Benefits of data archiving in data warehouses
Contents
2 Executive summary
3 Typical reasons for rapid data growth
4 Challenges associated with data warehouse growth
5 Traditional data growth solutions that do not work
6 Understanding data archiving
9 Benefits of data archiving
10 Guiding principles and technology requirements
11 Managing data growth responsibly with data warehouse archiving

Executive summary
Data warehouses are the pillars of business intelligence and analytics systems, often integrating data from multiple data sources in an organization to provide historical, current or even predictive analysis of the business. Information from multiple internal or external transactional systems is extracted, transformed and loaded into data warehouses as atomic data. This cumulative data and the analytics systems that leverage it provide the technology and methodology that help organizations discover and develop meaningful insights.

Due to the consolidated nature of data warehouses, these data stores often suffer from rapid growth. Typical reasons for this phenomenon include expansion of data warehouses with new subject areas or data marts, compounded data growth from organic or inorganic business growth, or a “let’s keep it all, someone might need it” attitude toward historical data.

This unchecked data growth often results in ever-increasing infrastructure and operational costs, poor data warehouse performance, and an inability to support complex data retention and legal hold requirements.

A data archiving solution helps organizations address these challenges by allowing IT staff to intelligently move (and purge) historical and inactive data from production databases into a more cost-effective location while still providing the capabilities to query, search or even restore data if needed.

A tiered archiving strategy provides additional benefits in terms of managing performance and cost-effectiveness. Data archiving can also alleviate data growth issues by:

- Removing or relocating inactive and dormant data out of the database to improve data warehouse performance
- Reducing the infrastructure and operational costs typically associated with data growth
- Leveraging proven policies and processes to cost-effectively manage multi-temperature data
- Improving disaster recovery and backup/restore plans to consistently meet service-level agreements (SLAs)
- Supporting compliance with data retention, purge or hold policies

This paper describes a data lifecycle management strategy for data warehouses that is designed to manage high-volume data growth cost-effectively, and avoid performance degradation.
Typical reasons for rapid data growth
The data warehouse is commonly an organization’s largest database. This is due to several factors:

Big data and the explosion in data volume: With the advent of big data technologies that help organizations generate insight from large information assets, companies are keeping unstructured and structured data that might have been thrown away in the past. Apache Hadoop and similar technologies continue to gain momentum and adoption, and will provide new ways of processing large amounts of such data, extracting intelligence from multi-structured data sources, and integrating the results into existing data warehouses for further analysis and reporting.

The “data tomb” effect: Data warehouses may become the dumping ground for historical data from various transactional systems, with little regard to the true value of the business intelligence within this dead data. This “data tomb” effect may be caused by the lack of an optimal archiving and data retention strategy in the originating transactional system itself.

Expansion into new subject areas: Companies frequently expand data warehouses with new subject areas and new data sources, making them part of a central repository for the enterprise or interconnected data marts. While this expansion can provide insights for crucial business activities, it can also lead to significant data expansion.
**Business growth:** Larger organizations are often subject to compounded data growth from mergers and acquisitions, as well as organic business growth. Consolidation of multiple implementations into one results in a larger system.

**Lack of retention and disposal policies:** Unfortunately, the business side of an organization may not provide IT teams with enough clarity on data retention and disposal policies. Most organizations have a “let’s keep it all, someone might need it later” mentality for historical data, which prevents them from exploring cost-effective data retention, hold or purge processes.

Each of these factors provides an impetus for IT organizations to adopt data lifecycle management strategies and efficiently manage categories of data according to their value in a data warehousing architecture.

**Challenges associated with data warehouse growth**

High-volume data growth and large warehouse implementations present multiple IT challenges and business risks. While many data warehouse solutions and architecture choices exist in the market, every approach poses several common challenges (see Figure 1).

**Cost of ownership**

The impact of exponential data growth on infrastructure and operational costs can be huge, often taking up most of an organization’s data warehousing budget. Larger amounts of data require larger capacity, resulting in more hardware and storage requirements—as well as higher costs to maintain, monitor and administer this infrastructure. Large data warehouses generally require bigger servers and appliances, which may also increase software licensing costs for the database, database tooling, integration or business intelligence (BI) tools.

![Performance and capacity challenges associated with data warehouse growth.](image-url)
In addition, IT departments must factor in the costs of a mirrored disaster recovery system, the data backup infrastructure, processes to copy large data sets within the SLA window and replicas of the database across test environments.

