IDC
Analyze the Future

WHITE PAPER

Integrating Data Protection and Archiving for Efficient Information Management

Sponsored by: Symantec

Steven Scully
Laura DuBois
Vivian Tero

July 2010

IDC OPINION

IT organizations are looking to modernize their data protection environments due to continued growth in the amount of data being created; accelerating use of server virtualization and shared storage; and increasing business, legal, and regulatory requirements. There is a lot of pressure on IT managers to meet these operational constraints and service-level objectives and, at the same time, ensure that the organization fulfills its regulatory and legal obligations and manages risks. IT managers have an opportunity to realize these objectives using the same technology investments. To do this effectively, corporations should have sound and harmonized information management and storage operations strategies. This approach should be supported by a robust and integrated storage infrastructure. Organizations looking to do this should consider the following:

- Align their information management program with the storage infrastructure operations, but don't boil the ocean. Focus on the high-risk content stores and applications, such as remote offices, network file shares, and SharePoint sites. Collaboration and planning between key stakeholders — legal, compliance, records management, IT storage, and IT security — will be useful in defining the appropriate retention, disposition, and archival policies. Collaboration between stakeholders will also drive priorities and help translate these policies into storage management protocols. Policies and protocols should also address the impact of virtualization and cloud computing in the IT environment.

- Consider an integrated tiered storage infrastructure using disk-based archiving and backup and deduplication as the foundation technologies. This tiered storage strategy would include a lean backup architecture, where policy-based archiving is used to move static, less frequently referenced data from primary storage and production environments into disk-based archives. The disk-based active archives will manage business records and fulfill regulatory and legal obligations to preserve data. Tape-based archives can be the third tier and will be used solely for long-term archiving.

- Deduplicate the data where it makes sense for their environment in order to optimize storage capacity, enhance the performance of the backup and archiving applications, and reduce the strain on network bandwidth resources during the backup and archiving processes.

- Utilize reporting and analytics tools and policy-based workflows to manage the health of backup, archiving, and deduplication processes.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P:508.872.8200 F:508.935.4015 www.idc.com

SITUATION OVERVIEW

IT organizations worldwide are dealing with dramatic growth of data. This information growth continues to be a key driver in the datacenter. At the same time, the increasing shift to virtual datacenter infrastructures is rapidly changing the IT environment. In response to these trends, IT organizations need to modernize their data protection and information management processes.

While the growth in enterprise disk storage capacity shipped slowed slightly in 2009, IDC is forecasting that disk capacity shipped will grow at over 50% through 2014. The growth in overall storage capacity comes with the complexity of managing and protecting all the data it contains, leading IT organizations to look at:

- Modernizing data protection processes to achieve a lean backup infrastructure
- Reducing the footprint of the data in a legally defensible manner by judiciously using deduplication and archiving technologies
- Adopting information retention and eDiscovery programs
- Monitoring and reporting on all aspects of data management

The pressure to protect the growing amount of data while adapting to the evolving datacenter infrastructure means that data protection and recovery solutions will continue to be a top priority for IT organizations in the years to come. IDC is forecasting that spending on data protection solutions will increase at a compound annual growth rate of 5.0% from 2009 to 2014. The spending on archiving solutions will grow even faster, at a rate of 11.5%, during the same time period as IT organizations implement more efficient ways of managing the growing amount of static and largely unstructured data created.

Modernize Data Protection

IDC research indicates that IT organizations are increasingly looking to modernize their data protection environments and solutions. A number of factors are driving this modernization, including dramatic growth in the amount of data being created and stored, server virtualization and shared storage, business and regulatory requirements for data retention, and mitigation of future legal liabilities arising from inconsistent data retention and disposition practices. Dealing with data growth has been a challenge for some time, while adapting to a virtual datacenter is a more recent challenge that has specific implications for data protection.

Virtual servers and desktops are increasingly being used to improve datacenter flexibility and scalability, but there are implications for organizations' data protection processes and architectures to ensure that every virtual server is protected and that the storage has the same flexibility and resiliency as the virtualized server environment. Some of these implications include:

- **Protecting all the physical and virtual servers.** Back when a server meant a physical box, it was easier to identify what needed protecting. A backup agent was simply installed on every physical server (box), and it was then incorporated into the backup process. Now that virtual servers can be created, moved, and shut down in minutes, making sure each server is protected has additional challenges.

- **Choosing a data protection approach.** Data protection vendors have been evolving with the virtual datacenter and now offer different options for protecting virtual infrastructures from backing up the host to each individual guest and from image-based backups to traditional file-based approaches. These various approaches provide different levels of granularity, recovery objectives, implementation costs, and the like. The challenge is to make sure each of these differences is clearly understood and that data protection expectations are being properly set.

