WHITE PAPER

Time Value of Data

Creating an active archive strategy to address both archive and backup in the midst of data explosion

April, 2014

By Floyd Christofferson, SGI

TABLE OF CONTENTS

1.0 Coping with the Information Explosion

In 1964, the New Statesman magazine first used the expression “Information Explosion” to describe what was thought at the time to be a dramatically increasing deluge of data, and the problems that this created. Then, of course, the flood was mainly data on paper, and the increasing proliferation of documents in all industries was becoming a nightmare to manage.

Thus the digital information technology era brought a welcomed relief from the problems associated with sorting through, finding, protecting and archiving an increasingly large mountain of paper-based information.

Or did it?

The United States Library of Congress contains hundreds of millions of items, including books in 470 languages, manuscripts, government publications, newspapers and magazines from all over the world, 2.7 million sound recordings, millions of photos, films and videos in a collection that fills about 745 miles of shelving. But for all it is estimated that, if all the print holdings were digitized they could be stored in only about 20 terabytes of storage space. All told, the entire collection of digital content at the Library of Congress is only about 3 petabytes.

All of the printed material on earth is estimated to be about 200 petabytes (200,000 terabytes) of data. That is a lot of information and a lot of data to manage. But it is dwarfed by the amount of digital data that is generated worldwide today.

According to IDC estimates, in 1999 the total worldwide volume of data was estimated at 2 exabytes, or two million terabytes. By 2011, this number had grown over by 900 times, to about 1800 exabytes (1.8 million petabytes). But the network effect of the digital explosion is that the rate of data growth continues to grow dramatically year over year to the point that even analysts are struggling to size it.

So while information technology has alleviated the problems of finding and sorting paper, it has also facilitated the explosion of data creation and the exponential growth of the problems associated with managing, sorting, protecting and archiving data.

Increasingly, the data explosion is in unstructured, file-based data. Eighty percent of Fortune 1000 companies surveyed by the Taneja Group said that more than 50% of their data is unstructured. Being unstructured means that these increasing volumes of data are not organized. And as data volumes grow, so do the costs and complexity of managing such data. Something as seemingly simple as deciding which data are OK to delete can be an overwhelming challenge at when faced with millions or billions of files.

In an analog world, an archive is a resting place for old content that is not currently needed. The problem is that when data is put into an offline archive, it becomes difficult to access, and complex to manage.

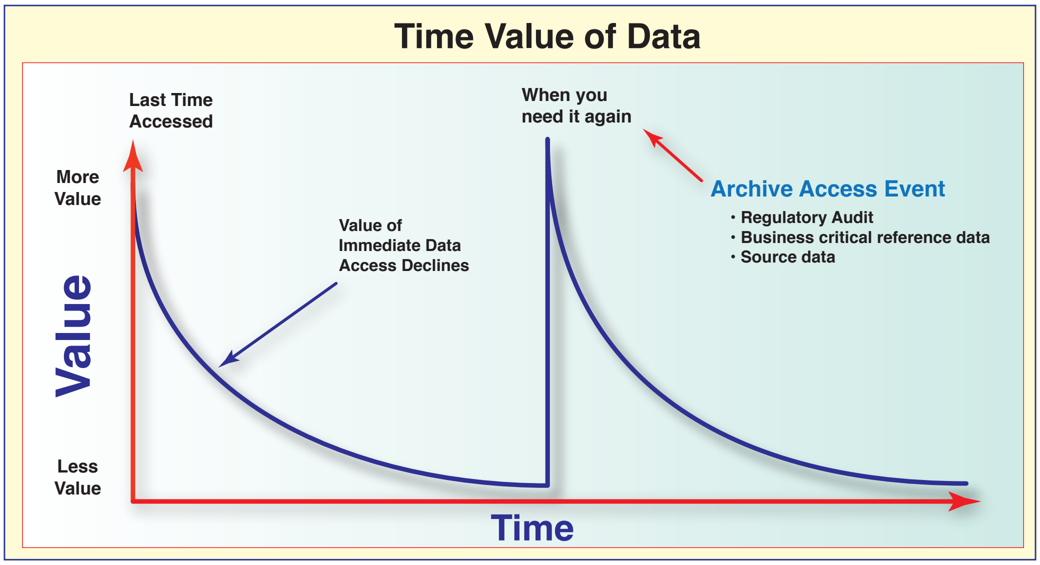
But in a digital world, all data needs to be easily accessible at any time. Whether it be for Big Data analytics or for the monetization of data, ensuring that it is online and available in a cost-effect way is the key first step to getting more value out of digital assets.

