White Paper

IBM's Information Archive

Addressing Long Term Information Retention Requirements

By Brian Babineau

January, 2010

Contents

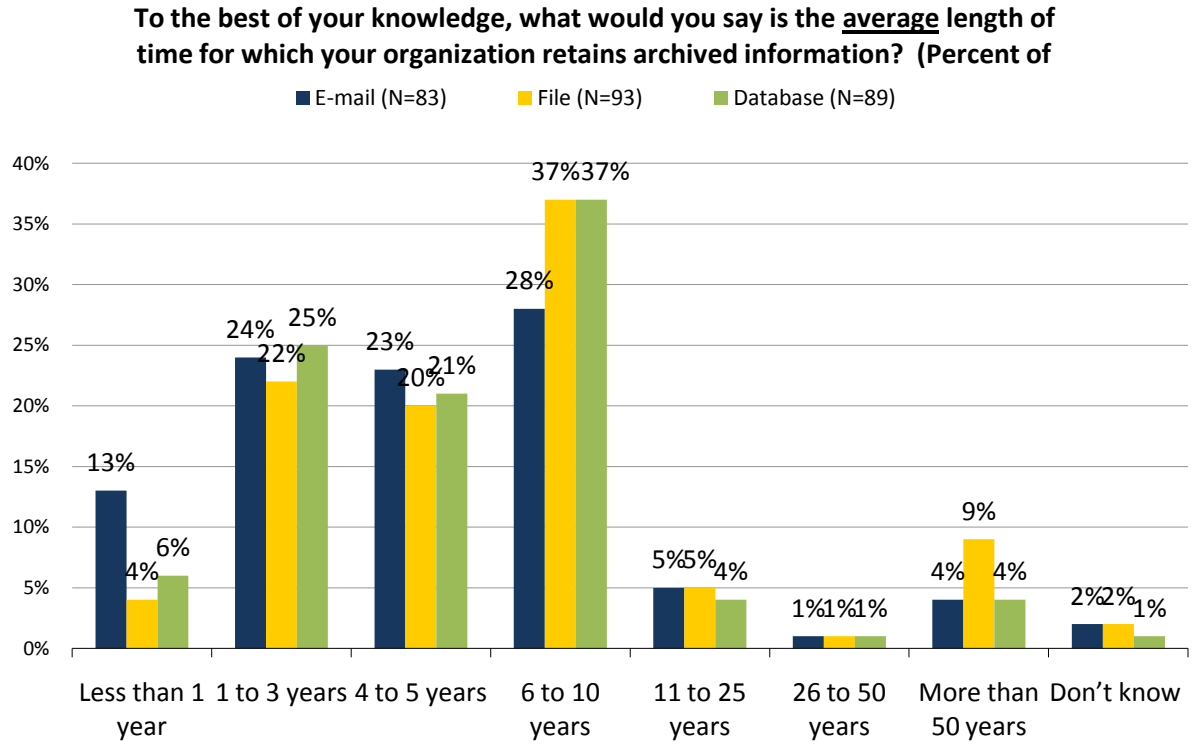
Introduction

Organizations retain information for two reasons: because they want to (it has some inherent business value, business intelligence, analytics, reporting, etc.) or because they have to (due to regulatory requirements, legal discovery, etc.). These issues are exacerbated by data retention due to a lack of formal ways to determine what is or is not needed in addition to a lack of enforcement of consistent deletion policies. As a result, it is hard to find a business that doesn't have a reason to save information. On the other hand, it is easy to find one experiencing difficulties with reigning in information management costs.

The process of retaining information, more commonly referred to as "archiving," is challenging for several reasons:

- Companies continue to generate new information, at least a subset of which the business wants or has to save.
- Information is generated by multiple applications in multiple formats. E-mail, instant messages, productivity files, database records, SharePoint sites, wikis, and videos are just a few examples of information types that may need to be archived.
- Because of the importance of the information being archived, organizations have to treat it as "mission critical" and put proper data protection and business continuity processes in place to prevent downtime or data loss.
- Companies have to secure archives in accordance with industry- or government-specific privacy regulations.
- Archived information is kept for a long time. ESG research suggests that more organizations are saving archived information between six and ten years than for any other period of time (see Figure 1).[1] With most IT departments viewing the world of technology in three year increments, cost effectively storing and managing archived information requires an entirely new perspective.

