

Realtime
publishers

White Paper

Best Practices for Data Archiving

sponsored by



Best Practices for Data Archiving.....	1
Prudent Policies	2
Data Vaulting Policies	2
Defining the Data Life Cycle	3
Keeping Data Current	3
Maintaining SLAs	4
Effective Procedures	4
Simple, Reliable Operations	4
Keeping the Data Archive Procedures Updated	5
Secure Storage of Archive Data.....	5
Monitored Operations	6
Empowered Personnel	6
Balancing the Archive System.....	6
Building Operator-Friendly Data Archive Systems.....	7
Empowered to Keep Operations on Track.....	7
Equipped to Optimize the System.....	8
Selecting the Right Products.....	8
Data Archive Architecture	8
Full Range Agent Support	8
Rich Reporting Services	9
Adaptable to Changing Environment.....	9
Summary	9

Best Practices for Data Archiving

Corporations depend on the flow of information to conduct their business. From accounting to production to marketing to personnel, the information procured, generated, organized, and stored by an organization represents a significant portion of the output of its workforce. Protecting that information from unexpected disasters is not only prudent, in many cases, it's mandatory for legal and regulatory reasons.

To effectively protect its information, the organization must balance the cost of the protection against the value of the information. The right approach involves a synthesis of appropriate policies, procedures, personnel, and products. This paper explores the design and implementations of such systems. It will discuss many of the best practices that help such systems provide the best Return on Investment (ROI) for data archive solutions:

- **Policies**—Organizations need to plan in advance for disasters or fall prey to them. Its policies should define where archival data should be stored. Policies should provide a life cycle for the data so that it can be safely and securely released in a timely manner to make room for additional data. They need to clearly define service level agreements (SLAs) that state how quickly data must be restored when it is required.
- **Procedures**—The organization needs to implement procedures that are simple to administer and maintain. Such procedures should be easy to administer in a distributed environment, and flexible when meeting the changing landscape of the IT infrastructure. Organizations should prove capable by practiced restorations and be able to validate that the data is held securely at all times.
- **Personnel**—The staff should be empowered to monitor the system centrally with minimal effort and training. They should have the understanding and tools to optimize the system and ensure that data is archived as mandated and designed. They should be able to proactively tune the system and quickly troubleshoot errors so that data does not remain unprotected and at risk.
- **Products**— The product chosen to support the archival system should meet the organization's needs. The architecture should support the full range of operating systems (OSs) and applications used by the organization. It should support a distributed model for storing data and administering the system. It should protect data at all times through strong encryption and virus scans. The system should be easy to keep updated and keep pace with updates to servers and data sources it protects. It should provide clear alerting, reporting, and auditing of data archives. It should also provide the tools to optimize the process and help protect corporate data in an effective manner.

Prudent Policies

Policies help the business convey its level of expectation to the staff. Policies should be crafted to make clear what will be expected from the people and systems that protect the data. One of the first considerations is where the data should be archived.

Data Vaulting Policies

In order to adequately protect the data, it must be stored in a location remote from the primary data center. Disasters, both natural and man-made, have proven that keeping all of an organization's data in one physical location is to leave it at risk.

The choice of data vault facilities involves many considerations. First, the facility needs to have good connectivity to the primary data center. The bandwidth and stability of the network connection between the data vault and data center will determine the rate and reliability of the backup procedures. With organizations that span the globe, this consideration becomes more important.

This leads to another consideration. With international enterprises, it is often prudent to distribute data vaults in multiple locations in multiple countries and continents. This can help contain the cost for international Internet connectivity and reduce the latency in the data connections by reducing the physical distance.

Additionally, the data vault must be secure. Access to the data, whether electronically or physically, must be guarded and carefully audited. The facility must be able to control access to the backup media and provide the means for recording who accesses that data and when.

The backup software can help facilitate many of the aspects of this policy. First, the software can help manage data stored in multiple vaults in multiple locations. If the administration of this data can be managed from a centralized console, it remains easy to locate and track. The system should be distributable. This will allow people in diverse locations (potentially in different time zones) to maintain control of the system around the clock. Such systems remain easier to man and more responsive to the needs of the organization. The system should also centralize reporting of the data access records. This will help validate the security of the system and make it easier to prove compliance with corporate and regulatory standards.

Defining the Data Life Cycle

When archiving data, a balance needs to be struck. It is tempting to keep all the data that a company has ever used forever. One can revisit decisions and actions and determine what went right and wrong. However, some information is better soon forgotten. Much of the data that is stored is transitional—valuable while an order is in process, but transient and worthless once the order is completed. Some data must be stored for a defined length of time by regulation. Other data is just copies and summaries of data stored in other systems. Regardless of the data's source or value, it costs money to keep copies of it.

