Proven strategies for archiving complex relational data

Contents

Executive summary

Continued database growth, long-term data retention regulations and storage requirements are increasing operational costs. As a result, more CIOs are examining the potential benefits of implementing cost-effective strategies for managing data and information throughout its lifecycle. Because most of the historical data stored in high-performance application databases is inactive, more companies are realizing the value of database archiving as an essential component in an effective data management strategy.

Allowing application data growth to continue can degrade performance, limit availability, increase costs and slow disaster recovery. Yet, archiving complex relational data poses many challenges. Organizations need the capability to archive historical data and safely remove it from the production environment. They also need capabilities to store archives cost-effectively, retain access and dispose of data appropriately to comply with data retention regulations. In short, organizations need an effective archiving strategy for managing application data throughout its lifecycle.

In searching for the ideal archiving solution, some organizations consider in-house development, but soon realize that the short- and long-term costs do not offset the expected return on investment. Implementing an off-the-shelf archiving solution does offer advantages in faster acquisition and implementation. However, the ideal solution must provide the archiving capabilities needed to support current and future data management requirements across often heterogeneous application environments.

This white paper explains how the IBM® Optim™ Data Growth Solution provides proven technology and comprehensive capabilities that make archiving complex relational data both practical and desirable. Archives can be easily managed and stored on the most cost-effective media, and historical data remains accessible until retention periods expire. Maintaining only current active data in production environments improves service levels, increases availability and improves disaster recovery initiatives. Archiving allows organizations to reclaim valuable capacity and lower costs.

As one of the solutions in the IBM Optim Integrated Data Management (IDM) portfolio, the IBM Optim Data Growth Solution addresses a critical operational need for enterprises that rely on large complex relational databases.

Database growth: impact and opportunity

Industry analysts and leading computing publications report that the rapid accumulation of data worldwide continues at an unprecedented rate. Forrester estimates that on average, data repositories for large applications grow by 50 percent annually.[1] Enterprise databases are doubling and even tripling in size and exceeding projected estimates, resulting in serious consequences.

Accumulating more data means degraded performance and slower response time. Availability is limited because routine database maintenance requires more processing time, often taking systems out of service. Managing larger databases also has an impact on critical disaster recovery time. If your company relies on data-intensive, mission-critical business applications, then unmanaged database growth directly affects your bottom line, as you risk losing customers, frustrating users and missing new revenue opportunities.

Managing and storing enterprise data are becoming top priorities for IT executives. CIOs face an increasing demand for 100 percent availability of mission-critical applications, 24/7 eBusiness support and scalable access to enterprise data on a variety of platforms. However, corporate databases are growing so large that just maintaining traditional service levels has become a demanding challenge for many IT organizations, and Forrester Research estimates that 85 percent of data stored in databases is inactive.[2]

So, what happens when accelerating database growth is allowed to continue?

For most companies, the solution to exponential database growth is to upgrade servers and acquire more storage capacity. Faster, more powerful processors can speed access to information, while the underlying databases continue to grow. However, this approach is rapidly losing its viability for many reasons.

Poor performance and response time. As database size increases, the performance of mission-critical applications deteriorates. The larger the database, the more time it takes to load, unload, search, reorganize, index and optimize. Response time slows. Access to decision-making information becomes more difficult. Service levels decline. In the past, IT could count on a linear increase in capacity to keep the service levels within an acceptable tolerance. However, with the overwhelming demand for high-performance systems being more critical than ever, IT departments are requiring more frequent and larger increases in capacity to satisfy the demand.

Intensive database tuning provides another approach to improving performance. Often, a DBA can achieve better performance by reorganizing the database, adding indexes, implementing partitioning and even de-normalizing parts of the database (though this last option may create costly ripple effects on application maintenance). However, as databases continue to grow, the effort yields less return and only postpones the inevitable need for an effective, long-term solution.

Limited system availability. Managing more data means that maintenance tasks require more time to complete. The corporate data explosion has stretched backup and reorganization windows to the point where data availability is seriously threatened. In fact, most companies struggle to complete their scheduled backups. Without a way to safely remove older data from production databases, this problem will only worsen over time.

