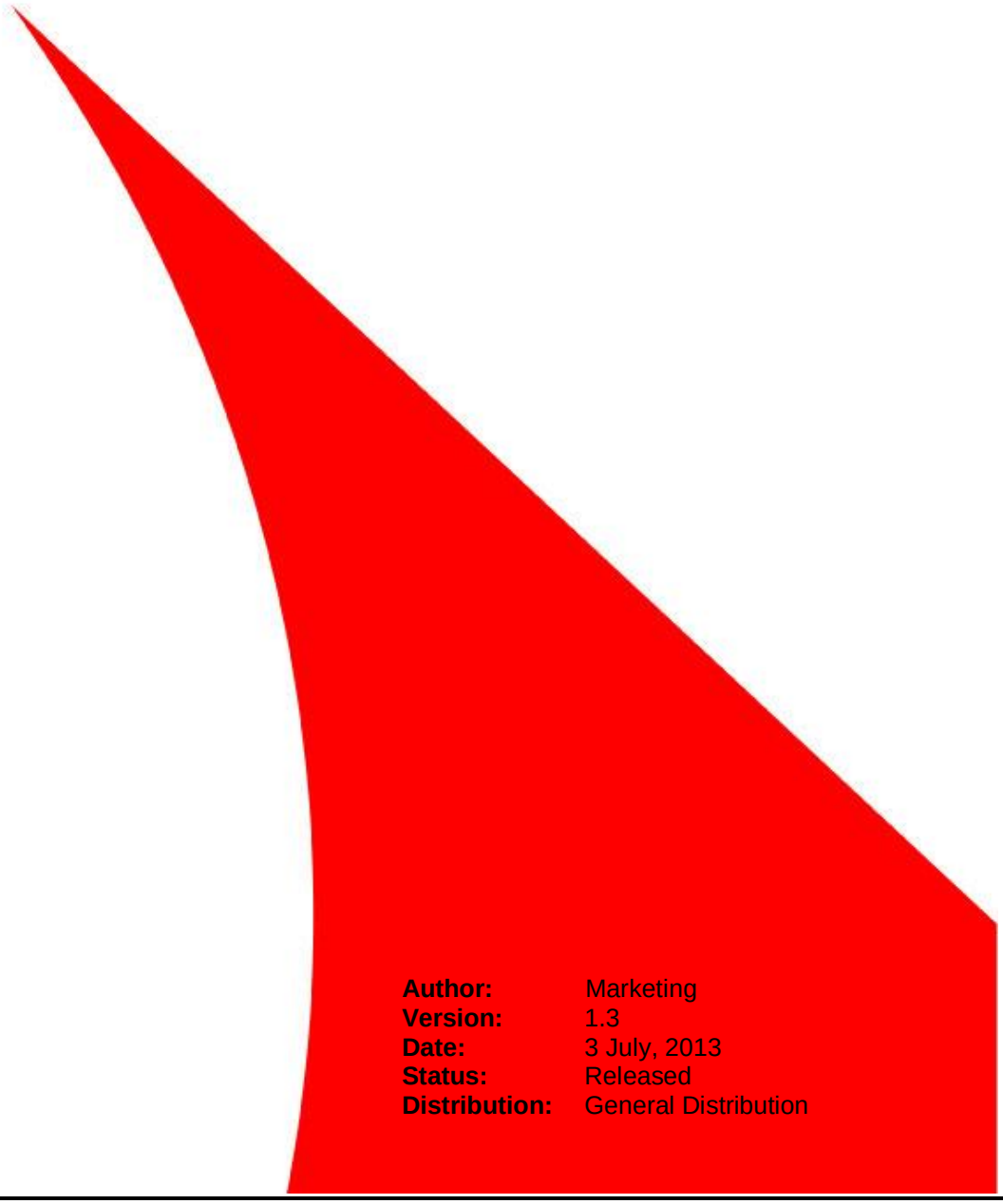
DESIGNING AN ACTIVE ARCHIVE

Guidelines and Considerations for Archive Success



Author: Marketing

Version: 1.3
Date: 3 July, 2013
Status: Released
Distribution: General Distribution

Table of Contents

Table of Figures

1 Active Archive Overview

Enterprise data continues to increase exponentially and has led to a rapid rise of online primary storage and its associated backup. Organizations are looking for an alternative to the standard IT response of adding more hard disk. One solution is to remove older data from primary storage and store it within an archive. However, traditionally archives have been seen as slow and only useful as a deep archive, where data is stored for long periods but seldom, if ever, accessed.