2.0 Time Value of Data—The Backup/Archive Dilemma

Taking proactive control of data is an essential requirement to realizing its full value. In a world where Big Data analytics is revolutionizing businesses in all market segments, ensuring that the data is online, available and also protected is essential. Only with the proper strategy and the necessary hardware and software tools is it possible to make the management of active, inactive and protection of data seamless and essentially transparent to the user. Only when looking at all three of those data types as a whole will the full "Time Value" of data be realized and be manageable. As an added benefit, this approach usually results in reduced operational costs.

The Time Value of data is just what it sounds like: When data is fresh, it is typically more valuable. The value of some data decays over time (old receipts, outdated reports, etc). The value of other data increases over time (NASA footage of the first men walking on the moon). The methods for understanding the relative value of the data is not as simple as these examples might suggest, however. What if the outdated report is also a necessary bit of evidence needed for compliance or historical purposes? At what point does one decide that it is not just an outdated file?

In addition to understanding the Time Value of data, it is also important to understand whether that data needs to be active for instantaneous access, or whether a delay of a few minutes, or hours is sufficient for retreival. According to IDC, 40% of fixed data is active or is accessed infrequently. Forrester Research says that 85% of production data is inactive, with 68% having not been accessed in 90 days. So while such data needs to be accessed sometimes, it doesn't need to be filling up expensive production disk capacity.



The immediate value of data will go up and down over its lifecycle.
This is called the Time Value of Data.

So, the Time Value of data includes not only whether it's important over the long-term, but whether the data needs to be immediately accessible. When these variables are not proactively managed, and when backup (protection of active data) gets confused with archive (retention of inactive data), data paralysis occurs. Backup and restore times become difficult to manage because they involve inactive data as well as active data. Seldom-accessed data becomes hard to find. Operating costs skyrocket when additional production disks are needed just to keep up with the relentless growth of data.

What's worse, the excessive growth of data stored on primary or production disk arrays can be a contributing factor to data becoming segmented into incompatible silos. This makes collaboration between different data silos either impossible, or at best a manual process prone to error and wasted effort.

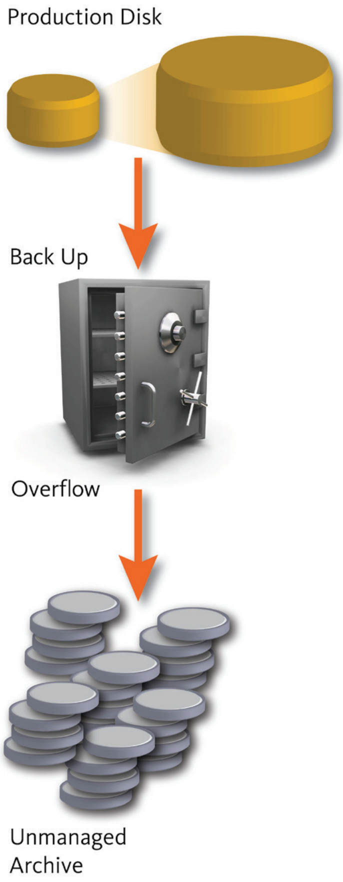
Users deal with files but are forced to work within file systems. The job of a proactive data management strategy, or an active archive strategy, is to let the users focus on their work, and not waste time, infrastructure, or energy on managing the data they need in order to do their work. In other words, an active archive strategy keeps data online and accessible to users without the complexity and cost associated with conventional one-tier architectures.

3.0 Key Concepts—Distinguishing Between Backup and Archive

At the heart of resolving these problems is sorting out an integrated approach to archive and backup. This makes data accessibility the priority for users while still allowing for cost containment and data protection priorities for IT managers.