Figure 1. Average Archive Retention Period By Content Type (E-mail, Database, File)



Source: Enterprise Strategy Group, 2007.

[1] Source: ESG Research Report, *E-mail, Database, and File Archiving Surveys*, November 2007.

Organizations can tackle a few of these issues immediately by deploying information-specific archiving solutions with the storage of their choosing. However, cost and complexity arise when another information type has to be archived and, due to legal preservation reasons, needs to be saved on immutable storage for a period that differs from the original information type. These individual archive stovepipes may be supported by a number of technology vendors, which can lead to management complexities, higher cost, and uncontrollable risk. A short-term perspective on information archiving can easily create unnecessary expenses and operational inefficiencies.

IBM's Smart Archive strategy, bolstered by a series of hardware, software, and services offerings, is designed to help customers transition from short-term archiving decisions and their associated risks to a longer-term, simpler approach to information retention. The offerings enable organizations to unify several aspects of the archive process (including the collection, analysis, management, retention, storage, and access of information) while providing a variety of technology consumption models inclusive of integrated appliances, managed services, and cloud-based services to help archiving fit into existing business and IT operating procedures—and not the other way around.

Archiving strategies such as IBM's allow businesses to implement individual products, safe in the knowledge that there are supporting and complementary offerings available for future deployments. Because archiving involves saving information for extended periods of time, the storage infrastructure is a logical place to start: the right storage infrastructure will help reduce costs resulting from the challenges outlined above. This paper details these costs, as well as how IBM's Information Archive—an integrated archive storage appliance and a cornerstone of the Smart Archive strategy—addresses them.

Information Retention Realities

Multiple Business Drivers

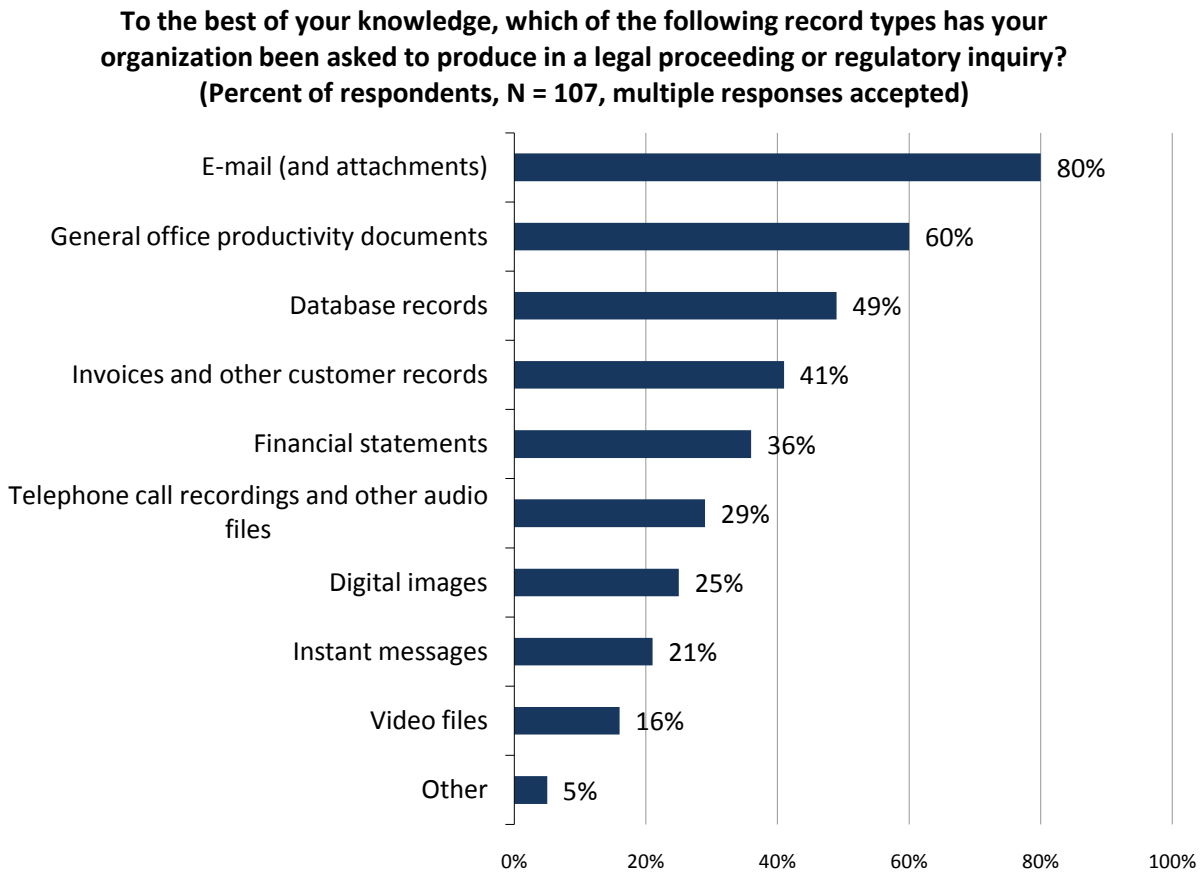
To understand storage-specific archive costs, organizations must first understand why they are saving information. ESG groups archiving drivers into four major categories:

