Craig S. Mullins
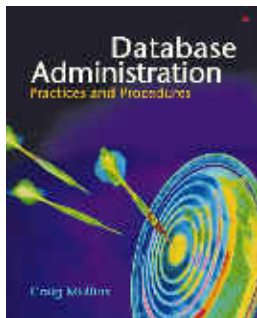
April 2008

The DBA Corner
by Craig S. Mullins

Database Data Needs Long-Term (100 Year) Archiving Solution

The Storage Networking Industry Association (SNIA), a not-for-profit global organization, recently released the results of its 100 Year Archive Requirements Survey. The survey, which was produced by SNIA's Data Management Forum, contains many enlightening nuggets of information. I'd like to share some of them with you and interpret them from my perspective as a database professional.

The survey was conducted worldwide and boasts participation from 276 different organizations. Respondents were practitioners who specialized in long-term archiving across a broad spectrum of practices, including IT, records management (RIM), archivists, legal, security and business executives.

Key survey findings presented in the report include the following:

- Long-term digital information retention needs are real: 80 percent of the respondents have information they must keep over 50 years, and 68 percent of the respondents said they must keep this data more than 100 years.

- Long-term generally means greater than 10 to 15 years, a period beyond which multiple physical media and logical format migrations must take place.

- Database information was considered to be most at risk of loss.

- 70 percent of respondents said they are "highly dissatisfied" with their ability to read their retained information in 50 years.

I find these results quite fascinating, particularly the one indicating that database information is at the greatest risk. I do not find it fascinating because I don't believe it, but because I'm surprised that so many actually recognize the problem. If you read the industry trade publications you would be excused for thinking that long-term retention and archiving is a problem that mostly impacts email and other unstructured data. Doesn't it seem like most of the hype has been focused on email?

But database data is at risk. All too often databases are designed, implemented, and turned over to production without any up-front planning for long-term retention and archival. Data just keeps piling up in the database until performance becomes an issue. And only then does database archiving become an issue.

At this point, sometimes the data simply gets purged without any concern for long-term retention. This happens less frequently in this day-and-age of governmental regulations though. Alternate approaches include sending unload files or database backups to an offsite location and then purging the data from the database. But this is not a true, well-planned archive. I mean, have you ever tried to actually use those "archived" files? Database structures change over time. Could you even figure out what data is on those unload files, let alone an image copy backup, at some point 10, 25, or even 100 years in the future?

Okay, I can hear you saying that you'll be retired and it will be someone else's problem, right? But that is not a professional response, nor is it the appropriate approach for your organization. Database archiving needs to be approached as a management discipline where the data retention and preservation needs for archiving are collected when the database is being designed.

With database archive policies in place, when data is no longer needed for the daily operations of the business, it will be periodically archived into a specialized system designed specifically to meet the needs of long-term retention and preservation of database data. This has the effect of minimizing the impact of data growth, while ensuring the on-going preservation of the data. These specialized needs include:

- Policy-based archiving, where a policy is set up and data is logically selected from the database for archival and retention based on that policy.

- Long-term retention, to ensure the data is preserved for long periods of time and taking into account that the life of the storage media may be shorter than the life of the data.

- Support for very large amounts of data in archive; because data growth continues unabated, the archive must be able to handle lots and lots of data.

- Ability to maintain archived data even as the operational systems from which it was archived change or are retired.

- Complete independence from the applications, DBMS, systems, and operational metadata. The archive must be useable and accessible in a stand-alone manner because the operational environment from which it was archived may not exist, at least not in the same form.

- Protection of data authenticity, because the data, once archived, should never be changed.

- The ability to access archived data when needed, as needed, preferably using an industry-standard mechanism (e.g., SQL).

- Policy-based discarding of data after the retention period expires. Data is an asset as long as you have to retain it for a business or regulatory purpose, but it becomes a liability if you store it even one day longer than you must.

So the takeaway from this survey is that digital information is at risk of being lost. And database data is most at risk. Inaction is not acceptable when the long-term survival of our data - our institutional memory - is at risk. DBAs take heed! Ignore database archiving at your peril.

For more information and to view the 100 Year Archive Requirements survey, visit http://www.snia.org/forums/dmf/programs/ltacsi/100_year/.

From Database Trends and Applications, April 2008.

Home.