The Buffer Pool

Regulatory Compliance and Database Archiving

By Craig S. Mullins

Organizations are generating and keeping a more data now than at any time in history. Why is this so? First of all, the amount of data in general is growing. The general consensus among industry analysts is that enterprise data is growing at the rate of 125% annually. But perhaps more interesting is that as much as 80% of the information in those databases is not actively used.

But why are we producing so much data? True, technology advances 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 may need to be retained for both internal and external reasons. Internal reasons are driven by company needs. If an organization requires data to conduct business and make money then, of course, that data will be retained. Today's modern organizations are storing more data for longer periods of time for many internal reasons. Typically, data is stored longer than it used to be to enable analytical processes to be conducted on the data. Data warehousing, data mining, OLAP, and similar technologies have delivered more and better techniques for extracting information out of data. So businesses are inclined to keep the data around for longer periods of time.

But external reasons, typically driven by the mandate to comply with legal and governmental regulations are another significant factor driving the need to store more data for longer periods of time.

Legal Requirements to Archive

The corporate accounting scandals of the past few years have caused an onslaught of new laws to be written. These laws place regulations on how businesses are to treat their sensitive, business-critical data. Additionally, older laws that have been on the books are being enforced more rigorously than in the past. Basically, government regulations are being adopted to ensure that corporations are "doing the right thing" with their data. And a common component of regulatory mandates is the enforcement of longer data retention periods.

The number one driver of data management initiatives today is likely to be government regulations. The growing number of regulations and the need for organizations to be in compliance is driving data retention. There are the big regulations such as the Sarbanes-Oxley Act, HIPAA and BASEL II, but regulations are being enacted at the state and local level, as well. One such example is CA SB 1386, a California law that requires companies to disclose to their customers if their private data is compromised.

Indeed, there are over 150 international, federal and local laws that dictate how long data must be retained. Many of these laws greatly expand the duration over which data must be retained. Until recently most organizations dealt with mandatory retention periods of only a few years for important business data. If the data was kept around longer it was because of business reasons, not legal requirements. But the situation has changed due to the bevy of new regulations. Depending on the industry, what was once five or seven year retention periods may now expand into decades or even longer. Increasingly, retention periods are being determined almost exclusively by government regulations and not from business needs.

To comply with these laws corporations must re-evaluate their established methods and policies for managing and retaining data. What worked in the past to retain data for a few years is no longer sufficient over a much longer period.

According to research conducted by Enterprise Strategy Group – in its report titled "Digital Archiving: End-User Survey & Market Forecast 2006-2010" – digital archive capacity will increase nearly tenfold between 2005 and 2010. Total worldwide digital archive capacity in the commercial and government sectors will grow from about 2500 petabytes in 2005 to more than 27,000 petabytes by 2010. And they state that the major factors driving this growth will be regulatory compliance, corporate governance, litigation support, records management, and data management initiatives.

Clearly, organizations will be retaining more data over longer periods of time. And this will create the need for new policies, procedures, methodologies, and software to support storage, management and access of archived data.

The Lifecycle of Data

So how can we determine when data needs to be archived? In order to accurately answer that question we need to understand the different states of data as it progresses through its lifespan.

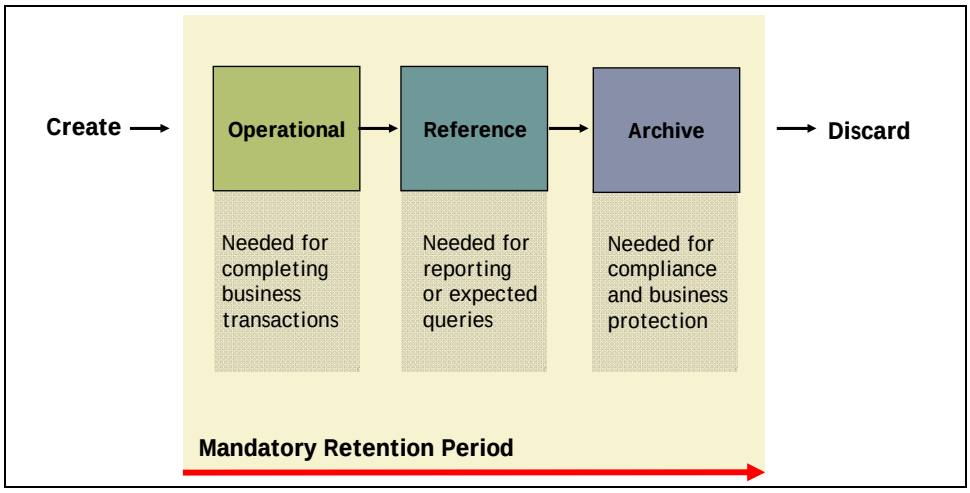

Figure 1. The Lifecycle of Data

The diagram in Figure 1 delineates the various states of data over its useful life. Data is created at some point, usually by means of a transaction: a product is released, an order is processed, a deposit is made, etc. For a period of time after creation, the data enters it first state: it is *operational*. That is, the data is needed to complete on-going business transactions. This is where it serves it primary business purpose. Transactions are enacted upon data in this state.