- ***Compliance***. Almost every business, regardless of industry or geographic location, is subject to record retention regulations. Record retention processes must transition to incorporate digital information, but doing so also introduces ambiguity. Most rules mandate which records must be saved and for how long, but they do not specify how to do it or which technology to use to meet requirements—yet failure to comply with record retention requirements can result in extensive fines, severe damage to the business reputation, and business dissolution.
- ***Corporate governance***. Corporate boards and executives often define policies and practices separate from legal compliance to ensure accountability, proper reporting, and fulfillment of their responsibilities to shareholders and stakeholders. Information retention processes are frequently included in corporate governance initiatives to demonstrate diligence and fiduciary responsibility.
- ***Electronic discovery***. There are two retention catalysts when it comes to saving information for legal purposes. The first occurs when information is relevant to a specific legal or regulatory matter and must be retained and preserved until the matter is resolved. The second involves corporate policies that prevent the deletion of certain information in the event that it is needed to support or defend a claim. These mandates often lead to everything being saved, forever.
- ***Business reference.*** Organizations recognize the value of institutional knowledge and strive to retain information about products, processes, customers, programs, etc. for future use. To be useful for business reference, information also needs to be appropriately cataloged and classified to facilitate easy searching and retrieval. The more information available to employees, the better their decisions will be. In this situation, many companies are simply trying to figure out how to keep more information accessible at lower costs. One of the easiest ways to do so is by archiving or moving the content from the primary application infrastructure to a lower cost one. Companies are also trying to catalog or index as well as classify information to facilitate simple search and retrieval.

Dealing with Unique Circumstances

Due to various business drivers, information retention requirements—and thus storage needs—can differ widely. As an example, companies often have to archive several different information types from a variety of sources. This occurs regularly in electronic discovery: companies are asked to identify, preserve, and produce any information relevant to a specific topic or date range. Potential evidence may take the form of e-mail, video, or a database (see Figure 2).[2] An archive storage system must be flexible enough to support all these content types: structured (database, unstructured (files), and semi-structured (e-mail).

Figure 2. Frequently Requested Record Types During Electronic Discovery Processes



Source: Enterprise Strategy Group, 2007.

Organizations must also cope with a myriad of retention policies and parameters, including government- and industry-specific rules requiring different records be kept for varying lengths of time. When dealing with multiple electronic discovery matters, there is no formal timetable for legal holds, further complicating retention policies. After figuring out what information to keep and for how long, organizations must then decide what to do when a retention policy does expire: delete the information permanently or keep it?

Some archived information must be stored in such a way that allows its authenticity and integrity to be proven. This occurs frequently when information is archived for compliance and electronic discovery purposes and can be facilitated by a storage system that supports encryption and immutability (Write Once, Read Many data formatting) configurations. Other content types may not have such strict requirements, but companies choose to leverage these formatting and security options in the interest of data protection. There are, of course, many instances where archived data just needs to be kept somewhere.

[2] Source: ESG Research Report, *Electronic Discovery Requirements Escalate*, November 2007.

