Mesabi Group

Commentary

April 28, 2006

Database Archiving: A Necessity, Not an Option

For those enterprises with mission-critical databases, database archiving is rapidly becoming a necessity, not an option. Database archiving solves a number of seemingly unrelated issues. Those include improving the performance and availability of live production databases, managing data retention policies — notably compliance — effectively, and preserving database data as long as required. Above all, database archiving, now more than ever, keeps the business running. Database archiving deserves a thorough examination.

Database Archiving Is a Three-for-the-Price-of-One Solution

What do these three database-related problems have in common?

- Production databases with unnecessarily huge data stores suffering from poor performance as well as unacceptable backup and restore times; moreover, the increasing size of the IT infrastructure is magnified by the number of copies of the live production databases — online standby, replica for disaster recovery, test, development, and so on
- Heightened requirements for managing the retention of database information, especially compliance-related information
- Figuring out how to gracefully retire — but not destroy — database information that has to be just-in-time available for periods longer than the physical lifetimes of the hardware in which it is originally stored

The answer is that database archiving can provide a solution to all three problems.

Of course, not all three problems have equal weight. Operating day-to-day revenue-producing business-critical OLTP (on-line transaction processing) databases calls for constant vigilance to maintain their performance and availability.

However, how to manage the retention process for database information is attracting much more attention in the light of sharpened compliance requirements, such as Sarbanes Oxley.

Moreover, it may seem attractive to defer the problem of long-term retirement of data indefinitely, but the resulting inventory costs can quickly grow quite large. What better time to do address data retirement than as part of an overall

26 Country Lane
Westwood, Massachusetts 02090
Telephone: 781 326 0038
Fax: 781 326 0038
www.mesabigroup.com

database archiving strategy that needs to be done anyway?

Finally, solving all three problems by one database-archiving strategy can simplify IT architectures — and complexity, as server consolidation has shown, costs money.

What Database Archiving Is All About

Database archiving migrates database information from live production databases to database archives and then manages the integrity and accessibility of information in those archives. Note that database archiving, like Information Lifecycle Management (ILM), treats data differently during the different stages of its lifecycle. To put it another way, database archiving can be used as a method for implementing some of the key functions of ILM on databases (i.e., structured data not flat files).

Database archiving divides database data into three different pools of storage — active changeable, active archive, and deep archive (Table 1).

The active changeable pool is the familiar mission-critical production database which consists of open transactions that have been newly created or those being updated as a part of the daily operations. The active archive pool consists of closed transactions that can still be used for ongoing business purposes, but can be better managed for data retention purposes, including compliance. The deep archive pool (what used to be called archived data) is data that must be retained but is no longer needed for the ongoing business. The deep archive pool now has enhanced capabilities for both long-term data preservation and restore, as these capabilities may be necessary for long-term compliance.

Why Live Production Databases Become Overweight

Mission-critical and other essential business processes typically operate by carrying out transactions that create and change 'open' data. At some point in time, the original purpose of a data item — a customer order, insurance claim, or reservation — has been accomplished. At that point — when invoices or claims have been paid, reservations fulfilled, or return periods have expired — the data is 'closed.' Closed data is fixed-content, read-only data.

In today's databases, closed data tends to remain in a live production database's data store for quite some time — years or indefinitely. Although it may be difficult to tell just when data becomes closed, sheer IT inertia is one reason for the buildup of closed data. Moreover, with mandatory compliance becoming more critical, "saving everything" has become an "easy" way to compliance.

Slimming Down is Good for Databases

Slimming down a mission-critical database by migrating closed transactions to what can be called an active archive can yield significant benefits with respect to both performance (the production side of the house) and availability (a key goal of data protection).