**Performance and availability**
Large volumes of data and varying workloads can put a lot of stress on data warehouse systems. With a majority of production data typically in an inactive state, the performance and system availability of data warehouses suffer greatly as a result of unchecked data growth.

When the response time of critical queries and reporting processes starts to degrade, extract/transform/load (ETL) loads take longer and may extend past the SLA windows. Database backups run endlessly and the IT staff must operate in reactive mode to contain these issues. These situations pose a significant risk to business continuity and system availability, because downtime can result in a lengthy system recovery period.

**Cost-effective compliance**
Many data warehouses also feed data back into the transactional systems, acting as systems of record in these cases. These systems may be subject to audits, retention, legal hold or e-discovery requests. Simply purging historical data is not acceptable as a method for keeping up with data growth because compliance regulations may require data to be retained for a certain number of years, put on legal hold to satisfy discovery requests, or audited. Keeping all of the data in production databases is not a cost-effective way to retain data for compliance reasons. Also, if a data warehouse was used to make business decisions, it may be targeted for legal disclosure under e-discovery rules.

**Traditional data growth solutions that do not work**
IT organizations may try to use conventional methods for managing data growth, but these methods are habitually ineffective or fail to generate a cost-effective solution. Common techniques include:

**Hardware upgrades:** Trying to keep up with data growth has a huge impact on capital expenditure and frequent hardware upgrades. The traditional solution is to add more server nodes, or perform forklift upgrades to replace the data warehouse infrastructure. While hardware upgrades are inevitable, there are other ways to defer these costs and reap better performance from existing infrastructure—which may amount to huge savings.

**Traditional backups:** Large, monolithic backups are highly redundant with historical and inactive data taking up most of the space. Backups are not substitutes for archives; archives are online or near-line and queryable. Backups cannot solve data growth problems because they require creating a replica of the production data, and need to be taken frequently (on a weekly or monthly basis), which adds more overhead to the growth problem. If IT teams use backups to archive data, it can be difficult to retrieve the data within a short period of time. Information retrieval also poses a challenge when the data schema in the original system has evolved.
**Database partitioning:** IT departments sometimes try to manage data growth by implementing a partitioning schema in the traditional database management system (DBMS) to separate active data from historical data. However, partitioning in this way still may not reduce the overhead on the database because the indexes remain the same size. Partitioning does not help reduce overall storage costs and maintenance windows; it also makes it difficult to restore or re-create selective data records located in a dropped partition from the time when the database was on an older version. Certain analytical DBMSs don’t even support database partitioning.

**Homegrown solutions:** Building a mature data archiving and purging solution in-house can be a very expensive and time-consuming effort. The scripts and code require proper handling of database referential integrity, error recoverability, high-performance execution and consistent application of business rules and policies across a potentially large number of systems. Despite the huge investment, these solutions are hard to maintain and do not provide much longevity in typical organizations where people and technology change regularly.

**Purging data:** In many industries, companies must keep large amounts of historical information (especially financial information) for compliance reasons. Data is subject to the same SLAs—including those for data retention—as the transactional system itself, and for that reason must be covered by information lifecycle policies for standard corporate data.

**Understanding data archiving**

Data lifecycle management is a policy- and process-oriented approach to efficiently control the flow of an information system's data throughout its lifecycle, from requirement to retirement. Data lifecycle management policies include ensuring optimal application performance and archiving historical data to manage data growth while ensuring access to both production and archived data. Before archiving data, it is important to classify everything based on usage activity.

**Data assessment and classification**

It is not uncommon for organizations to have millions or even billions of records across different fact tables that hold many years of accumulated information. However, it is quite common for users and DBAs to find that the most active data is typically located within the last six months to two years of transactions. Anything earlier is queried infrequently.

Data in the warehouse can be classified according to its temperature—the access frequency, volatility and query performance of the data. Hot data is frequently accessed and updated, and users expect optimal performance when accessing this data. As data ages, it tends to “cool off,” meaning that the probability of users accessing this data significantly decreases.
Archiving typically targets cold data and relocates it to a more cost-effective storage medium (see Figure 2). However, the data must still be available for regulatory requests, audits and long-term analysis—so the archived data should be queryable and restorable (in the original location or a staged location). Data assessment and classification based on business usage is an important factor in an effective archiving strategy.

**Data archiving**
Archiving in its simplest form involves the migration of information or data (typically historical) from an online application to a secondary (online, near-line or offline) system, making it accessible as a long-term storage repository. As a recognized information lifecycle management best practice, archiving segregates inactive application data from current activity and safely moves it to a different tier based on its value to the business. Consequently, smaller databases tend to deliver higher service levels with lower maintenance and operational overhead.