The problem is that backup and archive are often confused. Not necessarily in concept, but in day-to-day practice.

The process often goes something like this:



As data continues to grow, primary or production disk arrays fill up and must be expanded. As noted above, this is typically a mixture of active data with older, seldom-used data.

Backup is needed to provide protection for primary disk, and as the overall data volume grows, the backup windows grow as well.

Whether it is due to the inability to even get backups done within the available time windows or because the backup environment simply becomes too big, IT managers sometimes even resort to pulling excess backup data and putting it on a shelf as an "archive." The problem is that this "archive" is not a true archive at all, and is often unmanaged. Data with a high Time Value is mixed in with low-value data.

Also, when archive data is taken from backup data it becomes haphazard and incomplete. The data, which may have significant Time Value, is offline to users and is often irretrievable without significant cost, effort and time.

Worse, this approach means that the backup environment must continually grow to keep pace with the expansion of the production disk environment. That adds costs without actually solving anything.

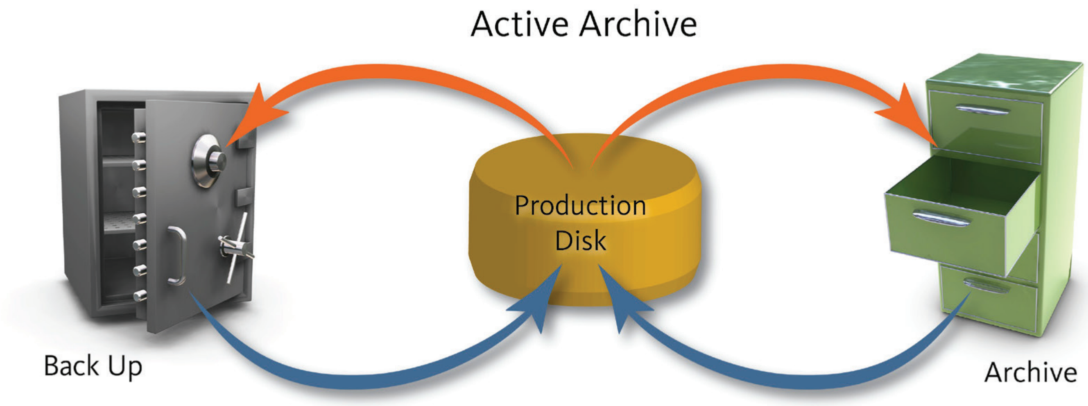
This is a common problem across most industries and impacts even those who are already extremely careful with their data production and protection. Indeed, it is because of the high volume of data and the difficulty in managing the distinction between high-value and low-value data that many IT managers are left with little option but to keep everything, which merely compounds their problem.

The solution is to create a clear distinction between backup and archive, and to decouple the needs of data protection from that of data retention.

Backup	Archive
Copies data for protection	Moves data off expensive primary disk
Supports short-term operations and recovery	Supports long-term business value & compliance
Supports availability	Supports operational efficiencies.
Short term in nature	Long-term in nature
Data typically overwritten	Data typically secured, not overwritten
Poor solution for regulatory compliance	Ideal solution for regulatory compliance
No historic relevance	For historic information
Not easily searched	Easily Searched