Active Archiving is the latest solution approach that can leverage the strong suits of any storage medium. Users now have the ability to extend a file system over a myriad of different storage structures to appear as a single, logical storage volume, allowing data to reside on the most appropriate storage level. Active Archive is a combined solution of open systems applications, disk, removable media and/ or cloud that allows users to access all of their data, and provides an effortless means to store and manage data. Small businesses, large corporations and government agencies all recognize the compelling need to ensure that valuable business assets are available through a cost effective and legally compliant strategy.

Unfortunately many organizations don't take into account the longer term perspective that is essential when designing an archive and can miss some of the most crucial considerations for a successful strategy. The purpose of this paper is to leverage the extensive experience of the authors in the design and deployment of professional archive solutions by offering proven guidelines and surfacing important considerations that should be factored into the design of any enterprise archive.

2 Archive vs. Backup

One issue that often causes confusion with even the most senior IT professionals is the distinction between backup and archive strategies. It is important to clarify this issue before proceeding with a more detailed discussion of archive architectures. Both backup and archive serve very important roles within the data center, but the fundamental nature and purpose of the applications are different and should be clearly separated.

Backup can be viewed as an insurance policy and should be used to protect and recover data that is actively being created or accessed. There is no point in repeatedly backing up unchanging archive data since this only leads to a bloated backup infrastructure with slower response times. Using a backup strategy as an archive also causes retention management and data authenticity conflicts which can violate industry regulations and increase corporate risk. It is far better to separate archive data from active data and store it in a independent environment where issues of compliance, retention and audit trail management can be effectively controlled. A well balanced data center will offload static data from the backup process to an Active Archive repository where it can be appropriately managed, reducing backup overhead and improving operational efficiency.

In short, backup is a tool for ***recovery*** in the event of system failure or user error and archiving is a tool for ***discovery*** enabling long term and compliant information access.

3 Archive Best Practice

Simplicity is often the key to creating a successful archive that is robust and cost effective. One proven approach is to adopt a framework that defines the overall structure of the archive, allowing individual organizations to adjust the parameters to fit specific needs. Endorsed by some of the storage industry's most influential analysts, Active Archiving provides this proven framework.

One methodology adopted by QStar Technologies is the use of 3-2-1 Archiving. Using a combination of software and hardware technologies, a 3-2-1 Archive advocates the retention of at least 3 copies of all critical business data, these copies should be kept on 2 different types of storage media, with at least 1 copy offsite.

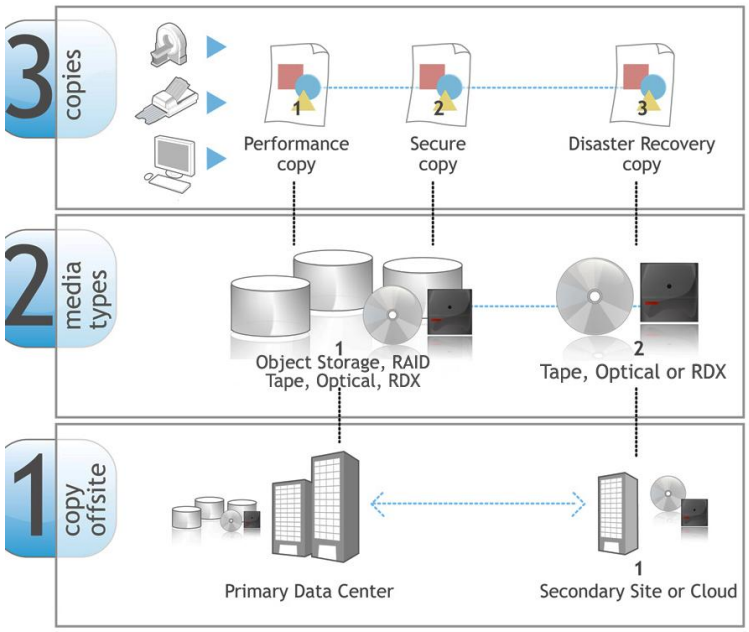


Figure 1 – 3-2-1 Archive

A 3-2-1 archive offers a number of compelling benefits. It helps to optimize the use of existing IT infrastructure, reduces the proliferation of duplicate data copies, avoids vendor lock-in through the intelligent use of RAID storage and removable media technologies, and offers unmatched flexibility to meet site specific requirements.

