In the last few years data archives have come to the forefront of IT storage management issues of interest. Specifically, the issue has been, how to manage growing archives cost-effectively while providing maximum data integrity and multi-user access in an efficient manner. A number of hardware companies have jumped into this previously mundane market with new storage platforms and technologies. This paper will explore the building blocks of archive storage and how to best select the technologies suitable for your application.

Policies

Policies are an important part of the archive system. They are needed to manage who has access to data archives, what gets archived, where it gets archived, how long the data must remain on-line, near-line and off-line before it is destroyed. However, policies are not enough to meet all regulations, nor are they enough to safe guard on their own. Corporations are also forced to determine what archive storage platform best serves their needs.

The first things to consider regarding the storage platform are the problems you're trying to address. What are the key issues that need to be overcome?

- Is there too much data to manage effectively on the current system?
- Is there too much data residing on costly high performance RAID?
- Can and how much data will be stored off-line?
- What will be the tracking method if data is stored off-line?
- How often will archive data need to be accessed, and what will the network traffic look like?
- Are there automated policies in place to migrate data to archives?
- Are there key data storage compliance issues that govern your application?

The latter needs to be strongly considered, as this may have a legal impact on your company should fines or penalties be levied against your non-compliance. Currently there are some 4,000 regulations regarding how and how long archive data must be retained. Not all storage solutions comply with these regulations and it is strongly recommended that your IT manager understand the rules that govern your industry to avoid fines and penalties.

Compare your needs against usable storage capacity in the chosen device. During this process you will want to analyze storage costs over archive time periods. Hardware costs will include the initial acquisition of the storage hardware, the storage media and any necessary storage servers. Software costs include the initial acquisition of storage management software and application software. In addition you will need to analyze annual maintenance costs for both hardware and software.

The Trustworthiness of archives

The key to data archive effectiveness regarding compliance issues is the trustworthiness of the data itself. In order for data to be deemed trustworthy, it must meet specific criteria: It must be deemed authentic-- either an original or an exact copy of the original. It must have been protected from alteration and all versions of the document must be accessible to determine the electronic paper trail. The data must be in a retrievable format and accessible for all authorized users, and if necessary, an archive system should provide necessary redundancies to protect data in the event of a disaster. Lastly, the data archive must use best practice audit trail procedures[1].

Many of the challenges to archive data trustworthiness come from the following; data stored on alterable media in a system unable to track data access and changes to data, the risk of record alteration due to frequent data migration from media to media, the lack of documentary evidence over the records data life cycle, and insufficient protection from alteration and or deletion.

Challenge to trustworthiness of data grows in proportion to the length of time required to keep data active. Media technology that is time independent helps to ensure trustworthiness over the life cycle of the media. The more stable the media, the higher the data integrity over time. This is to say that the longer data can reside on the disk and be accessible to the application, the higher the data integrity is over time. Conversely, if it is necessary to transfer the data to a new disk each time the technology is updated or replaced the larger the chance is that the data can become lost or corrupt.

Access Time

The necessity for quick access reduces with time for most data structure. Studies show data access drops significantly after the first 30 days of creation. Most types of data become fixed content after 60 to 90 days [1].

Legal discovery normally requires response in days, weeks or months, so there is usually no reason to have archives on fast access media. It only needs to be accessible, not accessible in seconds or even minutes. This said, removable media performance is appropriate for archive grade data once it has reached the end of the data life cycle.

Most archive systems, regardless of technology type, provide reliable near- or on-line access to data stored on the system at a much slower rate than on-line primary storage systems. The obvious advantage to a hard drive based archive system is fast access, however, the necessary data protection overhead slows the system to the point of having no advantage over archival optical technology. Depending on the data access requirements, speed is normally not an issue, however long-term non-alterability is almost always the issue.

Media Conversion Impact over Time

Archives are required to be available over time. The amount of time can be either externally imposed or self imposed, however, no data storage technology will last forever. Therefore you must consider the media conversion issues and the associated liabilities. Each time data is copied from one media to another it presents a risk to record integrity and demands verification and audit trail management. In the event archives are stored on non-removable media, data must be migrated every time technology is upgraded. This is also the case if data is stored on removable media that is not readable in future drive technology. During the migration process there is a risk of loosing or altering data unknowingly. By selecting a media that reduces migration frequency dramatically simplifies archive management and increases the chain of trust for electronic records.

Considerations for Archive Media:

- A durable medium- designed to be relatively impervious to environmental contaminants and protected by a robust cartridge
- Non-rewriteable non-erasable media- offering protection of electronic records at the media and storage management component level
- Removeability- offering the ability to off-line inactive archives or near-line active access to electronic records. Also facilitates in creating and retaining disaster copies of electronic records.
- Media longevity- the ability to archive data for longest possible shelf life of any digital media (Ideally the media shelf life spec should be twice as long as the data need be available)
- Backward compatibility- a history of successfully providing the ability to read older media generation with newer write/read generations. This reduces the number of data migrations, thus reducing the risk of data loss.

