Data Archiving

A Modern Day Necessity

Tribhuvannath Mishra, PMP
tmishra@outlook.com

Abstract

As businesses start growing so does the data associated with them. To manage uncontrollable data growth on your primary storage tier a strong Data Archiving strategy is required. As data ages, its value to the organization, access frequency, granularity of data needed, main consumers of data also changes over time. A good Archiving Strategy not only provides more cost effectively solution but also provide greater control on data to meet Government and regulatory agency requirements.

I. Introduction

Even though Archiving Solution and Data backup may sound similar but they are very different. Backup is needed for operational support such as disaster recovery. Archiving of data is required for technical and legal reasons. Archiving solution help in resolving memory and space issue by offloading sparsely used data to different storage while still making it available to user seamlessly. It helps to improve performance of live database while meeting all the complaint by ensuring statutory data retention rules.

II. DATA ArchivE Process:

Archival process start with defining the needs for organization. Business nneeds may vary based on organizational structure but most of archival strategy should satisfy conditions listed below:

- Low Maintenance: User shouldn’t be dependent on support staff to pull the data from Archive storage. It should be easily accessible and available On-demand for the users. Archival process should be automated in a way which requires minimal human intervention to maintain whole data cycle.
- Compatibility with Existing Applications: Proposed Archival solution should be compatible with existing Commercial Off-The-Shelf (COTS) Products and other in-house applications. It lowers the maintenance cost and also increase greater acceptability.
- Data Status Cycle: There should be a clearly defined data status cycle such as Active, Archive or Purge. Well defined data status cycle (Table 1) helps all the users and stakeholders to have clear understanding of location, age and amount of data in active database v/s archived data storage.

Table 1. Data Change Cycle.

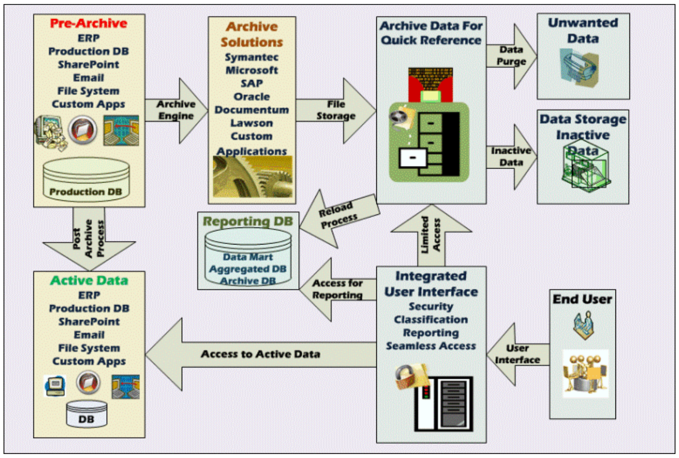
Term	***Definition***
Active	This DATA STATUS refers to current production data and reports.
Archive	This refers to the PROCESS of moving data from the Production Database to the Archive staging files
Load	This DATA STATUS refers to Loading the staged data to the Archive Database (
Retention	This refers to the period of time, which data must be accessible, either in the Production DB or the Archive DB. i.e.
Data Purged	This refers to process of removing unwanted data from both Production and Archive
Production Database	This refers to the Production Database where Active data resides.
Inactive Data Storage	This refers to storing the data offline. It can be accessed by authorized people only for emergencies and can’t be access by normal user.

- Scalability: Archive Solution should be designed to support future growth. Past data growth can be used to benchmark the growth pattern. Short term and long term strategy should be designed separately to meet internal and external growth demands.
- Legal Requirements: Archival process should be in compliance with data retention policy of the company and legal requirements set by the government. Some of the criteria are listed below:
  a. Data status change cycles.
  b. Archive Database Retention Period.
  c. Archive Data Reporting requirements.
  d. Archive Data Security requirements.
  e. Granularity of Archive Data

III. Major Components of data archive solution

Data from various existing system is archived using archive solutions. Data is moved to interactive reporting database, data vault or data shredder based on type of data. User can seamlessly access Active Data, Reporting Database or Archived Raw data based on access level and needs. Fig 1 below shows major components of a typical Data Archive Solutions.