4 Design Considerations

This white paper uses Active Archiving alongside the 3-2-1 Archiving model as a framework to explore the unique requirements and considerations of an enterprise archive. The design considerations listed below represent many of the most common issues that should be addressed when planning the structure and configuration of an archive.

- Archive Capacity
- Archive Growth
- Data Profiles
- Performance
- Disaster Prevention
- Data and Device Refresh Cycles
- Archive Management
- Regulation and Policy Compliance
- Total Cost of Ownership

Since the priorities and budgets of each organization can vary greatly it is not practical to offer detailed advice, however the considerations highlighted in this document provide a high level overview of critical issues and relevant guidelines for further analysis.

5 Archive Capacity

Quantifying the volume of unstructured file data in a network is an important place to begin when designing an archive because overall archive capacity will have a significant influence on practical design decisions. Determining archive capacity may seem a simple task, but many system administrators do not have a detailed understanding of exactly how much data is being managed within their network. Capturing the total data volume is based on actual disk utilization minus duplication, orphaned data and system objects.

The next step is to identify what percentage of the total data volume is active changing data and what percentage is active fixed or static data. Active changing data is defined as data currently being created, modified or frequently accessed. Active fixed data can be defined as data that is no longer being modified and has not been accessed in weeks or months. Some active fixed data may have very little value to an organization, but other fixed data is business critical and must be retained for long-term availability. Very valuable fixed data should be classified as archive data and requires a well defined strategy if it is to be reliably accessed for long periods of time.

Archive data normally has very different storage and performance requirements than active data and by identifying and separating these two data categories, a strategy can be defined which creates a cost effective archive that ensures the long term access and authenticity of the data. This process of data separation is undertaken by Data Management Software (often also called Information Lifecycle Management, Hierarchical Storage Management, Storage Resource Management, Data Classification etc). This is a set of data identification tools which automate the archive process and migrate data to a more appropriate storage technology.

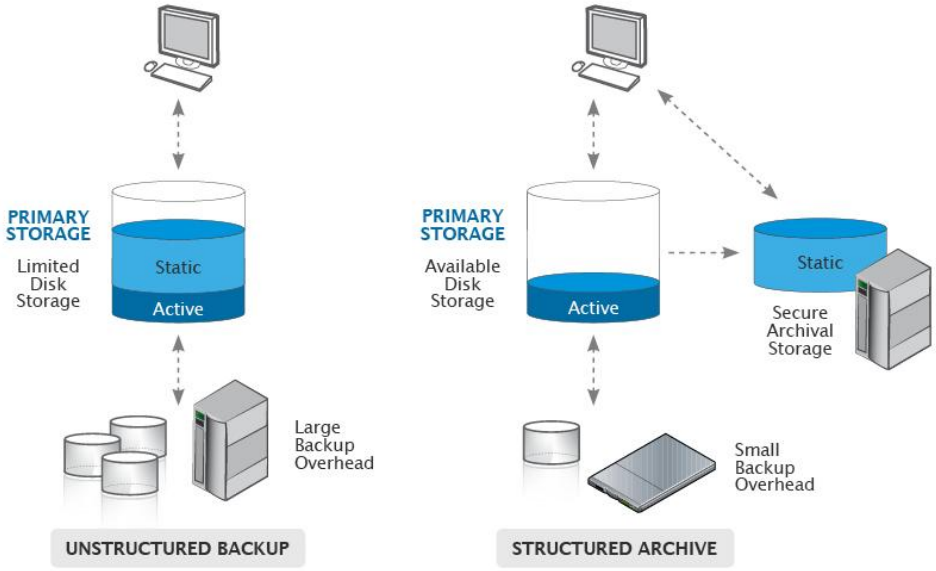


Figure 2 – Storage Optimization

Administrators are often surprised at how much of their online storage is consumed by unchanging data. It is not unusual to discover that as much as 80% of all files in a network can be classified as archive data. Separating archive data from active data can have major benefits in improved management, decreasing administration and reducing cost. For example, archive data that is no longer being modified need not be included in a normal backup cycle. Archive data can be protected outside the standard backup schedule so that backup operations can focus on the efficient recovery of active data that is changing on a frequent basis. This process alone can dramatically improve response time while reducing average backup windows, administrative overhead and the total cost of the backup process.