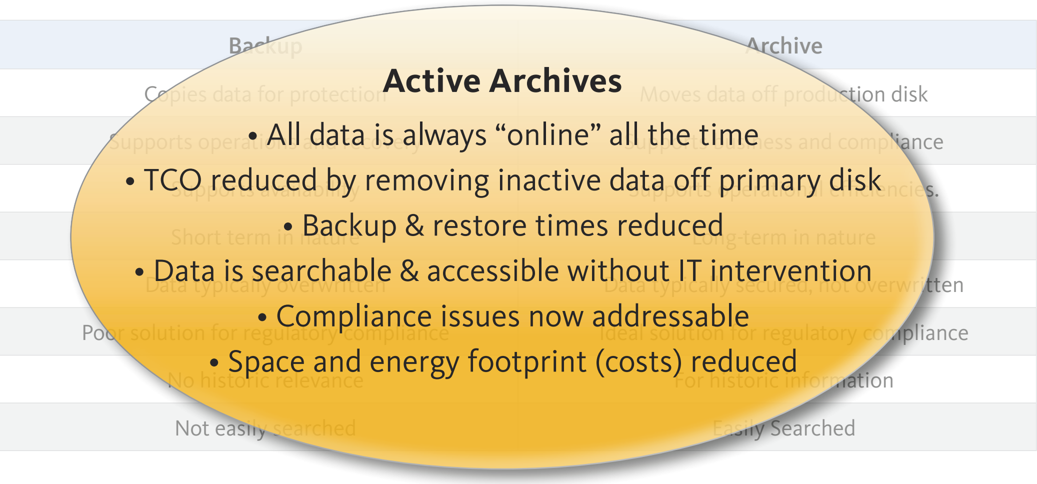
Backup strategies should be for short-term production data. They are to protect what is done in the short term in case of catastrophic failures. Archive or data retention strategies on the other hand are long-term by nature. Disaster recovery protection is still needed for this data, but does not need to be done within the tight time window required by backup.

4.0 Building an Active Archive Strategy



An active archive is one in which all data is always available in an 'online' state all the time. An 'online' state does not mean that it is taking up expensive primary disk capacity. In the context of an active archive, 'online' means that the data is virtualized across multiple tiers of different types of much cheaper data storage, decoupling it from higher priced production disk arrays.

In fact, when properly applied, an active archive strategy significantly reduces the overall storage and data management costs while at the same time increasing efficiencies and the ability of users to access all data. Using an active archive strategy, costs are contained because the production disk does not need to grow very often. Inactive data that still has retention value is proactively classified and moved into second or third tier storage which, although 'online' and visible to the user, is stored in tape, object or other lower-cost solution. These archives, while still online and available to users, can be managed with very different disaster recovery techniques to dramatically reduce costs.

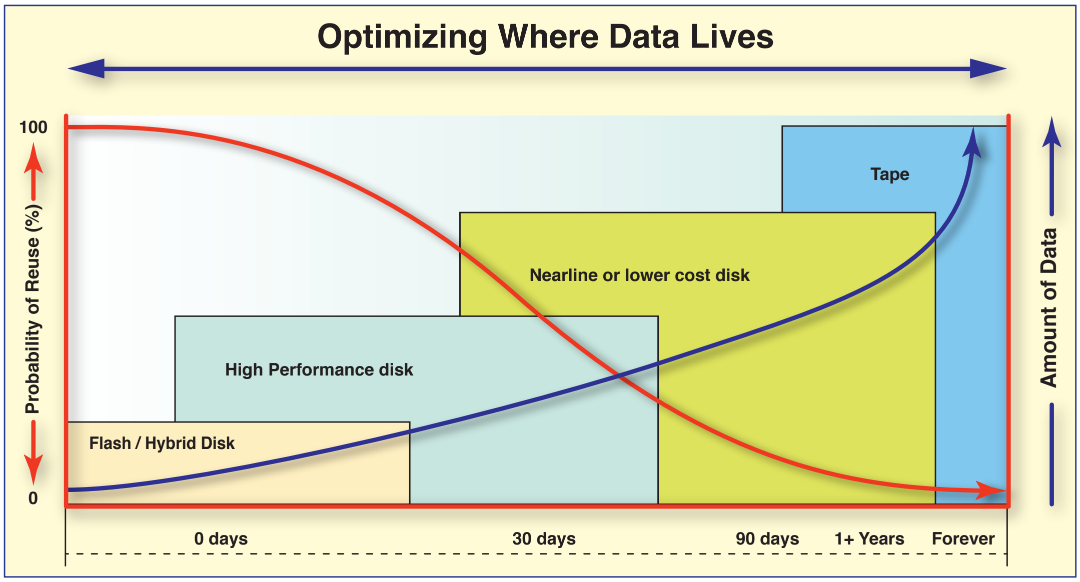


Active Archives

- All data is always "online" all the time
- TCO reduced by removing inactive data off primary disk
- Backup & restore times reduced
- Data is searchable & accessible without IT intervention
- Compliance issues now addressable
- Space and energy footprint (costs) reduced