Increasing costs. Purchasing capacity upgrades might appear to be cost effective as a "one-time" fix, but upgrades are needed on an ongoing basis to keep pace with database growth. As a result, IT organizations often spend millions of dollars in hardware and software license fees per year, just to expand server, storage and CPU capacity. Still, this short-term tactical approach is just a temporary solution for a much larger problem that is not going away — excessive data volume that continues to increase.

Slow disaster recovery. Many companies have made disaster recovery a high priority. In the event of a disaster, the key strategy is to get mission-critical systems operational as quickly as possible. With overloaded databases, all the data (including years and years of historical information) must be restored, just to get business-critical data back online. Restoring large volumes of rarely accessed data, in addition to the critical operational data, can slow the recovery process by hours or even days.

Database archiving: the challenges

Given the magnitude and the impact of database growth on performance, availability and cost, why are IT organizations reluctant to remove rarely-accessed data from production databases and store it separately? Database archiving may seem like a common sense solution, but it is not without challenges and risks.

While most applications have excellent strategies for validating and updating data, few provide a method for safely removing inactive data from production databases when it is no longer needed. Even if data could be safely archived, any subsequent need to access or restore the data can pose a major challenge, and many regulations require that archived data remain accessible.

Archiving referentially intact subsets of data. Archiving production data presents significant technical challenges. This is particularly true when the data is stored in one or more complex relational databases. For example, if data is referenced from two different databases, it must be archived together, while maintaining relationships. A capable database archiving solution must manage complex relational data and keep that data referentially intact.

When the data model is straightforward, creating referentially intact subsets of data is relatively simple. However, the databases powering most of today's enterprise applications rely on highly complex data models. Often, data is normalized across hundreds of related tables with relationships that include recursive relationships, bi-directional relationships and other structures that are challenging to traverse. To complicate matters, these relationships may be managed by the application rather than by database referential integrity (RI) rules.

Safely removing data from production. To remove data from relational databases safely and accurately, an effective database archiving solution must have capabilities for:

- Processing recursive or complex relationship cycles.

- Managing relationships based on partial columns, concatenated columns or data-driven relationships.

- Traversing relationships from a child table to a second, different parent table and so on.

- Performing all of the above tasks without violating referential integrity, creating orphaned data or corrupting the database.

When copying data, the IT staff is typically concerned about selecting a complete and accurate set of related rows to transfer. In addition, they worry about inadvertently removing production data that an application may still need, or worse yet, accidentally corrupting the database without an easy way to restore the data. So again, it may seem easier to let large databases continue to grow because no one wants to risk breaking a database that works, particularly when it supports a strategic application.

Retaining access to archived data. Many companies realize that they must be prepared for any request to access archived data. For example, if the need arises, a company may need to protect its interests in a lawsuit by producing records of financial transactions from previous years. In other cases, companies must comply with government regulations and legislation.

The major concern is that once data is archived and removed from the production database, it may be difficult to locate and access that data quickly when needed. Practitioners and business users need the capability to access archived data in its referentially intact, business context without having to restore it. They also need the capability to restore only the desired data. It should not be necessary to restore massive amounts of data to access a few required rows. Without reliable access and restore capabilities, companies find it easier to continue with traditional approaches for managing their data — allowing production databases to grow rather than risk "losing

the data."

Complying with data retention regulations. Data retention regulations require that companies across industries retain certain types of data for specific periods of time. For example, the Health Insurance Portability and Accountability Act (HIPAA) requires record retention for a minimum of six years and two years after a patient's death. In the financial industry, the Sarbanes-Oxley Act requires record retention for seven years after the end of the fiscal year. The cost of retaining large volumes of historical data online in production databases can be prohibitive; yet, the penalties for non-compliance are severe. The ideal database archiving solution must ensure easy access to the archived data on demand.

Managing data throughout its lifecycle. Database archiving enables managing and storing data, based on its business value and access requirements over time. To begin developing an archiving strategy, business users must define the value of each set of data at various points in the lifecycle. Data can be separated into access profiles, for example, acquisition, heavy access, medium access, rare access and disposal. By implementing database archiving and storage strategies that meet accessibility requirements, companies can reduce the cost of managing and storing data, while ensuring compliance.

Database archiving is essential for any data management strategy. The ideal solution allows for archiving rarely accessed data from a production database and saving it to a separate medium. At the point of diminishing returns, when the cost to provide optimal production-level access and performance outweighs the actual business value derived from the data, it makes sense to move your data out of the high-cost, fast response system and into lower-cost, slower response systems to better match its business value.