5.1 Deduplication and Archiving

Deduplication software and hardware appliances have become very popular in recent years. Object and block based duplication products can eliminate duplicate files or block patterns from a specific storage system or across an entire network. Controlling the number of duplicate file or blocks can significantly reduce storage capacity requirements, but it

must be emphasized that a deduplication strategy is not a substitute for the well structured archive and can create potentially complex legal conflicts. It remains vitally important for IT administrators to understand the relationships between different data types and business critical applications in order to define and manage intelligent storage strategies. For those sensitive business records that demand unquestioned record authenticity, the deduplication, compression and decompression of the physical files can place the authenticity of records in question and should be avoided in order to maintain the legal integrity of audit trails.

Deduplication can be a useful tool for unstructured file environments that contain active data, but it is generally not compelling for a structured archive. A well designed Active Archive, with a clear workflow, will not suffer from uncontrolled data duplication. In short, deduplication has its place for reducing unmanaged file proliferation, but it offers little benefit within a managed archive.

6 Archive Growth

The long term nature of an archive demands that administrators meet current requirements while also planning for future growth. Having said this, archive growth calculations can be difficult to measure accurately. It may not be adequate to base growth estimates on targeted data categories alone. Experience demonstrates that as an archive proves its effectiveness the net is cast ever wider to include new data groups not considered in initial calculations. For this reason, estimates should include the growth rate of currently identified archive data, as well as the potential addition of new records that may need to be managed in the future.

Constantly growing volumes of archive records demand solutions with significant initial capacity and the flexibility to scale over time. In order to achieve this degree of scalability the software and hardware environment must enable cost effective incremental growth. Archive solutions that require costly upgrades or major hardware replacement to achieve capacity expansion are contrary to the needs of a long term archive and should be specifically avoided.

The use of mixed media types promoted in an Active Archive also offers important advantages to a growing archive. Capacity targets can be achieved by balancing the distribution of content across the different media types including removable tape and optical library systems. Rather than purchasing new or expanding an existing RAID system, an Active Archive provides organizations with the option of taking older data sets offline to make room for more frequently accessed content. This flexibility is not possible with a monolithic RAID archive where the only option is the purchase of additional hardware and migrate data. Whether expanding an existing library or taking data offline, both techniques are non-disruptive and very cost effective for the incremental expansion of archive capacity without being forced to migrate data which can be costly, disruptive and places the integrity of data at risk.

7 Archive Data Profiles

Once the volume and scope of an archive has been determined, the next step is to sub-divide the content by establishing profiles for each primary data type. A profile is a set of attributes that characterize a document format, size, retention period, and access requirements.

Archives being used for a single task may only require two or three profiles since the document types may be limited. By contrast, archives being used by many applications will require additional profiles to reflect the wider range of document attributes. The objective of profiling is to add structure to previously unstructured data. Well-defined profiles create a direct connection between business priorities and the data center allowing administrators to intelligently manage both the physical and logical storage of the archive data in a way that is synchronized with business objectives.

For example, healthcare institutions handle both patient records, and high volume digital images generated by modern medical equipment such as x-rays, MR or CT scans. Most of this data must be retained for the lifetime of the patient. Fast data retrieval is a high priority in order to ensure that medical staff has ready access to patient information. Data integrity is also of foremost importance as it is an essential to maintaining a secure and unadulterated healthcare enterprise archive.

Engineering and manufacturing companies are also dealing with large and complex data, such as 3-D models, simulations, mock-ups and other technical documents, but the retention period and access requirements attributes will be different from those of medical records.

One advantage of data classification in an Active Archive is the ability to physically group similar profiles on the same piece of archive media. By doing this, older or less frequently accessed records could be taken offline making room for

newer data while still retaining the media for potential access. Using this technique, an engineering company could choose to archive all design, technical and administration records for one or more related projects on a single piece of media, allowing the project to be taken offline for future reference or distribution to another site.

8 Archive Performance

Data written to most archives is performed as part of a background or batch process. As files age or meet predefined event triggers, they are moved or migrated into the archive. This workflow means that the ingestion of records into an archive is seldom critical path when it comes to archive performance. By contrast, the timely retrieval of archived records is often critical to business operations. While it is true that most archive data does not require the same level of immediate access performance as active data, defining the level of performance that is needed to meet business objectives plays an important role in the storage hardware that is used and how the archive content is distributed across different tiers of physical storage.