- **Understanding the infrastructure impacts.** In addition to choosing an approach, storage administrators have to plan for the potential impact on server and network infrastructure, including an increase in the number of backup processes that will be consolidated and funneled through more limited I/O ports of each host.

- **Achieving application awareness and transaction consistency.** Especially in a virtual environment, the goal is application recovery, not just data protection. Virtual environments provide servers and the applications running on them a great deal of flexibility and mobility. This has elevated the discussion in many companies around recovery objectives, which increases the need for data protection and recovery solutions to be in better sync with the applications.

Consolidate and Centralize Data Protection

The major goal of these modernization efforts is to improve the efficiency of data protection and the performance of recovery and, at the same time, meet a corporation's legal and regulatory obligations. The best way to make this happen is to consolidate and centralize data protection applications across both physical and virtual servers.

Today, many organizations have multiple data protection approaches installed and in use. IDC research indicates that very large companies often have 10 or more solutions throughout the organization, which can come from mergers and acquisitions, changing data protection plans over the years (distributed moving to centralized and vice versa), and multiple remote or branch office data protection approaches. Some of the benefits that IT organizations can realize from a centralized approach include:

- **Reduced risk.** Having multiple and often siloed data protection approaches can increase the risk that servers are left unprotected. Organizations are finding local backups hard to manage and have difficulty staffing remote offices with experienced storage administrators. These organizations are increasingly moving away from local backups, leaving most local offices without any backup protection. A centralized backup strategy can address this issue.

- **Improved security.** Developments in the information security threat landscape, combined with upcoming data privacy regulations, are also driving the need for more centralized backups. For example, tapes containing sensitive corporate information are less likely to go missing from a local backup either at the local site or when the tapes are in transit to the central datacenter. Sensitive customer data is easier to protect if consolidated into a single process.

- **Simplified management.** Managing multiple data protection tools is complex and time consuming. A consolidated approach can reduce or eliminate the need to keep multiple applications up to date potentially across geographically dispersed locations, provide ongoing training for storage administrators who often are specialized in one specific application, manage multiple backup agents, and work with inconsistent alerts and reports.

- **Enhanced performance.** As part of data protection centralization and modernization, many IT organizations are increasingly using disk-based solutions with or instead of tape solutions to better address recovery time objectives and enhance overall performance.

Data Deduplication

Data deduplication has become an important storage technology in the past few years as IT organizations look to improve storage utilization and reduce costs. Storage solutions, either based on deduplication or with deduplication as a feature, are now available across the entire spectrum of storage offerings from many vendors, large and small.

Data deduplication gained much of its market attention around backup data and for good reason. Because a backup process typically is copying the same files again and again, it makes sense not to store another copy of a file if it has already been done once. Backup data remains a key opportunity for deduplication technologies, and almost every backup and recovery solution — from backup software to virtual tape libraries to disk-based backup systems — currently includes some form of data deduplication.

Data deduplication works by looking for repeated patterns in various chunks of data and eliminating the duplicates. Typically, an algorithm is used to generate a hash for each chunk of data, and if it matches a hash that has already been stored, the newer chunk of data is replaced by a pointer to the existing chunk already stored on the system. Three types of chunks are typically used for deduplication — file level (also called single instance), block level, and byte level — each with its own benefits and trade-offs.

Another technical aspect of deduplication is when it is accomplished by the solution, with inline and postprocessing being the most typical options. With inline deduplication, duplicate chunks are identified and removed before they are written to the back-end disk drives of the system. This requires more computing power and can impact storage performance, but it doesn't require additional space and doesn't perform unnecessary writes of data that already exists. Alternately, data can be written to disk, and the deduplication is then accomplished by a postprocess typically executed as part of a scheduled operation. Postprocessing solutions require less computing power, reduce the potential impact on storage performance, and can be scheduled at times that are convenient to the operation of the datacenter. However, postprocessing does require additional storage capacity to hold all the data before the duplicates can be removed and will execute additional reads and writes of the data.

A more important aspect of data deduplication for the data protection process is where the deduplication takes place. Deduplication techniques have evolved to the point where they can be accomplished at the source, the media server, or the target (often an appliance), or in some combination of the three.

- **Source deduplication.** Because all deduplication requires some processing power to compute and compare, source deduplication (sometimes called client-side deduplication) can be good when the clients have available processing power or when bandwidth is limited or expensive.

- **Media server deduplication.** Doing the deduplication at the media server can be good for more global deduplication because it offers a single point of contact that can deduplicate across multiple backup streams coming through the server.

- **Target deduplication.** Deduplicating at the target is effective when adequate bandwidth is available and the desire is to get backup jobs off the clients and media server as quickly as possible.