The operational state is followed by the *reference* state. This is the time during which the data is still needed for reporting and query purposes. It could be to produce internal reports, external statements, or simply exist in case a customer asks about it.

Then, after some additional period of time, the data is no longer needed for completing business transactions and the chance of it being needed for querying and reporting is small to none. However, the data still needs to be saved for regulatory compliance and legal requirements. This is the *archive* state. It is the requirements for data in this state which this article addresses.

Finally, after a designated period of time in the archive, the data is no longer needed at all and it can be discarded. This actually should be emphasized much stronger: the data **must** be discarded. In most cases the only reason older data is being kept at all is to comply with regulations, many of which help to enable lawsuits. In other words, when the legal mandate to retain data expires, in many cases the data ceases to be an asset and becomes a liability. So when there is no legal requirement to maintain such data, it is only right and proper for organizations to demand that it be destroyed.

Don’t think in terms of databases or technologies that you already know when considering these data states. The data could be in three separate databases, a single database, or any combination thereof. Furthermore, don’t think about data warehousing in this context – here we are talking about the single, official store of data – and its production lifecycle.

From here-on out we will use the terms introduced here for the various states of data throughout its lifecycle, with the emphasis being on archiving database data and the issues arising from doing so.

What is Database Archiving?

Database Archiving is part of a larger topic, namely Data Archiving. Data exists in many formats and for many purposes, and only a small percentage of it is actually in a database. Physical documents, electronic documents, computer files and data sets, e-mail, and multimedia files are all examples of data that may reasonably need to be archived at some point. Refer to Figure 2. Each of these "things" needs to be archived to fulfill regulatory, legal, and business requirements.

But each type of data requires different archival processing requirements due to its form and nature. What works to archive e-mail is not sufficient for archiving database data, and so on. In other words, type of data may need to command its own technology. This is most certainly true for database data. Why?
Well, data stored in a database is different than other types of data in many ways. The main advantage of using a DBMS is to impose a logical, structured organization on the data. A DBMS provides a layer of independence between the data and the applications that use the data. In other words, applications are insulated from how data is structured and stored. The interface to the data is through the DBMS data language, whether it is SQL for relational databases, DL/1 for IMS, or even XQuery for XML databases. So the archival of data from a database requires knowledge of, and operation in conjunction with, the mechanisms and interfaces of the DBMS.

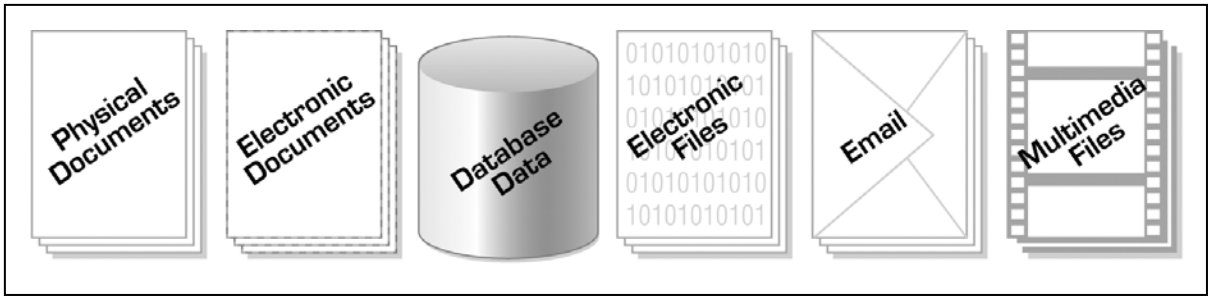

Figure 2. All Types of Data Need to be Archived

OK, if we now accept that database archiving is a subset of data archiving, let's define exactly what we mean by the term. ***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.

Let's examine each of the major components of that last sentence. We say ***removing*** because the data is deleted from the operational database when it is moved to the data archive. Recall our earlier discussion of the data lifecycle. When data moves into the archive state, query and access is no longer anticipated to be required.

Next, we say ***selected records***. This is important because we do not want to archive database data at the file level. We need only those specific pieces of data that are no longer needed for operational and reference purposes by the business. This means that the archive needs to be able to selectively choose particular pieces of related data for archival… not the whole database, not an entire table or segment, 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 (tables for DB2 or Oracle; segments and/or databases for IMS).

The next interesting piece of the definition is this: ***and storing them (the data) in an archive data store***. This implies that the data is stored separately from the operational database and does not require either the DBMS or the operational applications any longer. 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. The purpose may be external, in the form of a lawsuit or to support a governmental regulation; or the purpose may be internal, in the form of a new business practice or requirement. At any rate, the data needs to be readily accessible in a reasonable timeframe without requiring a lot of manual manipulation. I mean, let's face it, anyone can archive data if they don't have to worry about how to query it later, right?

Summary

As regulatory burdens continue to increase and organizations find themselves faced with ever more onerous data retention requirements, a robust, enterprise-level database archiving solution will become a required component of a comprehensive data management policy. Be sure to completely review and understand all of the requirements for long-term data retention before implementing a database archiving solution.

So, what do you think? Does your organization have the technology and resources at its disposal to archive your database data in accordance with legal requirements?