It's easy to say that archive data should be instantly available for an indefinite period of time, but this is seldom necessary for older, infrequently accessed content. Demanding high performance has major implications for workflow management and cost since higher performance storage is always more expensive to purchase and maintain. More typically, response times of a few seconds or even a few minutes can often be acceptable to meet business operation and legal discovery obligations. A tiered storage structure that includes content distributed across RAID, and archive storage offers exceptional performance flexibility.

An Active Archive allows archive data to exist on multiple tiers of storage at the same time. This can be as simple as using a small amount of RAID cache and adding more cache capacity at a later date, should increased performance be required, without having any impact on the backend tape or optical archive. If a larger percentage of the archive requires higher performance levels then policies can be created to retain data on both disk and tape (a performance copy and a secure copy). This can be achieved using replication techniques, where a change is reflected after a period of time has elapsed, or mirrored, where instantly two copies of the same data are created. The structure enables non-disruptive, on-demand system tuning with the added benefit of physical archive resilience.

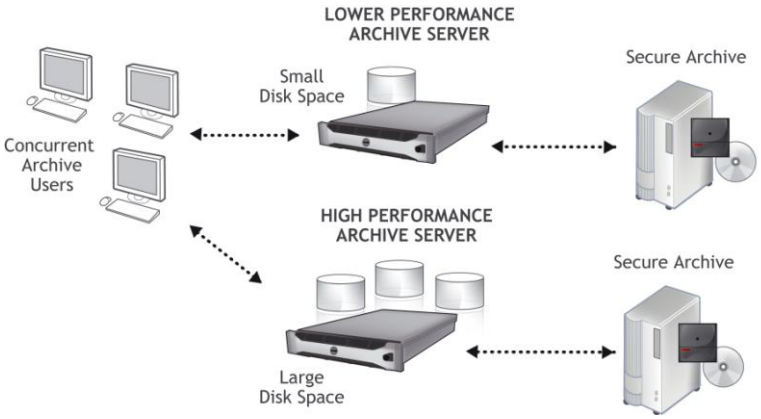


Figure 3 – Archive Performance Configurations

8.1 Concurrent User Access

One factor that should not be neglected when defining the performance service level of an archive is the number of potential concurrent users. As stated earlier, ingestion of content in an archive is typically not a performance bottleneck, but access can be if a large number of users are requesting content concurrently. This is particularly true when tape or optical libraries as the number of concurrent requests is further constrained by the number of tape or optical drives in the library. However, this can be offset by deploying a RAID system that retains all or part of the archive content as a storage tier in front of the library system. Typically this tier uses lower-cost SATA drive technology.

The ability to predict access patterns and user levels is important to properly configure the archive hardware. Using a mix of storage technology can address even the most demanding performance requirements while providing secure long term, removable storage on tape or optical media. An archive with insufficient "horse power" may cause request queuing, reducing performance and negatively impacting business operations. Understanding the access demands on

the archive allows the architecture to be balanced in a way that meets specific business requirements without expensive hardware overkill.

9 Disaster Prevention Strategy

Archive data worth retaining has intrinsic value to a company. Lost data can cost an organization a crippling amount of time and money to recreate and some records, such as legal or historic documents cannot be replaced which could result in significant fines. Even the most secure storage media will not protect data from fire, flood or sabotage. The only reliable method of protecting valuable archive data from any number of disaster scenarios is to maintain multiple copies of content at multiple locations. Every archive should implement a disaster prevention (DP) strategy that provides a second copy of the archive at a different geographic location. A DP strategy does add expense to the overall solution, but the investment is out weighted by the cost of losing business critical information and the potential legal, financial, and political ramifications.

The 3-2-1 Archive framework calls for one copy of the archive data to be retained offsite specifically for DP purposes. As with archive performance, it is easy to over-spec a DP strategy. While there are cases that may require a high availability DP structure, most archive environments have lower access service level requirements and this should be reflected in the DP facilities. Using removable tape or optical media offers a very cost effective and flexible approach when compared to deploying a fully mirrored RAID system.

There are three primary options that can be deployed in a tiered storage archive using 3-2-1 model. The first uses an automated copy regime, perhaps initiated during off-peak times. Data is copied from its primary media to a secondary media and once full removed from the library. The primary copy remains in the library and the secondary is stored offsite or in a fire safe. This is the most common form of data protection used, is the cheapest option, but will mean significantly longer amounts of down-time should the primary location be hit by a disaster.