The biggest potential obstacle organizations face when archiving information is balancing accessibility requirements with storage costs. The conundrum exists because certain situations, such as electronic discovery, require archived data to be more accessible than business records that are in their 20th year of a 30 year retention policy. From a compliance and corporate governance perspective, companies may choose to keep recent records more accessible than those created a few years ago to support internal and external audit processes. More commonly, data saved for business reference is usually kept reasonably accessible. The issue is that the more accessible the storage media, the more it costs. From an archive perspective, "accessibility" usually leads to a comparison between disk and tape. Disk is typically faster, especially when application access is random, but tape is less expensive. Companies need to figure out which information is best for each media type and how long it should be kept there.

Costs Abound

Companies planning to save more information for longer periods of time have to budget for more storage capacity. However, acquisition costs are not the only expenses involved in implementing an archive. Because they contain valuable information—in some cases, vital business records and evidence—archives need to be protected and require backup and recovery systems. If archives support mission critical processes, organizations should also account for disaster recovery solutions to mitigate the risk of downtime.

There are several operational expenses associated with an archive storage environment, including:

- **Power to run and cool the storage systems**. Depending on a data center's power constraints, this could be a fairly sizeable expense. Adding new devices can be disruptive if a company must procure additional energy or reorganize the facility to optimize airflow.
- **Management (labor).** IT has to find ways to set up, provision, configure, and operate the archive storage infrastructure, including the data protection and disaster recovery systems. In archive environments, resources are often needed to constantly add more capacity due to data growth, customize interfaces so that a system can support a specific type of content that has to be saved, and update configurations to address various retention period requirements.
- **Migrations**. Most organizations include migration expenses in their labor costs, but archive environments involve larger (due to the amount of data) and more frequent migrations. These migrations are driven by retention periods that extend beyond the useful life (or the warranty) of the underlying storage system. Additionally, some IT departments proactively move data between systems—migrating older data from disk to tape, freeing capacity for fresh content. Any form of data movement can be extremely resource intensive as IT has to set up the target and source systems and then monitor the progress to ensure that no data is lost during the migration. Losing data creates unnecessary legal and compliance risks.

IBM's Information Archive Creates a Robust Foundation

Overview

IBM's Information Archive is a new unified and secure storage repository for archived information. It is a critical component of the IBM Smart Archive, which is a comprehensive, unified, integrated, and information-aware *archiving strategy* from IBM. IBM Information Archive is a factory integrated storage appliance capable of storing up to 304 TB of raw data on disk. With disk drive capacities constantly increasing, ESG expects this number to increase substantially over the next 12 to 18 months. In addition to the disk, Information Archive also supports hundreds of IBM and non-IBM tape systems, which increases the solution's logical capacity into the petabyte range. Customers can start small and add capacity—disk, tape, or both—as their archive storage needs dictate. This scalable approach enables customers to consolidate their archive storage rather than constantly purchase new devices as capacity needs increase.

With NFS and System Storage Archive Manager (SSAM, a variant of Tivoli Storage Manager) interfaces, Information Archive can connect to a variety of data sources simultaneously, including primary applications, purpose-built archive applications including IBM's Content Collector and Optim Data Growth solutions, and general purpose

production file shares. Flexible interfaces, specifically NFS, allow customers to archive multiple content types in a single system, eliminating the need for a different archive storage solution for every data source. The Information Archive's versatile architecture, leveraging IBM's General Parallel File System (GPFS) technology, leads ESG to believe that IBM will be able to support other interfaces such as CIFS and HTTP, which will only extend the solution's reach in the future.

"Collections" Concept Extend Flexibility

To date, the market has been wary of large-scale, consolidated storage archives because of the various configuration parameters needed to support different business drivers and content types. While any company could fill up a consolidated archive, not all the data needs to be saved and managed the same way. For example, a portion of archived information may need to be kept on non-erasable, non-rewritable media while other data simply needs to be stored for a few years in a reasonably accessible format for auditing purposes. As a result, companies preferred to implement several purpose-built archive storage systems with different configurations as well as a multitude of performance and availability characteristics to address each of their information retention requirements (compliance, electronic discovery, governance, etc.) and content types. This stovepipe archive storage strategy creates more issues—including poor utilization and management complexity—than it ultimately solves.

To address the stovepipe issue, Information Archive leverages "collections," which are virtual repositories within a single appliance. In its first release, it supports up to three collections per appliance and each collection can be customized to address a specific archive need. Customers may elect to set up collections based on archive business drivers to handle different content types and data sources or some combination thereof. There is little risk in filling up a collection as each of these virtual repositories can handle an estimated one billion files or "objects" (an object is the combination of a file and its metadata). With the Information Archive, customers get multiple "virtual" archives while managing one storage system.

