Manage Data Growth and Optimize the Data Warehouse Infrastructure with Data Warehouse Archiving

By Bill Inmon
## Table of Contents

**Executive Summary** ........................................................................... 2

**The Evolution of the Data Warehouse** .............................................. 3
  - The Data Lifecycle within the Data Warehouse .......................... 3
  - Dormant Data in the Data Warehouse ........................................ 4

**Data Warehouse 2.0** ....................................................................... 6
  - Partitioning Data in the Data Warehouse ................................. 6
  - Using Storage Tiers to Manage Warehouse Data ..................... 6
  - Data Archiving to Optimize Storage Tiers ............................... 8
  - Indexing Archival Data ............................................................. 8
  - The Changing Structure of Data over Time .............................. 9

**Informatica Data Archive™**:  
**The Complete Data Warehouse Archiving Solution** ...................... 11  
  - Robust Archiving Techniques Enable Optimal Storage Tiers ...... 11
  - Multiple, Easy Access Methods to Archived Data ..................... 13
  - Automatic Indexing of Archival Data ....................................... 13
  - Automatic Management of Changing Data Structures ............. 13
  - Universal Connectivity ............................................................ 13
  - Integration with Other Archiving Platform, ECM, and Storage Solutions ................................................................. 13

**Conclusion** .................................................................................... 14

**About Bill Inmon** ................................................................. 15
Executive Summary

The world of data and information has been in a constant state of evolution since the first usage of computers in the late 1950’s. Over time, it became apparent that data, like so many entities, has a lifecycle and unique to each point in the lifecycle, a different set of characteristics, storage, and access requirements. The concept of a data warehouse evolved from the business need for reliable, consolidated and integrated data reporting and analysis from varying points in its lifecycle, across disparate data sources.

While in a gross sense, a data warehouse is simply a repository of an organization’s electronically stored data, it is important to recognize that any warehouse is only as good as the processes to find, access and move items into and out of the warehouse. For data, the essential components of a data warehousing system includes the ability to selectively store data, to retrieve and analyze data no matter where it’s located, and to manage the data dictionary.

Operating an efficient data warehouse requires the organization to understand the differences inherent in the information stored in the data warehouse according to its point within the data lifecycle. As data ages:

- The probability of that data being accessed drops. Simply put, the older data becomes, the less frequently it is used.
- The structure of data changes. As software grows increasingly complex to process and handle more data with greater efficiency, by necessity database architectures change. This is often seen in a steady stream of software releases taking advantage of increasingly more powerful hardware and software technologies.
- The amount of data being stored grows exponentially. Governed by both industry and government regulations, data must be stored and kept accessible for years. While only the first year’s worth of data is actively used, maintaining historical data can easily balloon data storage to as much as 20 times larger than the current production database.

This white paper will address the issues created by a complex data lifecycle within the data warehouse and how data archiving can better manage growing data volumes. By understanding the dynamic forces at work governing the explosion of data volumes in the data warehouse, and the technologies available today to effectively archive and retrieve data based on its point in the lifecycle, the operation and cost of the data warehouse can be made more manageable, productive and efficient.

Implementing robust archiving techniques will provide an optimal and cost-effective archiving infrastructure for the data warehouse that:

- Maintains data integrity across multiple formats
- Enables easy, on-demand access to archival data
- Provides universal connectivity and integrates with multiple archiving platforms to ensure superior and cost-effective scalability and performance.
- Efficiently stores archived data to save storage capacity, while facilitating fast data retrieval
The Evolution of the Data Warehouse

The most important achievement of the data warehouse was the ability to create a platform for integrating corporate data from multiple enterprise applications to facilitate analysis and reporting. This profound transformation allowed the organization to have—for the first time—a single, integrated corporate database. It is this complete set of integrated data that allows an organization to view information enterprise-wide, from a true organizational perspective.

By integrating more and more data from a growing variety of data sources, organizations grew more sophisticated in handling their data, exposing the need for an expanded set of information processing capabilities. From the basic data warehouse, simply a collection of aggregated historical data, the need for a second generation data warehouse architecture and design began to evolve.