Fig. 1. Componenets of Data Archive Solutions



IV. Activities for Planning Data Archive solution

- Identify the data not needed frequently but need to be accessed by user due to legal or business requirement
- Identify the data which is growing exponentially. The data which is growing very fast but it is not needed frequently will be prime candidate for archiving.
- Analyze impact on other integrated systems. Data in Big organizations are feed from various systems and data is also access from various internal and external systems. A holistic analysis of all these system is needed before implementing any solutions.
- Check Legal and Company rule for data retentions. Data must be retained in active or archived state based on these rules. Table2 shows some of the regulations and compliance for data retentions.
- Select between in-house custom application development or COTS product. Budget, available in-house expertise, and business impact may decide make or buy decisions.
- Design Post archive optimization process carefully so that not only archiving process run smoothly but also post archiving database optimization, indexing and recapturing of free space done properly to make active data operations optimized.
- Select Primary data storage, data vault storage and data purging software and hardware based on various factors such as volume of data, cost, sensitivity of data.
- Design archive reporting database (if needed). Some of the organization preferred to have separate reporting for archived data so that user can seamlessly get required data without knowing complexity behind it.
- Provide secured and Integrated User Interface with active and archived data. Archiving solution cannot be implemented successfully unless data is safe and made available to only those needed in timely manner.

Table 2 . Data Retention Regulations and Compliances

Regulation	**Data Retention Compliance**		
	Agency	***Impacted***	***Requirements***
PCI	PCI Security Standards Council (PCI SSC)	Financial Institutions Merchants Hardware / Software Services	PCI requires card issuers and acquirers to retain an audit trail history for a period that is consistent with its effective use, as well as legal regulations. An audit history usually covers a period of at least one year, with a minimum of three months available on-line.
U.S. Federal Rules of Civil Procedure (FRCP)	U.S. Federal Courts	All industries	Legal hold to preserve data unaltered. Rapid production of unaltered data and records
Sarbanes-Oxley	SEC	All industries	Section 103 requires firms to prepare and maintain, for a period of not less than seven years, audit work papers and other information related to any audit report, in sufficient detail to support the conclusions reached and reported to external regulators.
Dodd-Frank Act	United States Government various agencies.	Financial Institutions	Under Dodd-Frank, firms are required to maintain records for no less than five years.
21CFR (Code of Federal Regulations) Part 11	Food and Drug Administration	Pharmaceuticals and medical device manufacturers	Data security, integrity, auditability

V. Design Criteria for Archive Solution:

In addition to changing business needs, a lot of external and internal factors are to be considered while designing a cost effective archive solution. Some of the factors are listed below:

- Applications - Different types of applications need different archival solutions. Applications such as Emails, File system, SharePoint, Weblogs, Database ERP systems and Custom applications may need archival solution. For example simple 10 GB shared drive archival of individual email accounts for 30 days may be sufficient to meet the archival needs whereas hundreds of GB storage might be needed for a complex Financial ERP sales transaction which is to be retained for several years.
- Industry Type: Depending on the type of industry, sensitivity and retention period may vary. Data from Healthcare and pharmaceuticals , Finance, Telecom, Government and Judicial sector may need special consideration in terms of retention period and

data security. Some retail and advertising industry may not have to retain their data for long period of time.

- Hardware: There are two classes of network storage: storage area networks (SANs) and network-attached storage (NAS) each one of these need to be evaluated before selections. Cloud based storage is also gaining popularity. SANs and NAS storage are similar except wire and protocols used by them. NAS uses Transmission Control Protocol/Internet Protocol (TCP/IP) Networks wires SAN uses Fibre Channel. Both NAS and SAN can be accessed through a virtual private network (VPN) for security. NAS uses TCP/IP and Network File System (NFS) / Common Internet File System (CIFS) / Hypertext Transfer Protocol (HTTP) protocols while SAN uses Encapsulated Small Computer System Interface (SCSI) . Almost any machine that can connect to the local area network (LAN) or is interconnected to the LAN through a wide area network (WAN)can use NFS, CIFS or HTTP protocol to connect to a NAS and share files. Only server class devices with SCSI Fibre Channel can connect to the SAN. The Fibre Channel of the SAN has a limit of around 10km at best.