*Figure 2.* Multi-temperature data classification based on access requirement.
**Archiving in dimensional data warehousing or data marts**

Data warehousing uses different methods of data modeling. One popular approach—dimensional data warehousing—involves fact and dimension tables, whereas others use a more normalized data model. There are two types of history tracking in dimensional data warehousing:

1. **Fact data changes**: Granular fact records about a business event (such as a sale or transaction) are linked to a certain point in time, which are history-tracked and grow in large numbers over time. These high-volume, historical and detailed records are good candidates for archiving.

2. **Dimension data changes**: Data in dimension tables may also change over time and is known as slowly changing dimension (SCD) data. In this case, attribute changes in a dimension such as customer phone number or address may be tracked, can change over time and often result in a sizeable amount of historical data. The larger the dimension tables in volume and number of attributes, the larger the data grows in SCD records. However, fact record growth is higher than SCD records.

**Tiered storage archiving strategies**

Database archiving involves extracting a predefined set of historical data (often time-based) from a set of tables while maintaining its data referential integrity; moving this data set into either a secondary archive data warehouse or a file-based data archive; and purging the historical transactional data from the source database. For higher query performance and access to larger data volumes of data, warm data may be stored in another data warehouse instance, ideally on a lower-cost infrastructure. For rarely accessed data, storing this “cold” data in compressed and queryable data archive files may provide a more cost-effective solution compared to higher-tier storage.

Organizations may leverage a combination of these archive stores to balance access performance requirements and cost-effectiveness (see Figure 3). The archived systems would leverage lower-cost storage devices such as Serial ATA (SATA), network-attached storage (NAS), content-addressable storage (CAS), optical disks, tapes or cloud storage.

---

**Figure 3.** A three-tier archiving strategy designed to optimize cost-effectiveness and performance of specific data sets on different tiers of storage.
Access to archived data

While archiving strategy and architecture may look different for each implementation, there may be infrequent requirements to access the archived data. Archiving removes data from the production system, but this data is not lost—it is simply relocated based on its business value. In cases where a separate instance of an archive data warehouse is used, the queries could be directed to the archive instance directly. For scenarios where combined reporting with production data is needed, data federation technologies could be leveraged as well.

The data archive files created by the archiving solution should allow access using industry-standard interfaces such as ODBC/JDBC, XML or SQL, via any standard reporting tool. Users can then browse or search the archives using browser-based or other standard reporting mechanisms for auditing or compliance reasons. For heavier analytical requirements on larger sets of historical data, the archiving solution should allow users to restore archived data sets back to the original location or a staged location. In general, because archived data is infrequently accessed, restoring data is rarely required.

Benefits of data archiving

Lower total cost of ownership

Data archiving can have a great impact on reducing total cost of ownership for the data warehouse and help with IT cost-savings initiatives. By deferring hardware upgrades in production and disaster recovery environments, archiving enables companies to make the most efficient use of the existing infrastructure in a controlled data growth environment. Archiving strategies help the amount of data DBAs must actively manage and the amount of time they spend tuning or adjusting storage requirements—freeing them up to focus on more strategic projects. Archiving also holds the potential to reduce software costs (such as warehouse and database licensing costs) associated with larger data warehouses.

By leveraging lower-cost storage tiers or lower-cost data warehouse appliances with a tiered archiving strategy, organizations can purge inactive historical data once it has been archived—reclaiming space in production data warehouse servers. In addition, archiving helps control the cost of capital and operational expenses related to database backup processes because the redundant and static historical data will be reduced in periodic backups.

Improved performance and availability

Archiving and purging inactive data helps significantly improve query performance by reducing the amount of data and the number of indexes and table scans that must be processed. Smaller data warehouses also perform better with batch processing, long-running reports and ETL jobs—avoiding overruns into other production usage requests. Archiving makes performing periodic maintenance tasks easier and faster, and it streamlines restoration from backups in the event of a failure for better system uptime and user productivity.

“Organizations which fail to deploy strategies to address data complexity and volume issues for their analytics by 2012 will experience more than doubling costs of ownership for their data warehouse and mart environments in disorganized attempts to meet this new demand.”

— “Does the 21st-Century ‘Big Data’ Warehouse Mean the End of the Enterprise Data Warehouse?” 25 August 2011, Gartner
**Streamlined risk and compliance management**

Data archiving helps organizations comply with data retention and purge policies while providing queryable archives for audit or e-discovery requests into historical data. The technology and processes also support data legal-hold requirements. Plus, archiving enables organizations to apply business policies to govern data retention and disposal and provides long-term solutions for storing historical data.