Performance

Performance improvements come about because queries that would have competed with other transactions on an overweight data store are redirected to the active archive. As a result, like a data warehouse, the read-only active

26 Country Lane
Westwood, Massachusetts 02090

Telephone: 781 326 0038
Fax: 781 326 0038
www.mesabigroup.com

Table 1: The Three Lifecycle Stages for Database Information

	Active Changeable	Active Archive	Deep Archiving
What	Open transactions	Closed transactions	"Retired" closed transactions
Purpose	Manage ongoing business processes, especially those that are revenue-producing	Leverages an organization's informational knowledge for reporting and analysis; enables application of data retention policies, especially for compliance	Long-term preservation of data that may need to be recalled for compliance or auditing purposes
Lifecycle Stage	Youth	Middle age	Old age
Access Method	Online	Online	Online, nearline, or offline

Source: Mesabi Group April 2006

archive can deliver far higher querying performance by specializing in queries; while the original production data store benefits from its lightened load of transactions (e.g., avoiding the fabled "query from hell") and from a smaller data store.

Availability

The larger the database, the longer the time it takes to fully back up and restore, and the greater the risk of failure. Thus, splitting closed data from the production data store makes that store much more highly available.

Note that because the data split off is closed data, incremental backups (which involve changed data) are not speeded by addition of an active archive. However, slimming down the database may give the option of running more frequent full backups. That, in turn, means that — in the case of a full restore — fewer or no incremental tapes would have to be applied, shortening restore time and reducing the risk of a tape handling problem or system failure during restore.

The big benefit comes on the restore side of the house which is the purpose of an effective data protection strategy. The restore process is roughly proportional to the size of the database so obviously a slimmed down database requires much less time to restore.

An Active Archive is the New Home for Closed Transactions

An active archive consists of closed data that has been removed from the live production data store of active changeable information. At the same time, an

26 Country Lane
Westwood, Massachusetts 02090
Telephone: 781 326 0038
Fax: 781 326 0038
www.mesabigroup.com

active archive database is *still online* to the end user, which means transparent access to the data with a reasonable response time.

An authorized end user can access database-archived data in two primary ways:

- Through queries targeted to access only the active archive
- Natively through the original applications. The fact that the original data store has been split into active changeable and active archived parts will be transparent to the applications, but not necessarily to the databases (also known as database management systems) accessing the split data stores.

For example, in the latter case, a display of a particular customer's data — both open and closed — for the past year would use both the active changeable and the active archive pools, but the fact that a composite from both pools was used would not be visible to an end user.

The software that creates and maintains the active archive has special responsibilities in order to achieve both transparency to the database when desired and the ability to support optimization of actions against each data store. For example, the active archive software must maintain the referential integrity (consistency of links between data) that a relational database requires (relational databases are typically used for production applications).

What Active Archiving Is Good For

An active archive serves three useful business functions:

- It fulfills both operational and informational requests
- It enables the retention and ongoing administration of data, including compliance efforts
- It employs storage and other IT assets more efficiently and effectively

Helping Serve the Management Information Needs of the Business

Of course, the active archive can continue to serve the operational needs through the user's familiar application as before. Moreover, the organization can build new decision-support applications that directly target the active archive, such as financial statements with historical context.

Note that the data in an active archive is the gold standard copy of the data, the official production version. Therefore, unlike de jure data-warehouse data, active archive data has to be a faithful migration copy of the production data store's data. For example, the active archive may be at a finer level of granularity (say every ATM withdrawal) than a de jure data warehouse might need.

An organization has to decide if the active archive can serve as a de facto data warehouse or a source for feeding a de jure data warehouse or data marts. Moreover, the active archive may serve as a way to split the enterprise's data warehouse into "needs refreshing" and "doesn't need refreshing" parts.

A production data store may feed the de jure data warehouse initially (as a copy of closed data could be sent to the data warehouse at the same time that it was migrated to the active archive). However, after that time if a data warehouse needed reloading, the source would be

26 Country Lane
Westwood, Massachusetts 02090
Telephone: 781 326 0038
Fax: 781 326 0038
www.mesabigroup.com

the active archive as the official source of closed data. Moreover, data from the active archive might be copied to a temporary work space for special business intelligence analyses, such as a data mining analysis.