Finding the ideal solution

Many IT executives are already convinced that an effective database archiving strategy makes sense, and they are actively searching for viable solutions. The alternatives include building and maintaining a custom database archiving solution in-house or purchasing "off-the-shelf" database archiving software.

Developing an in-house archiving solution. Faced with the need to archive data and the limited number of proven database archiving solutions in the marketplace, many IT organizations initiate projects to develop in-house solutions. Even with adequate time and resources, designing, developing, implementing and maintaining a custom archiving solution for a particular application is an extremely labor-intensive and complex project.

Meeting the minimum requirements is just the first step in enabling an enterprise database archiving strategy. Additional challenges include enhancing and maintaining the custom archiving solution to support ongoing changes to the application and the database. Few business applications remain static over time and in some cases, database vendors or platforms may change, further complicating the matter.

Finally, there are two questions that must be answered before an IT organization can determine the full impact of developing a custom archiving solution in-house:

- How will the custom solution be developed to support the initial application and all other applications that could benefit from archiving? For each application, the IT staff must not only write and debug software to archive subsets of data with 100 percent accuracy, it is also necessary to design and develop a means for researching and restoring archived data.

- Would the company derive greater benefit if IT time was devoted to other business initiatives? Developing an internal database archiving solution diverts highly skilled IT resources from contributing to the direct business focus of the enterprise and the bottom line.

Implementing off-the-shelf database archiving software. This alternative offers a cost-effective approach that directly addresses the problem of database growth without diverting internal resources from IT business initiatives. The ideal database archiving solution must have the flexibility and enterprise scalability to be implemented with any application, custom or packaged, across your enterprise, without writing and maintaining program code.

Advanced technology for archiving complex relational data

Off-the-shelf database archiving technology, such as the IBM® Optim™ Data Growth Solution, delivers comprehensive archive, restore and storage capabilities that can be implemented quickly and easily customized. Optim's archiving capabilities:

- Preserve the business context of the archived data, even for the most complex relational data model.

- Include a "selective delete" feature to accurately remove precise sets of related data from a production database, leaving all other data intact.

- Offer a preview or "deferred delete" capability to prevent removing data accidentally.

- Support intelligent indexing, providing a sure and fast way to identify archived data for easy retrieval.

- Provide a fast and easy way to locate specific data across all archives and allow for browsing archived data directly.

- Support complete integration of the archiving strategy with any application, while meeting business requirements.

- Provide a "selective restore" capability that allows for accurately restoring only a few rows of data, as needed.

- Support multiple databases and operating platforms, including federated access and cross-database interoperability.

Using Optim, companies can archive and remove precise sets of rarely accessed data, save that data in an archive, and keep it "active" for easy access when needed. Mission-critical data remains in the production database, available at peak performance. Application performance and availability improve significantly, while reducing the total cost of ownership (TCO).

In addition to archiving data, referentially intact and complete, Optim provides built-in capabilities to satisfy the requirements for researching and selectively restoring archived data with 100 percent accuracy, regardless of changes to the data model. And there is no need to write, debug or maintain complex custom programs.

Optim represents a major advance by preserving the business context of archived data and making that data readily available for continued use. Archived data can be stored on the most convenient and cost-effective medium: online in archive database, near-line on a file server, offline to tape, disk-based WORM device or long-term permanent storage devices.

Crossing database and application boundaries. Optim provides proven database archiving capabilities for both the mainframe and distributed environments and supports the leading database management systems (IBM DB2® for Linux®, UNIX®, Microsoft® Windows® and IBM zOS®, IBM IMS™, IBM VSAM® and sequential files, Teradata®, Adabas®, IBM Informix® Dynamic Server, Oracle®, Sybase®, Microsoft SQL Server® and XML). Implementing an effective database archiving methodology allows you to archive data across applications, databases or platforms — all with a consistent interface and toolset that minimizes training and support requirements:

- Cross-database flexibility allows you to archive data from one database to another; for example, archive data from DB2 and restore to an Oracle database. You can even archive data concurrently from two or more different database management systems to create "federated" archives.