It is important to identify what data should be stored and for how long it must be kept. Although information life cycles can be difficult to manage, the IT staff needs to know how long data needs to be held in the system. For instance, suppose the document management system stores its data in a database that is saved every night by the archive system. How many copies of that database should be saved? Is the latest copy sufficient, or should the last five versions be kept? How often should old versions of the database be purged to make room for new versions? How much archive time and space should be allocated to dated copies of data?

The data life cycle policy will help determine the answer to these questions. It should provide IT with the guidance they need to determine the number of versions they should save and the amount of space they need to preserve them. They will be able to project the capacity of the data vault required and properly size the system. They will also be able to determine the bandwidth and computer resources required on a regular basis to perform backup operations.

The archive system can help in a number of ways. It can help track the data, in all of its versions, and make the age of that data clear. It can help report on the growth of the data archived and provide the basis for proactively projecting space requirements. It can help ease the process of securely removing aged data and freeing space.

Keeping Data Current

The amount of time that data is at risk is also a question of policy. Many business people are accustomed to the concept that data is archived every night, after they are finished for the day. But many transactional systems must be secured more often than that.

A database that handles orders, stock transactions, or other types of mission-critical tasks may not tolerate any loss of data. It might dump transaction logs on an hourly basis or even every several minutes. These logs must be stored quickly to keep them safe.

Policy should dictate how long a specific type of data can remain in the system before it is archived. This will vary from one type of data to another. People must then schedule the system to ensure that the data is copied as specified and provide evidence to make sure that the backups are occurring per policy.

The system should provide the means of efficiently scheduling and collecting this data on whatever schedule the policy mandates. It should provide reports that validate the operation of the schedule. It should provide timely alerts when the process fails, and the information and tools that will help the staff locate the source of the problem and correct it quickly.

Maintaining SLAs

Keeping the data safe is part of the equation. Restoring data on demand is another part. IT must understand how long data can remain inaccessible or unavailable before losses or damages. Some data can be offline indefinitely and no one will be materially affected. Such data is kept for rare occasions, such as regulatory audits. Other data, however, is required to keep the business running. The absence of such information can halt operations and cost thousands of dollars per hour.

Data that needs to be accessed more quickly needs to be kept in fast access, higher-cost media, such as disk arrays. Data that is archived and can be accessed more slowly can be stored in lower-cost media, such as tapes. The SLAs should define how quickly any given type of data needs to be restored. The policy should provide IT with the parameters they need to determine the type of media on which to store the archives, and when that media can be moved from more expensive storage to less expensive storage.

The archive system plays a key role in this decision. The system needs to be able to manage multiple forms of storage media. It should facilitate moving data stored online to offline. It should provide a record of the lineage of the data. Regardless of how and where the data is stored, it should simplify the process of retrieving and restoring the data.

Effective Procedures

While policy defines system requirements, procedures realize policy directives. The data and the systems that host it constantly change and juxtapose, so the procedures need to be easily altered to ensure the ultimate goal is always achieved.

Simple, Reliable Operations

Backup is seldom the primary concern of any member of the IT staff. Although very important, a well-run system requires little direct, day-to-day intervention. This allows the IT staff time to deal with other issues.

To facilitate this, the system should provide several features. It should be easy to schedule. Backup operations are often interrelated with other critical system operations—such as accounting system reconciliations, data warehouse processing, report generation, and other similar operations—so the system should make re-scheduling simple and reliable.

IT operations are often in flux. Servers are added, consolidated, and re-located. Procedures for deploying backup operations should be simple, and, preferably, administrated from a central console.

The archive system is central to keeping these procedures simple and reliable. The system should simplify the movement of jobs, whether moving from one server to another or following the server if it moves. It should make re-configuration of jobs a simple matter of using the central console to adjust parameters, such as the schedule. It should allow jobs to be deployed to new servers with little manual effort. It should self audit, providing reports that prove that the backups have occurred and offering timely alerts when jobs fail. The more responsibility the system lifts from the shoulders of the operations staff, the more cost effective it will be.

Keeping the Data Archive Procedures Updated

The servers and applications deployed throughout an organization change on a consistent basis. Existing OSs receive patches and upgrades. Applications also receive patches and upgrades. These changes can affect the way that data is collected or restored.

The servers and platforms on which data is deployed also changes. Mainframe applications may be moved to UNIX or other servers. One mail system may be replaced with another. Databases may be clustered or hosted in grids to improve availability. A document management system may be installed, changing the way documents are archived. Corporate acquisitions and mergers can add new applications and data sources that must be protected.