Managing Retention Is Essential

Only fixed content information can be retention-managed effectively. That means that policies can be put in place for deciding when the data should exit the active archive (if ever) and how (destruction of data or movement of the data to a deep archive). This means that only fixed content information that has a high-enough value will be kept in the active archive.

And only retention-managed data can be compliance data. Compliance data is a subset of retention-managed data that must be more tightly controlled, such as managing a chain of custody for the information.

Depending upon organizational requirements, compliance data might be kept in the original active archive or migrated to another archive, say a deep archive that encapsulates all the compliance information.

Management and Cost Benefits

As noted above, both the active changeable pool and the active archive pool can be optimized and scale independently. That makes it easier to manage the performance and growth of each.

Moreover, in some cases the active changeable and active archive pools can act as a "virtual" system, with the active archive requiring less performance. In this case, an active archive may use high-capacity, cost-effective disk (i.e., SATA) that can result in significant cost savings and still achieve necessary overall performance levels.

Data protection for active archives can be through copy creation (e.g., via replication), not traditional backups, because backups are only needed for open data. The implicit assumption with backups is that the data is subject to change. This copy creation can be carried out when the data is moved from the production data store to the active archive in an archiving cycle. (That does not mean that traditional backups cannot be used after an archiving cycle; only that there is an alternative.) Again, this improves availability and simplifies IT's job.

The active archive thus has a different set of service level agreements (SLAs) for both performance and data protection than does an active changeable pool of data.

Off to the Deep Archive

At a policy-driven time or event, data is migrated from the active archive to a deep archive. Of course, some data never makes it to the deep archive; rather, the data is simply deleted from the active archive.

Why keep old unused data at all? Some data may be compliance data that needs to be kept around for very long periods of time. That data needs to be made available upon demand. For other data, there may be a slight possibility that the data need to be recalled, e.g., for unusual auditing.

Two key problems are that the data may outlive the media on which it is currently stored and that it may out-live the application that can read it.

26 Country Lane
Westwood, Massachusetts 02090
Telephone: 781 326 0038
Fax: 781 326 0038
www.mesabigroup.com

Database archiving can help with both issues. Database archiving can manage migration from older media to newer media when necessary. Database archiving can put the data in an open standard format (say, XML) that can be recovered when necessary regardless of whether the application that created it still exists or not.

So What's Holding Enterprises Back

Many IT organizations confuse carrying out backup with effectively archiving the data. A backup is a data protection copy of data that is organized by date and time. The purpose of a backup is only to serve as a source for restoring production data when needed. Unlike archived data, a backup copy of fixed-content data may not be quickly accessible, and may not be in a format for retrieval once the original application is no longer available.

Archiving ≠ Backup

An archive is production data that it is fixed content and has been removed from a live production pool of data to another pool of data. An archive is the original production data, not a copy of the data that been made for data protection purposes.

Enterprises may also be worried about touching sensitive mission-critical systems and/or may feel that they have no window of opportunity to migrate the data to an active archive. However, the alternative is allowing fixed data to accumulate forever. The sooner the migration is carried out, the quicker, easier, and simpler it is — sips instead of a torrent.

Conclusions

Database archiving is the foundation for managing the lifecycle of database data. Database archiving relieves the burden upon live production databases and, in so doing, improves performance and availability. Archiving fixed content data enables the application of data retention policies and is essential if compliance is mandated. Database archiving is an ongoing, continuous process, not one that is done periodically as part of a database cleanup

Yes, using database archiving requires work in planning, such as classifying data. However, if an enterprise's databases are essential to carrying out its mission, then database archiving should be given careful consideration and high priority. The sooner that an organization gets started the better.

David Hill

Analyst Name: David Hill
Topic Area: Archiving

Mesabi Group LLC
26 Country Lane
Westwood, MA 02090
www.mesabigroup.com

This document is the result of research performed by Mesabi Group LLC. Mesabi Group LLC believes its findings are objective and represent the best analysis available at the time of publication. This Commentary is based upon research sponsored by Hewlett-Packard Company.

Phone: (781) 326-0038
email the author: davidhill@mesabigroup.com