**Guiding principles and technology requirements**

An enterprise-grade data archiving solution should meet four key technology requirements:

1. **Enterprise architecture**

Most enterprises rely on heterogeneous information assets, solutions and platforms from multiple vendors. A single, scalable data lifecycle management solution must support all of these major technologies, providing a common and reusable interface and processes. The solution should also be optimized for high-performance connectivity to multiple data warehousing solutions (such as IBM® PureData™ System for Analytics, which leverages IBM Netezza® technology; IBM DB2® and IBM InfoSphere® Warehouse; IBM Informix®; Teradata; Oracle; Microsoft SQL Server; and Sybase) with support for major operating systems including IBM z/OS®, IBM i, Linux, UNIX and Microsoft Windows.

Such an enterprise solution should also support a tiered storage architecture for optimal balance between storage cost, performance and access requirements. Pre-built integration with hierarchical storage management (HSM) systems like IBM Tivoli® or EMC Centera also helps ease implementation of a tiered archive strategy.

**IBM InfoSphere Optim: A single, enterprise-scale data lifecycle management solution**

IBM InfoSphere Optim™ software provides a central data management solution designed to scale to meet enterprise needs. Whether addressing a single application, a data warehouse environment or a global data center, organizations can use InfoSphere Optim solutions to streamline data management with a consistent strategy.

The unique relationship engine in InfoSphere Optim provides a single point of control to guide data processing activities such as archiving, subsetting, migrating and retrieving data. Reusable data management templates enable consistency and scalability, while advanced security features provide support for role-based access and activity permissions.

InfoSphere Optim supports major data warehouse environments, including IBM PureData System for Analytics, IBM InfoSphere Warehouse, Teradata and Oracle. It also supports enterprise databases and operating systems, including IBM DB2, IBM Informix, IBM IMS™, IBM Virtual Storage Access Method (VSAM), IBM z/OS, Oracle Database, Sybase, Microsoft SQL Server, Microsoft Windows, UNIX and Linux. In addition, InfoSphere Optim supports key ERP and CRM packaged applications such as Oracle E-Business Suite, PeopleSoft Enterprise, JD Edwards EnterpriseOne, Siebel CRM, Amdocs CRM and SAP applications, as well as many custom applications.
2. **Complete business objects**

From a database perspective, a business object represents a group of related rows from related tables across one or more applications, together with its related metadata (information about the structure of the database and about the data itself). Capturing the complete business object offers a complete view of the business activity surrounding a particular transaction.

Data warehouses are required to represent these relationships accurately, whether in a star schema, snowflake or hybrid data model. When the high-level entity, such as an order, is archived, the corresponding line items should be archived as well. If this does not happen, then data integrity is lost. Such connections form a complete business object—so enterprises should look for archiving software that represents and preserves such complex entities in a simple, easy-to-manage and high-performing way.

3. **Discovery and understanding data structures**

To archive complete business objects, enterprises need archiving solutions with robust data discovery and metadata mining capabilities. The solution should be able to discover, analyze and document data models with accurate schema and data relationships from data warehousing systems in multiple ways. It should allow IT staff to reverse-engineer a model from an existing source database by mining the database catalog. Without this, the data model representation would have to be built manually. If there is no physical or documented data model representation, the solution should have automated capabilities for analyzing data values and data patterns to identify relationships that offer greater accuracy and reliability than manual analysis.

Archiving solutions must also provide the ability to import an existing logical model and make changes to it for database archiving. They should also provide an easy way for IT staff to incorporate logical data relationships manually for any custom relationships not represented at the physical layer.

4. **Universal access to archives**

The archiving solution should offer universal access to archived data using industry-standard interfaces such as ODBC/JDBC, XML or SQL, and reporting tools using these interfaces, such as IBM Cognos®, SAP Crystal Reports, Microsoft Excel and others.

**Managing data growth responsibly with data warehouse archiving**

Data warehouses should not be allowed to grow into large, expensive historical data repositories. Managing data growth with data warehouse archiving helps reduce costs, improve performance and increase availability for business-critical analytics and BI solutions while maintaining compliance with data retention requirements. Together with IBM, organizations can make a case for archiving in their data warehouse implementations and evaluate the business value of managing data growth.

**For more information**

To learn more about IBM data archiving solutions and best practices, contact your IBM representative or visit: ibm.com/software/data/optim