- Cross-application support allows you to archive data from any number of your database applications and save the data in an Archive File that can be stored on a variety of cost-effective storage media and accessed by any application.

- Cross-platform support allows you to archive and restore data to and from mainframe and open systems applications.

Optim resolves the problems associated with streamlining large relational databases for any application and/or environment. To address database growth issues that affect the performance of some of the leading CRM and ERP business applications, Optim also provides "out-of-the-box" database archiving solutions for SAP® Applications, Oracle® E-Business Suite, PeopleSoft® Enterprise, JD Edwards® EnterpriseOne, Siebel® CRM and Amdocs® CRM.

Phased implementation. Most companies implement the IBM Optim Data Growth Solution using a phased approach, with the overall corporate objective to employ a strategic database archiving solution across the enterprise. The basic implementation strategy includes:

- Analyzing your business rules and objectives, archive requirements and policies, access and storage requirements.

- Identifying the data that can be archived to meet your business objectives. This process includes analyzing the effects of archiving application data and identifying customizations that may be necessary.

- Verifying the archive process, which includes selectively archiving and removing data from a production database, as well as accessing, researching and restoring archived data, as needed.

Once the archiving policies are established for a selected data-intensive application, much of the process is automated. As database archiving continues on a routine basis, Optim provides tools that help determine the optimum amount of data to remove. This systematic approach can then be extended to other mission-critical applications.

Typical database archiving scenario

Retaining the referential integrity and business context of your data is essential for archive and restore processing. Anything less than 100 percent accuracy is unacceptable, especially for data retention compliance and satisfying audit requirements. Optim's unique Relationship Engine™ technology offers proven capabilities for archiving precisely the right data, based on your specifications, with 100 percent accuracy — no matter how many tables or relationships are involved — and keeps that data referentially intact.

Defining the data to archive. Creating an Access Definition identifies the data to be archived from specific tables, including the relationships needed to maintain referential integrity. These specifications can include archiving policies (selection and date criteria) and user-defined indexes to speed access and restore processing at a later time. The capability to specify "archive actions" creates a tighter level of integration between database archiving and the application.

For example, a company may want to archive all information relating to customer accounts that have been inactive for at least a year or more. The archive must include all identifying information, as well as order and payment history, and the data must be extracted from many interrelated tables (see Figure 1).
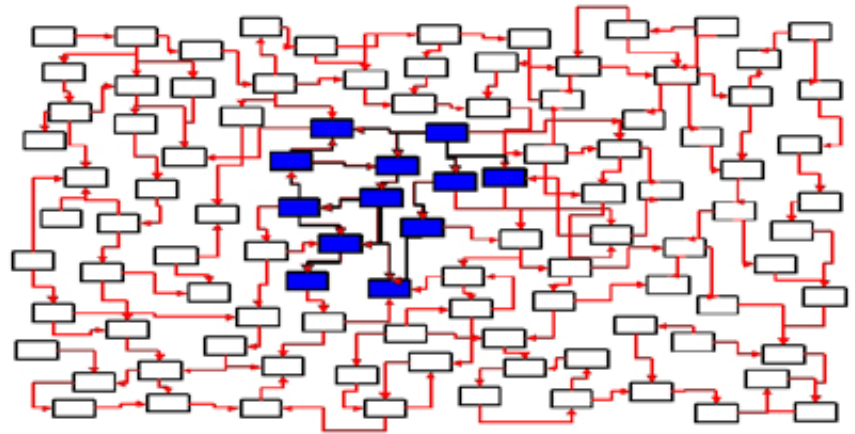

Figure 1. Define the set of related data to archive.