In order to keep pace with these changes, the procedures for preparing and deploying new jobs must be streamlined. Job changes and new applications should not be moved to production until the data can be adequately archived, so the time and effort to put archival procedures in place directly affects changes to the IT infrastructure.

The ability of the archive system to deploy on many platforms and support many applications is invaluable for keeping data archival processes updated. Most data archival processes will use agents to operate with specific OSs and applications. These agents need to keep pace with the changes to OSs and provide quick, simple deployment. The system should be designed to support all the various platforms supported by the enterprise, and provide options that allow IT to make strategic decision about what platforms are used to host critical data. The archive system should also track changes and provide evidence that all designated data is in fact being protected.

Secure Storage of Archive Data

The purpose of the data archival system is to keep data secure so that the corporation will not suffer from its loss. The system must also secure the data so that it cannot be stolen when in its archival form.

As a best practice, the data must be kept secure at every stage of its collection and storage. Data that is archived should be checked for viruses. It should be stored in an encrypted format. This will help protect it at every stage of its storage. The data should be transmitted securely within the network. The data vaults should have strict controls on who can access archival data and maintain complete logs of who accesses the archives and when. The protection should be extended to both physical and electronic access to the data sources.

The archive provides the technological support for keeping data safe. Systems that handle encrypting data and checking it automatically for malicious viruses will help preserve the integrity of the stored information. Systems that can manage control of the archive data will protect it from being appropriated in the wrong manner.

Monitored Operations

There can be a tendency to assume that once a system is designed, it will work in that way forever. Of course, people who monitor system operations day to day know that is seldom the case. A well-designed set of procedures will help operations people find problems when they occur and correct them quickly so that the data remains at risk for only short periods of time.

The alerting system should notify operators quickly when an operation fails. It should provide enough information to help pinpoint the source of the error. This will help operations personnel affect repairs.

The monitoring system should provide an ongoing record of successful operations. This will provide validation of the successful implementation of the policy. It should also provide performance information. This type of data can be analyzed by operations to determine how to adjust the system and tune it for optimal performance.

Monitoring operations that span time zones, platforms, applications, networks, and countries can be a challenge. An effective archival system should provide a common means of collecting this alerting and reporting data. It should store it in a consolidated data repository where flexible analyses can be performed. It should be able to distribute alerts to operations personnel anywhere in the world to get their attention and correct issues with the greatest possible dispatch.

Empowered Personnel

Data archiving is seldom the sole responsibility of a single person or group of people. It is typically part of the many tasks given to the IT staff members who keep the servers running day by day. This can make operating the archive systems cost effective. It also means that a system designed to be easy to implement and maintain will be the most effective in this role.

Balancing the Archive System

The data archive system should implement the policies established by the enterprise. The reality is that people must balance the requirements of the policy with the realities of the budget and available resources.

People must understand the risks to the enterprise data and the value of that data, then determine the ROI for protecting it. It is not a simple question of whether there is a copy of the latest version of any type of data. It extends into how many versions of the data are kept. It includes how long it takes to retrieve the data and restore it. It includes the storage capacity of the archive system.

People will interpret the policy and design and implement procedures that enact their interpretation. They need enough information to make the correct decisions. For instance, they might decide to keep transaction log backups on tape. If the database needs to be restored, restoration from tape may add a number of hours to the process. Although storing the logs on tape may save money until a restoration is required, it may cost hundreds of thousands of dollars when a restoration is required.

If people are equipped with the right information, they can designate some storage to online disk arrays and other data to store on offline media to save costs. If they can easily automate the transition of data from online to nearline to offline storage, or delete transitional data entirely once it is no longer needed, they can balance the costs of storage.

The archive system can help simplify these processes. It can help automate the management of archival storage, moving it from one source of media to another. It can store the lifetime of data and delete it once the mark is exceeded, thus freeing the media for other uses. All these aids will help people “right” size the archive vaults and help contain the costs of the data archive system.

Building Operator-Friendly Data Archive Systems

Data archiving is built on many platforms using diverse technologies. It would be overwhelming for everyone in operations to be able to work on every platform and with every application that the system needs to protect. Rather, systems that abstract the details of working with the underlying platforms and help people maintain operations from a single platform will help lower training costs and reduce labor costs for maintaining the solution.

There are a number of innovations that help accomplish this goal. Agents can be created that abstract the specific details of copying data into the archive system from the operator. For instance, an operator may instruct an agent to back up a file share. The OS might support shadow copy. The agent may then use shadow copy so that the files are copied as they existed at the point in time that the backup was initiated. The operator does not need to know how to use the shadow copy feature of the file system. The agent makes that portion of the backup transparent.