Three Retention Policy Enforcement Choices

After setting up the collections, customers can select one of three data protection levels to best address their archive business requirements:

- **Basic** enables applications and users to delete documents before the retention period or retention hold expires. Users can also increase or decrease a retention policy and modify the collection's protection level if warranted. This option may be used when information is archived for general business reference purposes or to meet broad corporate governance requirements.
- **Intermediate** allows users to increase or decrease retention periods, but prevents deletion before the expiration of the retention period or the retention hold requirements. Users can increase the protection level to "Maximum," but cannot decrease it to "Basic." Customers looking to save important project information or enforce strict corporate governance requirements are likely to use this protection level.
- **Maximum** only allows users to increase retention periods and prevents information from being deleted until the retention period or retention hold has expired. This protection level cannot be altered and is ideal for companies that have to archive data to meet strict regulatory compliance requirements. This protection level may be compared to storage systems configured in WORM format.

Multiple Retention Options

With each collection having its own protection level, customers must then determine what data goes in the respective virtual repositories and how long to keep that information. This is determined by retention policies, which are also configured on a per collection basis. A retention policy can be a retention period determining how long a piece of information is to be kept. When customers know how long they want to save information (which is often the case when archiving for compliance reasons), they can leverage a "time-based" retention period. For example, a company may be forced to save all client communications including e-mail and instant messages for three years from the date of their creation. In situations where the ending time period is unknown, Information Archive customers can leverage "event-based" retention periods where information is kept until a particular occurrence. For example, employee contract information might be kept only as long as the employee remains with the company.

Another example of a retention policy is a retention hold, which prevents any information from being expired even if its retention period has ended. This option is most commonly used in electronic discovery situations where content must be preserved until the matter is resolved, even if the information is no longer needed to satisfy compliance, governance, or business reference requirements.

When configuring a retention policy within a collection, a customer needs to establish criteria used to identify how content should be managed and the actual retention period or hold requirement. This is based on file metadata properties such as the owner, date of creation, and format. As an example, a company may want to retain all Excel spreadsheets created by the finance organization for one year, any images generated by marketing for two years, and put executive employment contracts on retention hold because it is part of an ongoing legal matter.

As content is moved into a collection, it is analyzed and assigned a retention policy. Fine-grained retention policy management enables customers to set retention policies on a per file basis while saving the data in a single virtual repository. Information Archive also accepts and executes retention policies that are set by applications such as e-mail archiving, records management, and enterprise content management systems connected to it.

Controlling Cost and Risk

Data Reduction

One of the easiest ways to reduce archive related storage costs is to save less information. While this may sound counter-intuitive given the amount of information most companies have and want to save, it is entirely possible via Information Archive thanks to the appliance's deduplication and compression capabilities. Deduplication ensures that the same bytes of information are stored only once (reference pointers are used to track any redundant bytes sent to the system) while compression shrinks the size of a given set of bytes. Together, they pare down the amount of physical capacity consumed by archive data, reducing storage acquisition and operating costs.

Deduplication and compression capabilities are also configured on a per collection basis, providing customers a choice in how to maximize storage utilization. This is extremely useful in situations where one of the information reduction options doesn't align with the data type being archived. Some databases compress very well, but do not contain a significant amount of redundant data. Alternatively, attorneys may take a conservative approach in saving all information, choosing to leverage compression and not deduplication.

Multiple Storage Tiers

Information Archive also controls storage costs by supporting disk and tape within the same system and automating the movement of data between the two. Within the policy engine, customers establish criteria determining when they want to move content from disk to tape or vice versa. Criteria can include data growth triggers: when a collection reaches a certain size, older data is moved to tape freeing up capacity on disk for new information. Some customers may simply move all data based on when it was last accessed.

Moving older data to tape means that customers defer disk purchases, reducing archive storage acquisition costs. Policy-based migration also minimizes the burden on IT resources that often have to copy archives between storage media options to optimize storage resources.

Security is Imperative

Information Archive includes role-based security access, audit logs, encryption, and a locking cabinet for physical security. Its Enhanced Tamper Protection capability provides an additional level of security for data that absolutely cannot be altered or deleted during its retention period. When this feature is enabled, neither the customer nor IBM has root login authority. Common administrative and support operations are pre-programmed to remove the requirement for root access. Once turned on, the Enhanced Tamper Protection capability cannot be disabled.