System hardware upgrades can be both disruptive and expensive on fixed media. If the media is fixed such as hard drives, data will need to be migrated from failing disks to newer disks periodically. Hard drive based solutions require redundancy to protect data because hard drives are relatively volatile. This translates to higher costs, and it can also impact performance. If parity is used to protect data, there can be a significant performance hit. Mirroring is also an option, however mirroring increases the cost of storage by a factor of 2. Other operational conditions can affect performance, as is common when adapting an erasable media for write-once applications.

If the media is removable and can be read in future generation drive technologies, then the data remains stable and there is little or no disruption during the upgrade process. As with optical technology, disks have been readable in drive technologies over a period of more than 10 years and 5 generations of technology upgrades. The data archive will only last as long as the media it is stored on. Data archives have to be able to out live the application, operating system and the hardware they were created on. The reason for this is, data archives need to be around 10 to 100 years from when they were created. Computer systems are not designed to be around that long, and are generally updated and/or replaced every 2 to 5 years. Ideally the media selected should be portable from system to system without having to migrate the data from one media technology to the next.

Tape media, though removable and upgradeable is still somewhat manufacture and standards dependent. It is also quite sensitive and volatile. Tape cartridges must be tested and refreshed periodically or there is a high risk of data loss. During the refresh process, data must be migrated from potentially failing tapes to new tape cartridges. Again this process can be disruptive and could result in data loss. Tape is not random accessible and can be very slow to access specific files on a cartridge or tape set. Removable media is critical for off site storage and disaster recovery plans. With

removable media, application software can track off-line data and redundant sets of media can be vaulted in a safe location for future access in the event of a disaster. Hard drives can and are mirrored to remote locations, however this requires a redundant set of storage hardware and infrastructure that is costly, and may not be necessary for archive data.

Long-term cost of ownership

Some hard disk based archive solutions advertise that the disk drives will be upgraded over time before or at the time of failure. This is seen as part of a self-healing feature. The reality is, over time drives will fail. To predict how often this will occur, it is a simple matter of multiplying the number of drives in the system by the predicted failure rate of the drive to determine the chance of failure in a year. For example, if a drive manufacture specifies a 2% failure rate, and the system has 64 drives in the archive system, there is a 128% chance of at least one drive failure in the first year. Prior to, or at the time of failure, the drive must be replaced and the data restored to the new drive. This could translate to the risk of frequent data migration and jeopardize the data integrity. Most hard drives have an expected serviceable life of 5 years or less, yet much of the data governed by regulations must be kept for more than 10 years. This means that archive data is likely to be migrated at least twice during the archive period on a hard drive based system.

Energy costs are also a consideration for long-term cost of ownership. Hard drive technology spins constantly regardless of usage. This can add significantly to the total cost of ownership depending on your country or region. The cost of constantly powering fast access hard drives for data that may not be accessed on a regular basis may be a waste of money. Again, removable media technology provides a more economical solution, in that, drives only spin up when they need to read or write. This adds latency to data access, however the operational cost savings can be significant. In addition, media that is moved off-line uses no power whatsoever and is mounted only in the event of required access. Additional media can be put in the library after aged data is moved off-line, thus scaling to the needs of growing databases at the added cost of media with no additional hardware investment.

Proprietary API's for individual storage devices can also add to the risk of long-term ownership of data. Application providers are on the hook to support proprietary API's from companies that provide compliant or pseudo-compliant systems. For each supported API the cost of ownership on the application goes up. A better way is with a heterogeneous interface allowing for a single share point to all supported storage hardware, both primary and secondary, making interaction between devices as easy as drag and drop or copying between directories. This interface results in compatibility with many storage technologies from multiple vendors, allowing the user or automated systems to migrate archive data from legacy hardware to any supported archive storage product.

The bottom line is identifying what the business problems are that you are trying to solve. Once you determine the problems you can focus on the technology to solve the problems. Identify what regulation requirements your business needs to meet, and select your storage technology accordingly. Look at data access needs, how often will data need to be accessed and how long will it need to be accessible. If data doesn't need to be accessed often, you might not need high performance storage devices. If data needs to be stored for more than 5 years a technology that supports longer retention periods may be more suitable. Lastly try to rely on hardware and software that support open application program interfaces (APIs) and industry standards.

[1] Cohasset Associates Inc., "Trustworthy Storage and Management of Electronic Records", (2003): 8-9