- Existing COTS products: Investments done already on existing COTS products such as SAP, Oracle, Microsoft or In-house development may be a major deciding factor in choosing archive solution. Architecture for a unified and comprehensive archival process should provide data security features. Necessary authorization and authentication controls should be in place to govern user access and permissions to data to ensure strict data security.

- Quick data retrieval: A clear understanding of enterprise data profile helps to decide effective data indexing and classification strategy for quick retrieval of archive data when required.

- Data integrity: Data needs to be identical and reliable each data change cycle to meet most of government compliance and regulations. Saving data in its native format not only help in maintaining data integrity but also dramatically improve ability to retrieve data quickly when needed.

- De-duplication: Removing unwanted duplicate copies of data not only helps to reduce the amount of data to be stored but also result in reducing the overall cost of managing the data.

- Ease of Access: Users needs the ability to quickly and easily locate and access information seamlessly from active and archived data.

- Seamless Data Transfer: The purpose of archive and purge programs are to move data from active environment to archive storage to free up space on active application. Seamless transfer of the data from active to archive storage without affecting system performance increases acceptability of archive solution by users.

- Audit Trail: With strict government regulations, enterprises are required to maintain audit trail which should be able to provide a complete record of activity, changes and breach of any security policies.

- Agility: Archive solution must easily scale for volume and performance. Agility is critical for meeting business demands of new retention policies, change in data content, change in growth pattern and regulatory changes.

- Commercial Archiving COTS Products: Many industries face exponential data growth in very short term due to rapid growth of e-commerce transactions. To support this rapid data growth, most of the major hardware and software companies have already developed COTS product for Archiving solution. These COTS products are designed to support different business sectors and size of the data. Customized archiving solutions are preferred by small industries due to cost involved with most of COTS products. COTS products offer better reliability and integrations but it may cost more to customize and licensing. Some of the providers of archiving COTS products are Oracle, SAP, EMC, Symantec, Informatica, Infor (Lawson), HP. Some of the COTS products may specialized in few aspect of archiving like Microsoft for emails, Oracle for RDBMS, SAP for ERP, Symantec for sensitive data and many more. While selecting the COTS product apart from cost, a careful evaluation of other software, hardware and in house expertise is needed. Most hardware vendors work closely with archive COTS software vendors to provide specialize storage hardware for the archiving solution.

VI. BENEFIT OF EFFECTIVE ARCHIVAL SOLUTION:

- Improve application performance and lower IT costs by archiving inactive data and legacy applications.

- A smart implementation of archiving solution provide easy and appropriate levels of access to archive data for end users.

- With archive solution in place, the size of active data can be effectively controlled and in turn would aid in streamlining maintenance/upgrade processes and results in IT staff efficiency.

- A proper archival strategy ensures effective maintenance of data status change cycle. It also helps in maintaining government regulatory compliance by proper data retention and retirement procedures.

VII. KEY FOR SUCCESS:

- Staff/Resource : Analytical and dedicated team always help in successfully implementing projects. IT projects are heavily centric towards teamwork and coordination amongst various IT resources is essential for its success.

- Process Driven: The Archival solution should be process driven and steps should be taken to clearly outline the process flow.

- Smart Automation: Automation is a necessity in projects which involve frequent repetitive steps. Archival project has many repetitive steps. Post implementation there would be subsequent cycles of data migration from active to archive. All these steps have to be automated to avoid human error and ease of use as well.

- Provision for future model changes: The Archival framework should be code compatible with future releases.

- Budget commitments: Budgeting requirements are most important thing to keep a watch on during the project. Usually, it is a good practice to have some reserve for rainy day. Frequent analysis and revisiting the fiscal needs is always helpful to keep project on track with respect to the budget.

- Clearly defined Data Status Change Cycle: A clear cut definition of when, how and what data get archived will help every stakeholder understand data status clearly at given point.

- Scalability : The Archive solution should be built by keeping the future data/model changes in mind. Many a times data can grow exponentially. The Archive modeling should be implemented in a way to avoid latency issues associated with changes on the Legacy system.