Together, these security features mitigate the chance of unauthorized access and data tampering while audit logs track configuration information as well as any actions taken within the system. Customers can leverage the audit log data to prove that no one intentionally or maliciously tried to alter data without the appropriate permissions. They can also be used to track the success of collections and retention policy enforcement, which is imperative when companies need to demonstrate the authenticity and integrity of archived information often needed in regulatory and discovery situations.

The combination of the security features with "Maximum" Data Protection level make Information Archive an ideal target for compliance data that cannot be deleted during a retention policy. The system stores data in WORM format, prevents unauthorized access and tracks any attempts by users to change or delete information (which, of course, would be unsuccessful given how data is stored). To further demonstrate these capabilities, Information Archive was recently audited by a third party, Cohasset Associates, a consulting firm specializing in records management. The findings of the audit determined that Information Archive capabilities were satisfactory to address SEC Rule 17a-4(f) requirements—a regulation widely known as having the most stringent storage media requirements pertaining to the retention of business records. In short, this regulation requires broker/dealers to store requisite business records in non-erasable, non-rewritable format for specified (in SEC Rule 17a-3) retention period. And, several interpretations of this regulation have stated that electronic storage media as well as combination of electronic storage media and software that store information in WORM format can be used to meet the requirement.

Note that the ultimate determination of compliance falls within the customer's responsibility to notify its regulatory examining authority when they change the storage media which stores their business records. The examining authority may choose to question and inspect the media or not respond at all—the latter of which is usually interpreted as an "acceptance" of the solution. If the customer has to produce any of the business records stored on the system, they will need to prove and defend the means in which these records are stored and preserved on non-erasable, non-rewritable storage media.

Availability is Paramount

Similar to other storage systems, Information Archive includes redundant components and RAID protection to ensure high availability and data protection within the system. To prevent data loss due to corruption, customers can back up the archived data using traditional file system methods including NDMP as well as copying the data directly to tape using the Tivoli Storage Manager component. Additionally, the Information Archive database—which contains all of the metadata and file mapping information—can be protected with standard onsite and offsite backup policies.

For disaster recovery, Information Archive supports enhanced remote mirroring, enabling two real-time synchronous or asynchronous data copies to be maintained on separate Information Archive devices. Like many other Information Archive capabilities, data replication capabilities are configured on a per collection basis. If one collection contains critical data used to support compliance or discovery processes, customers do not have to replicate the entire system, saving storage costs and consuming less bandwidth.

Simplified Management and Operations

Customers may experience the greatest cost savings through Information Archive's management efficiencies. The first set of management efficiencies comes from ease of use and set up of the appliance itself. Wizards guide the user to create and configure collections as well as establish protection levels and retention and data management policies. Monitoring and troubleshooting tasks are integrated in a single interface, enabling IT to easily track system health.

The second set of management efficiencies is the direct result of Information Archive's ability to consolidate archive storage into a single system while addressing a myriad of archiving business requirements. When customers think about protecting their archives from downtime due to a disaster, they only have to set up replication on one system. Retention policies are set up in a central location as opposed to several devices with different protection levels. Floor space consumption is kept to a minimum and power draw can be minimized when customers deploy Information Archive with tape.

A Good Start to a Smart Archive

A New Approach

Organizations can continue to save information the way they always have: save everything forever, keep it on tape, and hope that no one ever has to access it. While this approach may sound simple, it is extremely complex and cumbersome as information retention requirements change. IBM's Smart Archive strategy is designed to help companies maintain a simple approach to information archiving while maintaining the flexibility to address whatever information archive challenges exist today and what may arise in the future.

In order to keep things simple, IBM's Smart Archive strategy includes products (hardware and software), services (consulting, implementation, etc.), and consumption models (appliances, cloud, etc.) so that customers can figure out what combination is best suited to meet their information retention needs. As an example, customers may turn to any number of IBM's Information Management software offerings to unify the identification, analysis, classification, and management of several information sources. In turn, these software solutions can send data into the Information Archive for long term, efficient retention.

There are several other examples of how IBM's different products and services can be used together to create a smart archive. However, customers should not lose sight of the fact that the whole point of IBM's Smart Archive is to do things differently in a more cost effective and more risk adverse manner than they are today.