Copying data to an archive. The Access Definition that specifies the data to be archived is used in the archive process to copy the data to an Archive File (see Figure 2). During this process, the specified archive indexes are created automatically.
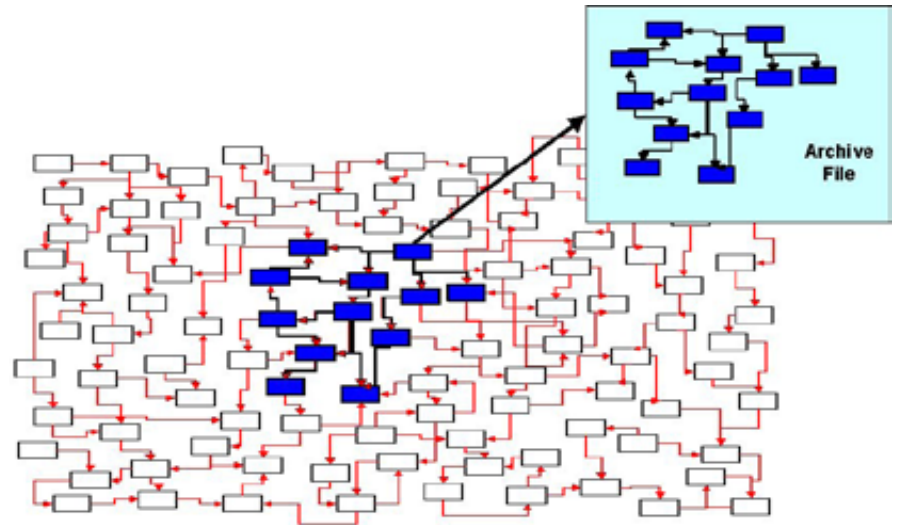
Figure 2. Archive the precise subsets of data based on user-defined criteria.

Optim saves not only the data, but also the metadata describing tables, columns and relationships used to create the Archive File. With this information, archived data can always be restored in its business context weeks or years later — even if the data model changes over time.

Removing data from production databases. IT organizations are reluctant to remove production data from complex relational databases, mainly because accidentally deleting essential data could bring mission-critical systems to a halt. For this reason, it is often considered "safer" to let databases grow than to risk corrupting the data using a delete program that does not ensure accurate results.

Optim addresses this concern by providing options for removing archived data from a production database. Because Optim solutions are "referentially aware," authorized users can remove small portions or entire sets of related data, quickly and accurately. In addition, Optim applies standard database processes for all operations. DBMS security rules are never bypassed.

Selective Delete processing. Optim's Selective Delete capability provides a powerful and safe way to remove specific subsets of the archived data from a production database, leaving all other data complete and intact. Selective Delete offers maximum flexibility during the delete process. For example, an IT organization may decide to archive all order and payment history for customers whose accounts have been inactive over the last year.

Additionally, the master account information for each customer, such as names and addresses, must be retained in the production database. The archive can be created to include all identifying data for each customer. At the same time, Selective Delete can be directed to remove only the related rows from the Order and Payment History tables, leaving the master account information intact.

Immediate or Deferred Delete processing. An archive process request includes options for archiving and removing data from a production database immediately after the data is archived. As an added verification measure, delete processing can be deferred allowing time to review and confirm that the appropriate tables are selected.

Deferred Delete is a relational delete capability that offers advantages when IT organizations want to verify the data before it is deleted, or want to include delete processing in their routine database maintenance tasks. The Deferred Delete can be used at any point after an Archive File is created. The relational delete process uses the specifications defined in the Access Definition to remove data from the production database, complete and referentially intact.

Researching and restoring archived data

IT organizations are often concerned that it will be difficult to locate or restore data after it is archived. In addition, as more data is archived over time, the ability to manage the archived data can pose another maintenance challenge.

Managing archives. Optim provides a user-managed Archive Directory for grouping and maintaining archives. Familiar file maintenance features allow IT organizations to set customized standards for managing archives, making the data readily accessible.

Accessing archived data. Optim provides comprehensive capabilities that allow easy access to archived data, ranging from built-in viewing and reporting functions to programmatic data access. Database archiving processes can be tailored to your company's specific data access requirements, taking the worry out of finding specific data among numerous large archives.

The browse capability allows for viewing archived data in its relational or business context. Using simple or complex selection criteria pinpoints the exact data needed. The archive indexing capability ensures the archives can be searched quickly, without any impact on the current production system and without restoring any data. Archives can even be converted into the industry-standard, comma separated value (CSV) format, providing additional flexibility for accessing the archived data.

Optim can be seamlessly integrated with your applications to provide on-demand access to archived data. For example, in distributed environments, data access is made possible using the industry-standard ODBC and JDBC interfaces; an API enables data access in mainframe environments. For the fastest access, archived data can be stored in a secondary database.

Optim's out-of-the-box database archiving capabilities offer seamless, real-time access to archives. End users can browse and restore archived data transparently from within an application using familiar menus. This level of integration requires little or no end-user training or DBA involvement.