- Documentation: Everything implemented should be documented properly and document should be source controlled so that latest document is always available whenever needed. Over documentation also kills the purpose of documentation as its get harder to search for useful information from pile of junk and redundant information. One of archive project we found four different documents but couldn't figure out even after reading each of them which one is relevant to current state of archiving solution. We created another document from scratch. Documenting customizations and some of the key aspects of archive process can help not only to troubleshoot problems but also knowledge transfer to other staff.

VIII. Reasons for Failure:

Storage Strategy: Many times the archive solution is implemented on the same server as the live system. This approach leads to contention of resources and introduces latency issues.

- Volume of Data: If we ignore the future growth prospect of the data getting archived from the Legacy System then we are bound to run into performance bottleneck. For example, a company dealing with student test data didn't predict exponential data growth which in turn resulted in latency issues with the application.

- Data Mix (Legacy Data vs. New Data): If the archive solution was previously implemented and it has an existing data set then introducing a new revised solution might introduce data redundancy and data type issues when we try to merge the two data sets. For example, a finance company designed the archive table structures specific to a particular version of software without providing provision for future updates resulting in duplication of process even when there was simple changes.

- Frequent Model Changes: Due to various updates and patches pushed by application components underline data model may also be changing. A close monitoring of cause and effect of these changes may avoid too many cycles for the Archival Solution and make it less prone to failure due to architecture changes. For instance, an ERP installation failed to upgrade their product due to heavy customization on the code.

- Vendor License Restrictions: Many a times ERP vendors have restrictions on data changes and migration activities. If Vendor doesn't allow certain aspect of customization, user should avoid including this change as it may violate licensing agreement for support.

- Frequent Manual Interventions: Like any other software product, if the archival process involves frequent manual steps then it will create the following issues - missed steps, incorrect sequencing, introduce human errors, unnecessary delay due to manual steps.

- Project Management: Project manager for archival process need to have knowledge about the business domain, key elements of archival process and good sense of available staff resources along with project management skills. Failing to understand these dynamics during planning stage can derail the project at later stage.

- Staff/Resource: Archival Solution implementation team has to have clear picture of the steps involved and should have the technical expertise to meet the implementation requirements.

IX. Conclusions

In today's rapidly changing information age, exponential data growth has made data archival a critical necessity. Government's stricter data retention laws and regulations make it even more challenging to achieve an archive solution. Data should be stored in a way that it could be retrieved not only when needed but also in secure and reliable fashion. The ideal archival solution should ensure the availability, retention, compliance, and the secure access to business information needs in cost effective way. It should include a comprehensive and unified framework which creates a long lasting enterprise architecture protecting the enterprise investment in software, hardware, process and organizational support and meet future scalability challenges.

X. References

[1] Archiving and Vaulting Solutions -http://www8.hp.com/

[2] Storage Concepts: Storing And Managing Digital Data (Volume 1) by Hitachi Data Systems Academy (Jul 18, 2012)

[3] Implementing an InfoSphere Optim Data Growth Solution - An IBM Redbooks publication

[4] The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling by Ralph Kimball and Margy Ross (Jul 1, 2013)

[5] [4] Data Modeling and Database Design by Richard W. Scamell and Narayan S. Umanath (Jan 17, 2007)

[6] Data Modeling Made Simple: A Practical Guide for Business and IT Professionals, 2nd Edition by Steve Hoberman, Michael Blaha, Bill Inmon and Graeme Simsion (Sep 15, 2009)

[7] Archiving Your SAP Data (2nd Edition)

[8] CNET - Future-proof your data archive

[9] informationweek.com - Plot An Effective Data Archive Strategy

[10] Enterprise Backup, Recovery, and Archive Products and Solutions - www.emc.com/

[11] Implement an Archiving Solution That's Right for You - http://www.symantec.com/

[12] Why enterprise data archiving is critical in a changing landscape- http://www.informatica.com/

[13] Cloud-based data archiving service - http://www-935.ibm.com/

XI. About the Authors

Tribhuvannath Mishra, PMP, is a successful project management professional with Engineering degree. He has decades of experience in evaluating, designing, development, and deployment customize Data warehousing, Reporting, Archiving and ETL solutions for Government and Corporate clients in Retails, Manufacturing, Telecommunication, Healthcare, Insurance, Finance, and Education sector based on their specific needs. When he finds spare time, he loves to writes poetry and book.