Craig S. Mullins

Return to Home Page

Spring 2007

IT Compliance

Database Archiving for Long-term Data Retention

by Craig S. Mullins

Organizations are generating and keeping more data now than at any time in history. Many factors contribute to this reality. One contributing factor is general data growth. According to industry analysts, enterprise data is more than doubling every year. Additionally, as much as 80% of that data is not actively used to conduct business.

Why are we producing so much data? Advances in technology have better enabled our ability to capture and store data. But technology alone is not sufficient to account for the current rate of data growth.

Data is retained for both internal and external reasons. Of course, when an organization requires the data to conduct business and make money, then that data will be retained. And today's organizations are storing more data for longer periods of time for many internal reasons. Typically, data is stored longer than it used to be in order to enable analytical processes to be conducted on the data. As such, businesses are inclined to keep data around for longer periods of time.

But external reasons, typically driven by the mandate to comply with legal and governmental regulations also compel businesses to store additional data. Indeed, data retention is a significant aspect of regulatory compliance that requires focus and attention. The need to retain data is impacted not just by the normal culprits, like Sarbanes-Oxley and HIPAA, but also by over 150 international, federal, and local laws that govern how long data must be retained. Organizations need to develop plans for archiving data from the operational databases as their data retention requirements expand over longer and longer periods of time.

The Lifecycle of Data

As data moves throughout its useful lifecycle, it progresses through five distinct phases: data creation, operational, reference, archived, and discarded. Data is created at some point, usually by means of a transaction. For a period of time after creation, the data enters an operational state. The data is required to conduct business.

The operational state is followed by the reference state. During this phase data is still needed for reporting and query purposes: internal reports, external statements, or simply in case a customer asks about it.

Then, after some additional timeframe, the data is no longer needed for business purposes and it is no longer being queried. But the data must be saved for regulatory and legal purposes. This is the archive state. Then, after a designated period of time, the data is no longer needed at all and must be discarded.

The data states say nothing about where the data is stored or what technology is used. But it makes senses to move archive data out of the operational database for many reasons.

Database Archiving

First, let's define database archiving. ***Database Archiving*** is the process of removing selected data records from operational databases that are not expected to be referenced again and storing them in an archive data store where they can be retrieved if needed.

We say ***removing*** because the data is deleted from the operational database when it is moved to the archive. If the data is still required for operational requirements it is not ready to be archived. When data moves into the archive state, query and access is no longer anticipated, so removing it is not problematic.

Next, we say ***selected records***. We do not want to archive database data at the file or table level. We need only those specific pieces of data that are no longer needed by the business, but also related data. The archive must be able to selectively choose particular pieces of related data for archival… not the whole database, not an entire table, and not even a specific row. Instead, all of the data that represents a business object is archived at the same time. For example, if we choose to archive order data, we would also want to archive the specifics about each item on that order. This data likely spans multiple constructs within the database.

The next interesting piece of the definition is this: ***and storing them (the data) in an archive data store***. Archived data is stored separately from the operational database and does not require either the DBMS or the applications. Archived data is separate and independent from the production systems from which it was moved.

The final component of the definition that warrants clarification is… ***where they can be retrieved if needed***. The whole purpose of archiving is to maintain the data in case it is required for some purpose. So the data must be readily accessible without requiring a lot of manual intervention.

Database Archiving Requirements

Let’s examine the many capabilities required of a database archiving solution. Perhaps the most important consideration is that the archived data must be ***hardware and software independent***. Independence

is crucial because of the duration over which the archived data must exist. With a lifespan of decades it is likely that the production system from which the data was archived will no longer exist – at least not in the same form, and perhaps not at all. What changes have your production applications gone through over the course of the past ten or twenty years? It is completely unreasonable to expect that the operational environment will exist to enable access to archived data. We constantly change our databases. And the archive must be able to support multiple variations of the data structure as it changes.

The archive solution also must be able to storage a ***large amount of data***. As we store more data, we will archive more data. And when we combine this with long regulatory mandated data retention periods we have an explosive combination.

The archive must be able to ***manage data for very long time periods***. Many data retention requirements are stated in decades. So the archived data will outlive the systems and the programmers that generated them. The archive also will outlive the media we store it on. No media lasts forever: consider that the lifespan of tape is 7 years. So, the archive must be able to re-purpose the archived data from one type of media to another. And ideally it should do this automatically as the media reaches the end of its useful life.

Data, to support regulatory compliance, must remain unchanged once it is archived. So the archive must be able to ***protect against data modification***. Only read access should be available to the archived data (with the exception of periodic administration). Archived data

must be guaranteed to be authentic. And mechanisms to prevent surreptitious modification are necessary, too.

Finally the archive requires ***metadata*** to be useful: both metadata defining the archived data, as well as metadata defining what to archive and when. The archive must be able to store multiple versions of the first type of metadata. As the operational schema changes the archive must track and function across these variations in schema. The second type of metadata controls which data is archived, when, and from where. This is the metadata that drives and defines the archive itself. Both types of metadata are needed for the archive to operate.

Summary

Taking all of these considerations into account, then, a secure, durable archive data store must be used to retain data that is no longer needed for operational purposes, and it must enable query retrieval of the archived data in a meaningful format until it is discarded.

Operational databases are no place to maintain historical data over long periods of time. Database archiving will become more prevalent over time and wise organizations will start planning their database archiving needs today.

.

Home.