The Data Lifecycle within the Data Warehouse

As organizations became experienced with a first generation data warehouse, database administrators noticed that most of the queries were going against the most recent six months worth of data. This first manifestation of a data lifecycle within the data warehouse came with a growing awareness that as data aged, the probability that such data would be accessed dropped. The older data became, the less frequently the data was accessed.

Even more importantly came the awareness that as the data warehouse aged, the volume of data increased. Data in a data warehouse grows at an explosive rate. In the first one or two years of a data warehouse the volume of data often grows at a 200% to 500% rate per year. This rate continues to accelerate until the fourth or fifth year when the data warehouse rate of growth drops to about 100% to 200% per year. But by that time there is already a significant amount of data that has been collected in the data warehouse. For a variety of reasons, the data warehouse caused an explosion of data that was to be managed by the corporation.
This explosion of data in the warehouse has two major impacts:

- The impact on performance as data grows causes degradation across the enterprise, creating bottlenecks that negatively impact each user’s ability to access data in a timely manner.
- The growing cost of adding disk storage and increased cost to maintain an IT infrastructure to support it.

As long as IT organizations can maintain systems with just the relevant amount of current data that is regularly accessed for daily operations, performance is optimal. But as the system accumulates a large amount of historical data with only a small portion of that data being used, performance worsens.

Performance degrades because the system must process and handle large amount of data that is not used. An analogy would be cholesterol in the body. In the circulatory system of the young marathon runner there is very little cholesterol and the young athlete has a very efficient heart. But in a 65 year old couch potato, there is an accumulation of cholesterol causing stress to the heart which has to expend more effort to maintain proper circulation. The same is true of a large data warehouse where the system contains a large volume of unused data. The system has to manage huge amounts of unnecessary data, and in doing so uses machine cycles that would otherwise not be necessary.

By maintaining this explosion of data in a data warehouse, IT infrastructure and maintenance costs grow exponentially even though the percentage of the data that is actually used decreases. What complicates matters is that after a certain volume of data, costs rise dramatically as supporting this data begins to require more than just physical disk. The infrastructure begins to require additional processors, complex disk arrays, additional software, and of course, staff time to operate and maintain the growing systems, causing the associated IT cost to increase exponentially.

Dormant Data in the Data Warehouse

An analysis of usage patterns shows that most queries use only the most current data, with a larger and larger portion of the data warehouse not being used. Within just two years of collecting data, most organizations find that only the first six months is being analyzed, leaving approximately 18 months of data untouched—a trend that continues unabated as data is collected over longer periods. The result is that the vast majority of the data in the data warehouse is simply never touched by anyone.

![Figure 2. As the volume of data grows in the data warehouse, the amount and percentage of data that is actually used drops.](image)
The organization has just discovered what is termed “dormant data”. Dormant data in a data warehouse is like a 2000 pound anchor on a two man rowboat. It simply causes problems far more disproportionately than one would ever imagine. One way to understand the impact of dormant data in a data warehouse is based on the probability of access. In a first generation, mature data warehouse, there is typically some current data that is used very frequently and a lot of data that is rarely or never used at all.

The next stage in the evolution of data warehouse architecture now becomes apparent—it makes both economic and technological sense to move dormant data out of the production system to some other storage media, in a different data tier. There are three main reasons for moving dormant data out of the first generation data warehouse environment:

- The cost of the data warehouse infrastructure is greatly reduced by the movement of data from the first generation data warehouse into another less expensive storage media.
- By moving dormant data out of the first generation data warehouse into the different storage tiers available in the next generation data warehouse, the organization can now handle much larger data volumes than could ever be handled by a first generation data warehouse.
- Performance improves by alleviating the stress created by maintaining a huge database infrastructure.
Data Warehouse 2.0

Based on limitations in the first generation data warehouse, a second generation, DW 2.0, evolved to recognize and support the lifecycle within the data warehouse. There are several substantial differences between the first generation data warehouses and DW 2.0, most notably the recognition that as data ages, its characteristics and access requirements change. As a consequence, the infrastructure in DW 2.0 is divided into different storage types based on the age of the data. Data is first placed on a high performance storage type and is moved over time from this high end storage type to the next lower cost and lower performance storage type based on the probability of data access. This second generation data warehouse recognizes the need for database partitions, indexes and storage tiers.