The second method is to replicate and mirror all data archived to a second site or perhaps to a public Cloud storage environment. In this scenario data is always in two places and so a primary site disaster will not prevent users accessing their archived data. This method is the most expensive DP approach, but provides the highest level of data availability.

The third method provides a balance of the previous two. The second site is equipped with a smaller online archive capacity, perhaps using a smaller library than the primary site. Archived data is still replicated or mirrored to the second site but only the most recently created data is stored online in the library. Older datasets on older media are removed and stored in a fire safe. Some data will be available directly from the secondary library but older files will require a manual intervention to retrieve data.

A DP strategy is essential to all archive environments, but the specific configuration for a given organization will depend on the required availability of the archive content, administrative resource and budget.

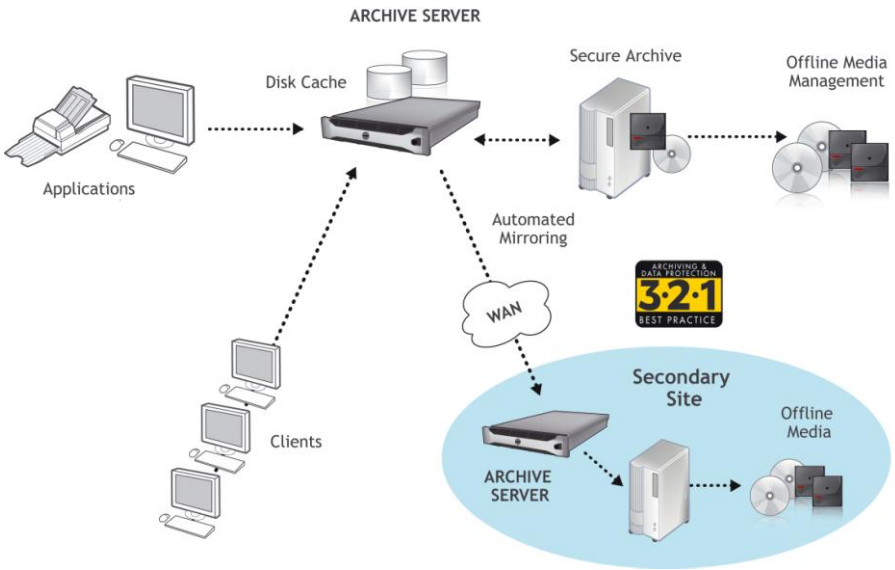


Figure 4 – Archive Disaster Prevention

10 Data and Device Refresh Cycles

Applying short-term IT solutions to a long term archive can prove to be inefficient, costly and dangerous to the authenticity of the data. In an archive environment the objective should be to develop an IT strategy that reduces the frequency of hardware and software replacement and minimizes data migration since these processes are disruptive, expensive and can compromise the integrity of the data. When content needs to be retained for many years it will be necessary to replace storage systems and migrate data to newer technology, so selecting technology with the longest possible system refresh cycles will reduce cost and mitigate risk.

Some organizations choose only RAID storage for the retention of their archive data. The advantage to these systems is their relatively low initial acquisition cost and good performance. A major disadvantage in an archive environment is their short life and high operating costs. Standard RAID systems have an operational life of 4-5 years over which time maintenance costs may increase substantially. This short life means that the RAID systems may need to be replaced many times over the life of the archive causing significant disruption and dramatically increasing long term operating costs.

By contrast, the use of storage technologies designed for long term data archiving such as tape or optical provide longer operational and maintenance life than the exclusive use of RAID storage. While no technology will enable data to be archived indefinitely, extending refresh cycles through the use of long term storage technologies is crucially important to an overall archive strategy. Choosing the most appropriate technology can mean the difference between replacing hardware every 4-5 years verses every 8-10 years. Choosing the most appropriate software and hardware technology will dramatically reduce the frequency of system refresh cycles within the context of the Active Archive.

11 Archive Management

For organizations that place a high priority on data compliance, there is a growing trend to define a role that is responsible for the compliance process. Often referred to as the “Compliance Officer”, this person takes responsibility for researching compliance requirements and for the design and implementation of both procedures and technology to meet compliance obligations.