Backup	Archive
Copies data for protection	Moves data off production disk
Supports operations and recovery	Supports business and compliance
Supports availability	Supports operational efficiencies.
Short term in nature	Long-term in nature
Data typically overwritten	Data typically secured, not overwritten
Poor solution for regulatory compliance	Ideal solution for regulatory compliance
No historic relevance	For historic information
Not easily searched	Easily Searched

Thus, data lives where it is most efficient. Online, active data is contained only on the primary disk arrays. Data is proactively classified so that which has a lower Time Value is automatically migrated off of production disk while still being available to users. Since data growth is not managed by increasing production disk, backup of those arrays does not mushroom either. Backup is reserved only for active data, keeping costs down while reducing backup and recovery times.



As data becomes inactive over time, it doesn't need to live on expensive storage. An active archive strategy automatically places data in the right tier as its Time Value changes.

5.0 Implementing an Active Archive

There are numerous tools that can be employed to implement an active archive strategy. These will vary by industry, by use case and workflow. Not all are needed in every circumstance. In fact, what is most important in devising a strategy is to approach data growth proactively from this whole-system perspective rather than reactively by throwing more disk to solve short-term problems. As we have seen, short-term solutions compound the overall problems, always leading to higher costs and risk.

Some tools to consider:

- **Digital Asset Management Solutions:**

A key problem in determining whether data is active or not is in the strategy used to classify it.
This problem is compounded when production data is distributed across multiple silos.

Leading digital asset management solutions such as LiveArc™ from SGI enable content to be indexed automatically in multiple ways as it is created and modified. Users can search for data, administrators can easily set policies to determine which data should remain on production disk and which can migrate to second or third tier storage.

Another key benefit of a digital asset management platform like LiveArc is the ability to bridge multiple namespaces, or data silos, to provide a global view across all storage, data and metadata types. In this way, IT managers have complete control over their environment and can implement back-end changes without impacting users. Users don't need to know or care where the data actually is in the hierarchy of storage infrastructure because it is always visible to them within the management interface on their desktop.

- **Storage Tier Virtualization Solutions**

Another key practice which can aid in developing an active archive is to virtualize tiers of storage with solutions such as SGI DMF™ and SGI StorHouse®. DMF and StorHouse enables multiple tiers of disk and tape to appear to the users as one large aggregated volume even though the data is actually distributed across multiple storage types.

For example, production disk is typically higher performing (and thus higher cost) disk, according to the needs of the particular use case. But since only a fraction of the data is active, expensive primary disk is used to house inactive data.

With solutions like StorHouse and DMF, the expensive high performance disk is linked with “nearline” or cheaper, capacity disk. This in turn can sit in front of a tape library or cloud storage.

The beauty of this system is that all the data appears to the user to be online in the expensive production disk all the time. But in reality, even though the file appears to be right where the user put it in the filesystem, it is actually migrated to lower cost disk which results in dramatic overall savings, without the need for users to wonder where their content is.

With an active archive solution, the rules for when or if data moves downstream to lower cost storage can be established by policies, such as file type, time since it was last accessed, etc. In addition, since DMF can manage multiple copies of the same file, backup becomes optimized to a significantly smaller amount of data.

An active archive strategy requires planning and tools, but when done properly can dramatically reduce the overall costs of managing a growing pool of data. More importantly, by de-coupling production disk from other tiers of storage, single points of failure are virtually eliminated. Individual components can be upgraded or changed without impacting overall utilization for users. Scalability becomes an asset in this scenario, not a headache.

6.0 About SGI

SGI, the trusted leader in high performance computing (HPC), is focused on helping customers solve their most demanding business and technology challenges by delivering technical computing, Big Data analytics, cloud computing, and petascale storage solutions that accelerate time to discovery, innovation, and profitability.

For more information please contact an SGI sales representative at 1-800-800-7441 or visit www.sgi.com/contactus.

Global Sales and Support: sgi.com/global