Restoring archived data. No database archiving solution would be complete without the capability for restoring archived data when the need arises. But data recovery is often difficult because archived data from hundreds of tables may need to be reassembled and restored, based on complex relationships. Then, the right set of rows must be identified and selectively restored. To complicate this scenario, application data models can change over time, making any restore effort more challenging. Without the capability to selectively restore data, you risk losing critical access to your archived data.

Optim allows users to locate and restore specific data in any number of archives and then restore it — referentially intact and complete. Optim supports a variety of restore options, including restoring archived data to the mirrored tables in the production database, back to the production database, or more commonly, to a separate database for review and reporting purposes.

The Selective Restore capability allows users to restore subsets of archived data selectively, using specific criteria that may differ from the criteria used to create the archive (see Figure 3). For example, suppose that a company archives transaction and payment data, where the last transaction date is older than two years. Seven years later, the company needs to restore the data, but only for customers from New Jersey. Although the data was archived using date criteria, it may be restored using geographic criteria.
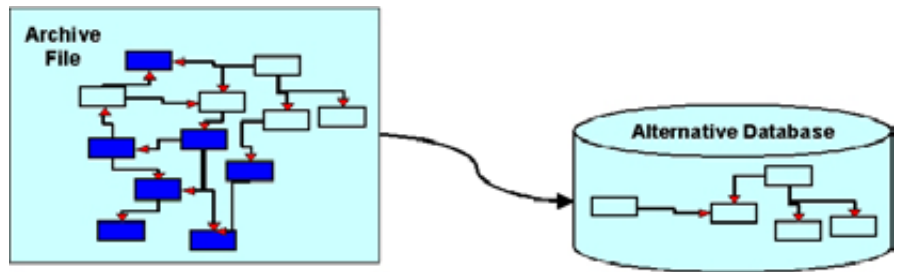
Figure 3. Selectively restoring data from an archive.

With Optim, it is not necessary to restore an entire archive to obtain only the small portion that is needed. Users can find and quickly restore the specific subset of transaction data.

IBM Optim delivers business value

IBM Optim Integrated Data Management (IDM) solutions support a scalable, modular data management framework designed to increase organizational productivity. Organizations can also look forward to improving the quality of service, lowering the cost of ownership and supporting the governance of data, databases and data-driven applications.

As one of the solutions in the Optim IDM portfolio, the IBM Optim Data Growth Solution addresses a critical operational need for enterprises that rely on large complex relational databases. Overloaded databases can be reduced by up to 50 percent or more during the initial archive. Database and application performance improve and valuable processing power is no longer required to support inactive data. Routine archiving continues to deliver long-term benefits. Archived data can be browsed directly and selectively restored as needed.

Optim offers a value proposition that significantly lowers your total cost of ownership by optimizing your company's multi-million dollar investment the relational databases that support your mission-critical business applications:

- Deploy a single, comprehensive enterprise archive solution to solve the problem of database growth across all applications in the enterprise.

- Maintain acceptable service levels by minimizing production database searches and improving response time.

- Reduce database maintenance time, reduce the downtime for application upgrades, and shorten batch process windows to increase application and database availability.

- Defer capacity upgrades and the associated expensive hardware and software license fees.

- Store archived data cost-effectively on a variety of storage media and retain easy access to archived data no matter where it is stored.

- Speed disaster recovery time by hours or days by keeping only business- critical data in production databases.

Using the IBM Optim Data Growth Solution to manage continued data growth, your organization can deploy a strategic database archiving strategy across the enterprise.

About IBM Optim Integrated Data Management Solutions

IBM Optim Integrated Data Management Solutions offer proven, integrated capabilities to manage enterprise application data from requirements to retirement. With Optim, teams can share data artifacts (like models, policies and metadata) to align data management with business goals and improve collaboration. Today, organizations of all types leverage Optim to improve performance, streamline database administration, speed application development, and enable effective governance. Optim delivers better business outcomes, at lower cost, with less risk, while providing capabilities that scale across enterprise applications, databases and platforms.

For more information

To learn more about IBM Optim Integrated Data Management Solutions, contact your IBM sales representative or visit: **www.ibm.com/software/data/optim-solutions/.**

IBM®