Partitioning Data in the Data Warehouse

One standard practice for managing the data warehouse environment is the ability to break the archival data up into partitions.

While there are many ways to partition warehouse data, the most common is to divide the data by date. One partition contains the data from 2003, the next partition contains data from 2004, the next partition contains data from 2005 and so forth. This mode of partitioning is natural because the data arrives by date.

Other strategies can also be employed, such as partitioning by organizational unit, by geography, and so forth. And data can be partitioned by more than one set of parameters. For example data can be partitioned by date and geography, or by date and organizational function, and so forth. By dividing data into partitions, data can be structured according to user access patterns. Searches that can eliminate data in multiple partitions at once can be conducted more quickly and efficiently, lowering the cost of accessing data and reducing processing demands.

Using Storage Tiers to Manage Warehouse Data

To further optimize the data warehouse infrastructure, data partitions can be located on different storage tiers, having different performance, access, availability, and cost characteristics, based on the access requirements of the data. There are many reasons for separating a first generation data warehouse into physically separate sub divisions that resides on and is managed on different storage tiers.
The most obvious and compelling of those reasons is economics. By separating out the first generation data warehouse into separate physical storage tiers, the small amount of data that is used frequently resides on expensive high performance disk storage, and the bulk of the data that is not being used resides on less expensive storage media.

Different storage tiering strategies can be employed. One possible strategy is to define storage tiers based on the performance requirements around data access and data updates:

- The interactive tier is where transaction processing takes place. The probability of access of data in the interactive tier is high.
- The integrated tier is the place where corporate data is created. In the integrated tier is found the classical first generation data warehouse. There is a reasonably high probability of access for data found in the integrated tier.
- The near line storage tier is optional. Some organizations need near line storage and some organizations do not. Typically, near line storage is a “cache” for the integrated tier. Data found in the near line storage has a low probability of access.
- The archival tier contains data with the lowest probability of access

However it is done, data needs to be physically removed from the core production data warehouse, where in the context of the above storage tiering strategy, the integrated storage tier would likely reside. This means that the data can be relocated to other storage types such as less expensive disk or file-based storage. However it is done, dormant data needs to be placed on a separate storage medium than that of the core production data warehouse. This leads to the evolution of data warehouse archiving.

Figure 6. One of the benefits of moving data to alternate form of storage is to reduce cost
Data Archiving to Optimize Storage Tiers

Data archiving can be employed to automatically and physically relocate data with lower business value in data warehouses to more appropriate and cost-effective storage tiers. Data can have lower business value based on a number of criteria, such as data access and performance requirements, the age of the data, which region or department the data pertains to, or partition usage. As low access data grows to consume the lion’s share of the data warehouse, the most logical progression is for this data to be physically and logically separated from the core production data warehouse.

Once the organization understands the issues of data management, the related economics, the issues of dormant data, and the evolutionary pressures created by data growth, the conclusion is inevitable that first generation data warehouses evolve to DW 2.0, and in doing so, the archival data storage tier is created.

The archival storage tier in the DW 2.0 data warehouse environment has many different characteristics that set it apart from the other parts of the data warehouse. The probability of access of data in the archival tier is low. Data is normally not updated in the archival environment. Database design may or may not be the same between the two environments.

The major drivers for data warehouse archiving are usually to reduce infrastructure cost by storage tiering, reduce maintenance cost, and maintain peak data warehouse performance. Simply relocating inactive data from the production data warehouses to lower-cost servers and storage achieves many of these goals, but your business requirements are likely to be more complex, such as how you access and retrieve archived data. You need to consider your organization’s budget constraints and performance and access requirements when selecting a data warehouse archiving solution.