The system should help operators adjust the schedule in a uniform manner, regardless of the platform on which they are archiving. This relieves them of the need to know cron to schedule on a UNIX server and Windows Scheduler on a Windows server. The console should be intuitive and easy to understand so that operators do not need to spend extensive time learning it and constant use to keep their actions accurate.

The control system must also be easily distributed. An enterprise console that can be manned in Asia for part of the day and North America the rest of the day is critical for international enterprises that must manage operations around the clock. It can also allow the operators closest to problems to access the data that they need to remediate the issues.

Empowered to Keep Operations on Track

As the enterprise grows, so will the operation of the data archive system. Jobs are scheduled, but over time, the duration of those jobs will vary. Some will grow as they capture more and more data. Other will shrink, as systems are slowly obsolete and replaced with others. Some will use network links that start small and are increased through use. Others will have bandwidth reduced as the requirements are lessened.

To keep abreast of the shifting requirements, people need performance reports on the archive system. Reports that clearly demonstrate the shifting utilization of the system will help personnel balance the system, adjusting schedules and using the resources in their most effective manner.

They also need to be able to make simple adjustments to the schedule with little overhead. A simple, easily understood console that allows the schedule to be adjusted quickly will empower the staff to keep the system well tuned.

Equipped to Optimize the System

Beyond the confines of the schedule, the system may need to adjust the destination of the data and the manner by which it is moved. Data vaults themselves may be reconfigured, with new devices with new features added to store information. Changes in the complexion of the IT infrastructure may call for the consolidation or distribution of new data vault locations.

A system that makes it simple to route data archives and move data over selected network paths will help people expand and optimize the data archive infrastructure as required. Systems that support wide ranges of hardware for archiving will help the staff make the best choices for archive storage, balancing the cost of new equipment with the value of existing assets.

Selecting the Right Products

Throughout this paper, suggestions have been made for the features that the right data archiving solution should support. The system should provide features that will make it simple and cost effective to manage the data protection needs of the organization.

Data Archive Architecture

The system should support the varied and complex needs of the organization. It should allow for multiple data vault locations to be located in multiple sites. It should allow data to be stored securely, keeping it virus free and strongly encrypted at all times. The architecture should allow the entire system to be managed from a single, simple console. The console should be available to operators wherever they are located.

The system should simplify the tasks of designing, scheduling, deploying, executing, and monitoring backup tasks. It should make it simple for operators with myriad other tasks to monitor the backup system and correct problems quickly and reliably. It should allow simple re-configuration of the system to match changes in the infrastructure.

The archive architecture should help manage the storage of data, moving information from more costly online storage to less costly offline storage only when it is safe and prudent. It should remove data that has reached the limit of its usefulness and free storage for other uses.

Full Range Agent Support

Most enterprises are heterogeneous, so the data archive system needs a wide range of agents. These agents should support the full range of OSs and applications used by the enterprise. As that list is not static, the system should demonstrate that it covers most common platforms and applications, thus providing the depth that allows the enterprise to freely choose the best options for providing services without concern for protecting the data those services husband.

The agents should simplify and abstract the process of extracting and securing data. They should reduce the demand on the operations staff to master a wide variety of technologies. They should allow operators to design, configure, and deploy backup jobs from a single console, agnostic to the server on which they will operate.

Rich Reporting Services

The data archive system must supply rich alerting, reporting, and analytics for the operation of the system. The alerts should immediately show operators where troubles have occurred and provide them with pertinent information to allow them to quickly correct the fault and protect the data. The reporting should provide validation that data is secured. It should meet any requirements for auditing. It should also provide analytics to help IT manage the processes, project and control storage needs, and contain costs.

Adaptable to Changing Environment

The data and data sources protected by the archive system will always be changing. Effective systems will make changes simple and reliable. It will help to document the changes and keep IT abreast of the current deployment of the backup solution. It will facilitate the restoration of the proper version of the data fast and easy.

Summary

Data archiving is required to protect an organization's investment in mission-critical data. The process begins by defining policies that help value the data and determine how much time and effort should be expended to store it and how quickly it should be retrieved. Procedures should be devised to implement that policy and ensure that the data is kept secure. People must be empowered to design, deploy, operate, and maintain those procedures.

The choice of the appropriate product to support data archiving can help make the process both secure and cost effective. By understanding how the product supports the policies, procedures, and people in an organization, IT has the parameters they need to make the proper decision about the product they should choose to protect their organization.