For smaller organizations this role may be a part time responsibility, but it still makes good sense to establish the position since it can serve more than one purpose. First and foremost, it will help to ensure that the organization is meeting all necessary compliance regulations. It demonstrates corporate commitment to legal and regulatory bodies. And it better prepares an organization to respond professionally in the event of litigation. Those organizations with a Compliance Officer are in a stronger position to defend the integrity of their audit trails and data, avoiding the potential of costly penalties or jail sentences.

The more tools a Compliance Officer has at his disposal, the better able an organization is to demonstrate systemic archive integrity. This is typically a combination of clearly documented process and the effective use of technologies such as WORM recording and data encryption to ensure authenticity and document audit trails. The proper administration of an archive should not be neglected. It is a vital role that underpins the integrity of the entire archive environment.

12 Regulation and Policy Compliance

Important considerations for the management of archive data also include external regulations and internal corporate policies that define the authenticity requirements and retention periods for specific data types. Many organizations today are required to comply with government and industry regulations on data retention for legal or public safety purposes. In addition to external regulations, organizations often have their own data retention policies to ensure the availability of corporate resources and to protect against potentially damaging litigation. Understanding the role that regulations and policies play within an archive environment is essential to establishing a legally defendable data archive. Retention considerations are fairly straightforward to address since they can be included within a data profile strategy outlined in the previous section. However, data authenticity is a broader issue that demands additional consideration.

True data authenticity can only be established through a combination of processes and practices that provide well documented audit trails for the management of archived data. The physical storage technology also plays a crucial role. In industries where data authenticity is most critical, some regulations actually call for the use of unalterable, Write Once storage. WORM (Write Once Read Many) media is available in tape and optical formats, providing a level of data authenticity that is unmatched by rewritable magnetic disk technologies that use software based WORM emulation. In

addition, data encryption tools can be deployed at a file or media level to control access to the archive data to further strengthen audit trail management. The use of WORM technologies provides a strong foundation for compliance by created a well defined and defensible archive strategy that incorporates best-of-breed technology for long term record authenticity.

13 Total Cost of Ownership

Cost is always a key factor in the design of any IT solution and an archive is no exception. However, an archive is a long term proposition so costs must be evaluated over the entire operating life of the system if it's to have any real relevance. Simply assessing acquisition cost of raw storage capacity provides very limited insight into the true financial investment of archive deployment and ongoing operation. The best approach when assessing the cost of an archive is to look at the Total Cost of Ownership (TCO) over time. This must includes the consideration of the following factors.

- Initial Software Acquisition
- Initial Hardware Acquisition
- Initial and Ongoing Media Acquisition
- Ongoing Software Maintenance
- Ongoing Hardware Maintenance
- Periodic Hardware and Software Refresh Costs
- Administrative and Management Overhead
- Computer Room Floor Space
- Ongoing Hardware Power Consumption and Cooling

Some of the above TCO criteria are fairly straightforward to capture and assess while others may be more elusive and subject to a range of variables. None the less, each of the TCO criteria is important in creating a comprehensive financial perspective and should not be neglected. Using these criteria as part of a TCO model can provide an excellent vehicle for comparing different archive alternatives and discovering the financial strength and weakness of different solutions.

14 Active Archive Design Summary

While this document highlights many of the key issues that should to be considered in the design of an Active Archive, the overall process need not be complex. It is essential to begin with a clear understanding of the business objectives and to apply them within the context of a structured framework. As a design goal, "flexibility" should be the watch word for any archive. Operating over the course of years or decades, a rigid archive infrastructure will not be able to adapt to the changes that will certainly happen over time. Locking an archive design into a single technology, vendor or application will almost certainly result in compromised success and snowballing costs.

Active Archives are now an essential resource for businesses to execute their objectives and mitigate regulatory and legal risk. An archive design demands careful planning with due consideration of key configuration issues, but the long term benefits derived from well constructed archive strategy are tangible and compelling.

15 QStar Technologies

Since 1987, QStar has delivered enterprise class data management and archival storage software solutions to customers around the world. QStar consistently meets increasingly sophisticated requirements with an industry leading technology platform, which has the capability and flexibility to meet the demands of today's challenging business climate.

QStar's Network Migrator and Archive Manager software are part of a complete archive and data management platform that is hardware, system and data independent. This unique architectural approach enables customers to optimize their existing IT infrastructure while minimizing disruption and capital expense. With thousands of customers across a wide range of industries, QStar provides strategies and solutions to manage a changing technology landscape while protecting valuable digital assets for the future. www.qstar.com