Your IT organization will probably access archived data less frequently than active data. But you may still have to periodically retrieve the combined archived and current data directly from the original application interface. In this case, the data should be archived to a format that facilitates relatively high query performance—such as another data warehouse instance, located on a lower-cost infrastructure.

On the other hand, if inactive data is quite old and is ready to be retired, you may have to access it only rarely. In this case, access from a reporting or e-Discovery tool, rather than from an application interface, may be adequate. Slower query performance can be tolerated, and the data may be archived to a more optimal, compressed format, such as a compressed file.

Indexing Archival Data

Another significant component of building a data archiving environment is the practice of creating ‘passive indexes’. In the active parts of the data warehouse, the practice of creating indexes to enhance performance is very common. In the archival environment, however, projecting business requirements for future data access can be difficult. Generally, the archival environment is examined whenever a business need arises. But the business need may not be recognized for 20 years after the archival data is stored. Therefore, the processes used to build indexes in the data warehouse do not apply to the archival environment. To that end, there is the design practice of creating what are called ‘passive indexes’.
Typically passive indexes are built using the likely or possible criteria for fast future retrieval of archival data. Part numbers, customer names, order numbers, phone calls made, and an episode of care—all are likely pieces of data that could be indexed. An analysis of common usage patterns can help determine what data is likely to be referenced in the future. Archiving software should be able to analyze the data and automatically create indexes during the archival process, optimizing it for future access.

The Changing Structure of Data over Time

It is because every organization undergoes change, and every change is ultimately reflected in the data structure, that the database designer expects the data structure found in the archival environment to not remain constant.

As data is added to the archival environment, the issue of managing data stored in different releases of software technology over a long period of time arises. Suppose that an organization starts to store data in the archival environment in 1990 under release 2.0 of a product. By 1996 the data is stored under release 3.1 of the product. More time passes and by 2005 data is stored under release 8.i. In 2010 data is stored under release 11.4.

Such a progression is absolutely normal. The question now becomes—can the current software release read and recognize data that was stored under an earlier release? Usually software vendors can handle the previous release of the software. But when it comes to going back to a software release that is a decade old (or even two decades old), a time comes when a vendor can no longer support a past data architecture while providing the new functionality that has become essential.
There are many approaches to the handling of changing data structures in the archival environment. One essential element of the archival environment is that of metadata—the descriptive information that defines the context and structure of the archival data. Maintaining the right metadata is essential to handling structural changes to archival data. One solution to managing structural changes is to maintain multiple metadata versions corresponding to the structural changes over time. Another solution is to update the metadata periodically to synchronize the archival metadata with the core production data warehouse structure. Regardless of the approach, a data archive solution needs to handle structural changes to archival data based on the evolution of the production data warehouse over time and shield the user from the maintenance nightmare.

Figure 9. Over time, the basic structure of data changes

By evolving the data warehouse infrastructure to the DW 2.0 architecture, organizations become better able to balance data to meet access and system performance requirements. In doing so, the cost of the data warehouse is mitigated, enabling the data warehouse to more efficiently accommodate huge amounts of data. In addition, the data warehouse can store and manage data over a wide range of time. DW 2.0 manages data that is two seconds old and data that is 20 years old.

Figure 10. There is a natural evolution of data warehouses from the classic first generation to DW 2.0
Informatica Data Archive™: The Complete Data Warehouse Archiving Solution

Informatica Data Archive helps your IT organization to cost-effectively manage the explosion of data volumes in data warehouses. It allows you to easily and safely archive inactive data, and then readily access it when needed. Informatica Data Archive delivers the full range of capabilities that your IT organization needs to effectively manage data growth in data warehouses, including:

- Robust archiving techniques that ensure data integrity after archiving and supporting multiple archive formats to enable optimal storage tiers
- Multiple, easy access methods to archived data
- Automatic indexing of archived data
- Automatic management of changing data structures
- Universal connectivity
- Integration with other archiving platforms, ECM, and storage solutions, such as Symantec, Commvault, and EMC

By leveraging the power of the Informatica Platform, the industry's leading data integration platform, Informatica Data Archive enables organizations to handle the huge data volumes typical of very large global enterprises. The software provides superior scalability and performance, delivering data to the most cost-effective storage option based on their value. It also offers unparalleled interoperability. The software is based on an open, easily extensible architecture, enabling simple integration with third-party solutions.