Moving Forward Faster

Information Archive is an ideal starting place for a customer taking a new approach to archiving as it encompasses many of the facets of IBM's Smart Archive strategy. When a customer invests in an Information Archive, IBM Global Technology Services assist with installation and set up, accelerating the "time to archive." Often, archive projects get bogged down when customers realize they do not have the resources to properly connect data sources to the underlying storage system or spend the time to create retention policies. IBM helps Information Archive customers overcome these objectives and, for those organizations with existing archives, the services engagement can be extended to include data migration services.

IBM's services portfolio also includes managed services options for companies lacking the internal staff to run the archive environment. In some cases, customers may not be able to handle the explosive storage growth and associated data protection functions needed to properly manage an archive environment. Others may want to turn management of the entire archive infrastructure, including an Information Archive system and the applications which connect to it, over to IBM so they can focus limited resources on primary IT operations.

Enabling Access and Expiration

Saving the right—not all—information is a critical component of IBM's Smart Archive Strategy, as is expiring data when it is no longer needed for compliance, governance, legal, and business reference purposes. Determining what data to keep is a process handled by archive, content management, and other information management software applications. Information Archive helps customers expire content in a consistent manner through one its three protection levels assigned to a given collection. When customers begin to delete information, the storage savings exponentially increase because capacity is suddenly freed for new data.

Before data is expired, Information Archive facilitates secure information access by maintaining a clustered index of all the data stored within it. Applications and users can execute searches against the index, with results being displayed based on the user or application's security privileges. Those privileges may dictate if a user or application can view results across the entire Information Archive appliance or a specific collection. The index is always stored on disk (even if the data is moved to tape by the Information Archive), ensuring fast query response times even as archive storage capacity increases.

Heading for the Cloud

There is no bigger topic in IT right now than "cloud." While companies try to figure out their strategies in leveraging "public" or external cloud-based services, many are observing cloud-provider IT architectures using the insight to build internal or "private clouds." When examining what characteristics define a cloud architecture—flexibility, ease of provisioning, and real-time scalability—it is evident that IBM's Information Archive can be the cornerstone of a private archive cloud. Customers can connect most data sources to it, there are multiple "provisioning" options including setting up a new collection to address a new business requirement dictating the retention of information, and the appliance scales by adding capacity to the existing system or integrating another tier of and seamlessly migrating data to it.

When a customer is ready to leverage a public cloud, Information Archive can be deployed at a trusted third party service provider including IBM. A customer may do this to shift the day-to-day archive storage management to the service provider. In some cases, this transition may be triggered by shifting the software connected to the Information Archive from an on-premise deployment to a cloud or Software as a Service model. As an example, IBM's Content Collector software is already cloud-enabled which gives customers a choice in how they want to implement and consume archiving software. By having an architecture that supports cloud implementation characteristics, Information Archive provides customers with flexible deployment options.

The Bigger Truth

Archiving used to involve saving information for extended periods of time. Now, there are many more variables involved in archiving processes as companies have to deal with multiple information types, different access requirements, evolving retention policies, and lengthening retention periods. All of these impact archive storage needs and costs.

Companies could continue saving everything forever on the same storage media or buying separate storage systems designed to address individual archive requirements. However, these alternatives can get very expensive over the long term. With its "collections"-based architecture, IBM's Information Archive is flexible enough to address different requirements in a single system solving the stovepipe archive storage issue. And, whether you measure scalability by capacity (which Information Archive optimizes through information reduction features), resource management (which Information Archive has covered by supporting over 300 terabytes of disk and petabytes of tape), or number of objects (which Information Archive can handle up to three billion of in its first release), it is clear that customers can grow their archives with ease. Information Archive is also optimized to protect and secure archived information over the long term with capabilities such as Enhanced Tamper Protection.

As part of IBM's Smart Archive strategy, IBM's Information Archive provides customers with a variety of ways to solve evolving information retention challenges—an approach that has the potential to dramatically reduce capital and operating costs even as archive capacity explodes.

Enterprise Strategy Group | **Getting to the bigger truth.**

20 Asylum Street | Milford, MA 01757 | Tel:508.482.0188 Fax: 508.482.0218 | www.enterprisestrategygroup.com

TSW03051-USEN-00