categories

- And a technique needs to be created for moving the data to the appropriate storage devices

The first step to ILM is to develop a data classification scheme. An example of such a scheme, invented by Teradata, is called multi-temperature data. This technique deploys four categories: Hot, Warm, Cool, and Dormant. The temperature of data is defined as a function of the access rate for queries, updates, and data maintenance. In other words, hotter data is accessed more frequently than warm data which is accessed more frequently than cool data. Finally, dormant data is rarely updated or queried and it is part of a static data model.

The characteristics of your data can help to determine its appropriate temperature. For example, data tends to cool over time. And often the temperate of the data correlates well with data volume; the larger the volume, the cooler the data. But these are not hard and fast rules. What is needed is a technique to audit the frequency and type of data access for every piece of data in the database.

Generating an accurate view of all data access has traditionally been a very difficult exercise. The difficulty has always been a combination of the fact that databases store huge volumes of data, have very high degrees of complexity, and that mapping of data access must be done in a non-intrusive manner. Data access auditing solutions are available that can provide this type of information. The best of these solutions will perform non-intrusive enterprise data access classification by monitoring the database server and inspecting all SQL requests and responses. Using this non-intrusive database monitoring technology one can quickly get a view of all data access at a very granular level. One such solution provider is Guardium, Inc.

By monitoring database requests it is possible to create a detailed classification of which data is being used, how it is being used, and when it is being used. This information can be used to classify the data using multiple dimensions – such as time, usage, and frequency – and then to translate that into the temperature of the data so it can be placed on the appropriate storage mechanism.

Once categorized, data can be aligned with appropriate storage types. Hot data can be placed on devices offering high performance, reliability, and large capacity. Warm data can be placed on less expensive disk devices that offer good performance and reliability, but are not top-of-the-line. Cool data is not accessed very often, though it likely should still reside on direct access storage devices. Dormant data, which may not have been accessed for a long time can be moved to offline storage systems such as optical disk. Alternately, dormant data can be archived to a separately managed archive data store so it no longer impacts operational systems but where it can be safely accessed when needed.

Resources: University of California at Berkeley study

http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm#summary

From Database Trends and Applications, July 2008.

Home.