Robust Archiving Techniques Enable Optimal Storage Tiers

With Informatica Data Archive, you can archive to another data warehouse instance or to a highly compressed file format that can result in dramatic storage capacity saving. The compression ratio that can be achieved is based on the size of the data—the larger the data size, and the more redundancy in data values, you may be able to achieve a compression ratio of 20:1 to 60:1 compared to the original data size. The choice of archiving to another data warehouse or a compressed file archive should be based on the age of the data and response time as well as frequency of access. If you still need to access the data with relatively high frequency and with high performance, then archiving to another data warehouse instance is more appropriate. However, if data will be rarely accessed, for infrequent reporting or audit requirements, then archiving to a highly compressed file is the more optimal solution. Archived data can be stored on a file system located in lower cost storage or even storage in the cloud, for economies of scale. As data ages and access requirements change over time, Informatica Data Archive automatically converts and relocates the data from one archiving format and location to another, enabling multiple cost-effective storage tiers.

Informatica Data Archive enables you to archive transactional and detailed data only, which are the fastest growing. This is done while maintaining data integrity and links to dimensional and aggregate tables that may still be stored in the production system. Eventually, some older dimension records may also be archived as well. Informatica Data Archive has deep knowledge about what types of tables should be archived to support an optimal archiving strategy. Informatica Data Archive can also handle partitions that were created in the production data.
warehouse and maintain those data partitions in the data archive, to maintain scalability and performance. Figure 11 illustrates a data warehouse archiving strategy where detailed data are slowly relocated to another database and subsequently to a more optimal compressed file format, which results in extreme reduction in storage capacity.

Informatica Data Archive provides an easy to use graphical interface to define archiving jobs easily without extensive configuration, scripting, or programming. Figure 11 shows the Informatica Data Archive wizard-based interface to allow users to easily define and monitor archiving jobs.

A data warehouse archiving solution that offers multiple archiving formats and accessibility options allows IT organizations to determine the appropriate tradeoffs among archive size, performance, application accessibility, and cost.

Your IT organization must also be able to restore archive data to its original location. Otherwise, there is no way to correct mistakes during archiving or to accommodate changes to access requirements. If archived data later needs to become active again and for some reason modified and annotated, then it also needs to be restored. For example, a customer order that is closed and reopened may need to be restored because it has become active again. Informatica Data Archive allows you to restore archived data at different levels of granularity, such as selected detail records, business entities, or an entire archive.
Multiple, Easy Access Methods to Archived Data

Regardless of the archive format, archived data needs to be easily accessible either from the original application interface or through standard interfaces for reporting or compliance audits. Informatica Data Archive supports standard SQL/ODBC/JDBC interfaces for reporting using any reporting or business intelligence tool. The solution also offers the option to access the data from an application-aware data discovery portal to easily search, browse, and view archived or retired data based on business entities and with a similar look-and-feel as the original application interface.

Automatic Indexing of Archival Data

When archiving data to another data warehouse instance Informatica Data Archive automatically builds and maintains indexes that exist in the production data warehouse instance. When archiving to a highly compressed file archive, data is automatically indexed and stored in an optimal format to facilitate efficient storage and scalable retrieval. No performance tuning and maintenance is required on the archival data, reducing IT staff time.

Automatic Management of Changing Data Structures

As the production data warehouse structure continues to evolve, Informatica Data Archive automatically updates the metadata and structure of the archival data warehouse. When archiving to a highly compressed file format, Informatica Data Archive maintains multiple versions of the metadata, corresponding to periodic snapshots of the production data warehouse structure. This enables point-in-time querying of the archival data based on the structure of the data warehouse at that point in time. By automatically managing the metadata and structure of archive data, based on the changing structure of the production data warehouse, Informatica Data Archive reduces the maintenance effort required on the archival infrastructure.

Universal Connectivity

If your organization is like many other enterprises, you have data warehouses and applications on multiple database systems on varying operating systems. To support your enterprise needs, Informatica Data Archive enables you to manage archive processes across data warehouses and applications on diverse databases, including relational (e.g. Oracle, DB2, Sybase, SQL Server, Teradata, Informix), mainframe (e.g. IDMS, VSAM, IMS), files, and packaged CRM and ERP applications on open systems (e.g. Windows, Linux, UNIX) or mainframes (e.g. z/OS, AS/400).

Integration with Other Archiving Platform, ECM, and Storage Solutions

Your company may already have an archiving solution for e-mails and files. Your IT organization may also have standardized on an Enterprise Content Management (ECM) solution to manage your unstructured data. To support compliance to regulatory requirements and ensure immutability and single instance storage of retained data, you may be using archiving platforms, such as Content Addressable Storage (CAS), which requires proprietary connectivity.

To enable your organization to respond quickly and accurately to audit requests as well as to cost-effectively retain data for longer periods, Informatica Data Archive allows you to manage and discover archived data of all types, both structured and unstructured, centrally. This is achieved through integration with existing archiving, content management, and storage solutions, including EMC Documentum, Symantec Enterprise-Vault and Discovery Accelerator, and CommVault Simpana and eDiscovery, to facilitate centralized management and e-Discovery of all types of archived data.
Conclusion

Based on an explosion of data in corporate environments, the data warehouse has evolved from a simple platform for reliable, consolidated and integrated data reporting to a sophisticated data infrastructure that recognizes a complex data lifecycle. By understanding the dynamic forces at work in data growth, storage and accessibility, and the technologies available today to effectively manage the operation and cost of the data warehouse, IT organizations should now be better positioned to implement solutions that make their data warehouse environments more manageable, productive and efficient.

DW 2.0, the second generation data warehouse, recognizes that as data ages, its characteristics and access requirements change. By dividing data into different storage tiers based on the age and frequency of access, from high performance storage for interactive data to lower cost, lower performance storage for low access or inactive data access, DW 2.0 provides a platform for managing warehouse data more effectively.

The key to capping your IT organization’s data warehouse management costs and risks is to relocate dormant data to a lower-cost infrastructure that the storage tiering in DW 2.0 architecture makes available. This is what data warehouse archiving solutions can do for you—archive data based on its point in the lifecycle, while maintaining data integrity and easy access to the data.

Informatica Data Archive enables organizations to handle the huge data volumes typical of very large global enterprises. By providing comprehensive and robust techniques to easily and safely archive inactive data, and then readily access it when needed, Informatica Data Archive delivers the complete archiving solution necessary to provide an optimal and cost-effective data warehouse infrastructure.

When your IT organization implements a complete, scalable, and flexible archiving solution, you’ll reduce the total cost of ownership of your data warehouses and other applications by:

• Reducing storage, server, software, and maintenance costs
• Improving data warehouse performance
• Increasing data warehouse availability
• Supporting compliance with internal, industry, and governmental mandates and regulations

Together, Informatica and your IT organization can align the business value of data with the most appropriate and cost-effective IT infrastructure to manage it.
Learn More

Learn more about the Informatica Platform. Visit us at www.informatica.com or call +1 650-385-5000 (1-800-653-3871 in the U.S.).

About Informatica

Informatica Corporation (NASDAQ: INFA) is the world’s number one independent provider of data integration software. Organizations around the world gain a competitive advantage in today’s global information economy with timely, relevant and trustworthy data for their top business imperatives. More than 3,900 enterprises worldwide rely on Informatica to access, integrate and trust their information assets held in the traditional enterprise, off premise and in the Cloud.

About Bill Inmon

Bill Inmon, the father of data warehousing, has written 52 books translated into 9 languages. Bill founded and took public the world’s first ETL software company. Bill has written over 1000 articles and published in most major trade journals. Bill has conducted seminars on every continent except Antarctica.

References

W H Inmon, DW 2.0 – ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING, 2008, Morgan Kaufman, Boston Mass

Inmoncif.com – a website with many white papers and other information about data warehouses and DW 2.0.