Data Archiving

Archiving is one of those areas that is broadly interpreted by IT organizations. When talking to organizations of all sizes, IDC often finds that many indicate that they already do archiving. Once they provide a better explanation of their archiving activities, archiving for many organizations can be as simple as saving a copy of each month's backup tapes and shipping them to an offsite vault for an extended period of time.

Backup and archiving are distinct but complementary operations that address storage operations and information risk management objectives. Organizations should not use them interchangeably.

Historically, organizations have utilized backup operations to quickly restore large volumes of data that are lost when hardware fails or during natural disasters. The combination of aggressive growth in data volumes, rise in the number of remote offices, adoption of mobile devices and endpoints, and deployment of new technologies such as virtualization and cloud computing in the corporate infrastructure is posing data management challenges in a typical organization's backup operations. These data management and operational challenges are manifested in longer backup windows, incomplete backup operations, and corrupted and incomplete backup data sets.

Backup and archive are two separate processes with different goals and should be considered as such. Properly implemented, backup and archive processes complement each other and enhance the overall data management process. Backup applications provide point-in-time copies of a defined set of data to tape, disk, or optical media and are used to recover all or part of the data set if needed because of logical or physical error or site disaster. Backup applications are designed to provide regularly scheduled copies to bring data back online in the event of an outage.

Archiving provides an automated and efficient way of storing, indexing, and retrieving data. It creates policy-based point-in-time copies of data sets. Individual files are indexed and individual file paths are also captured, stored, and managed for search, retrieval, and chain-of-custody documentation purposes. Archival policies are typically defined by the organization's regulatory, legal, and business obligations, as well as application performance and storage operational requirements such as optimizing backup windows. Archiving enables the organization to delete the original copies of the data from the primary storage and production environment while still allowing authorized users to search for and retrieve individual files, folders, or entire instances of the data set from the archival environments.

This single-instance storage feature allows the organization to free up storage space in the production and primary storage environment and improve application performance. The organization may write copies of the archived data to disk for nearline storage or to tape or optical storage. Active archiving is the strategy for moving archival data into nearline storage, thus facilitating close to real-time search and retrieval. Active archiving is typically used to manage infrequently referenced business records and preservation-intense information. The decline in the prices of high-density disks has made disk-based backup a viable option for more frequently accessed information, allowing tape storage to be used as a longer-term archive. Under this tiered storage infrastructure, tape archives are deemed the least accessible. Understanding this distinction is important during litigation.

Backup with Archiving

Archiving allows for policy-based movement of large volumes of static data from primary to secondary storage. Archiving does not maintain copies of data that has changed in between backups or retain copies of data deleted by users after the backup is replaced with a newer version. Backup and archiving are thus complementary operations, best done with separate but integrated tools to maximize their efficiency.

Implementing data archiving as part of an integrated data protection strategy can make backup better by providing a number of benefits, including:

- **Faster backups and shorter backup windows.** When older static data is archived off primary storage and onto an archive system, the backups of newer production data will go much faster. It's not uncommon to have 70% of data eligible for archive, which, if archived, would allow the backup of production data alone to run 70% faster. In addition, the backup data sets are smaller. Moving fixed content or less frequently referenced data off primary storage and production environments enables the organization to reduce the need to conduct full backups, which can be a time-consuming and compute resource–intensive process. Smaller backups are also less taxing on the network and allow for faster backups and shorter backup windows. Consequently, harmonizing backup with the archiving would enable organizations to reduce their backup windows in primary storage and production environments to 90-, 60-, 30-, or even 7-day cycles. This approach allows organizations to keep their backup infrastructure lean by letting the archive manage the data for the longer term. Moving the data into the archives also enables organizations to fulfill retention and eDiscovery obligations.

- **Better recovery.** In the unlikely event of the need to recover from a backup, the recovery will be faster because the backups are smaller and only the most current production data is restored. In the meantime, the disk-based archives will make the nonproduction data searchable and available for the user. Conversely, when nonproduction data is scattered throughout the backup data set, it all must be restored instead of just the current data needed for production.

- **Optimized storage.** Organizations are able to optimize the storage infrastructure and mitigate the need to overprovision primary storage capacity by using the following strategies in concert: automate data retention and deletion programs that define the disposition of business records and information in the primary storage, production, and archival environments; apply deduplication and compression technologies to reduce the data footprint; move less frequently referenced and static data into the archive; and employ a tiered storage strategy that allows for multiple disk-based storage tiers within the archive and an option to use tape as the third tier for cost-effective long-term retention. It is worthwhile to note that for remote offices, deduplication at the source would be less taxing on network bandwidth resources, making remote backup and archiving operations more efficient. Within archival environments, policy-based retention and disposition protocols that include processes for legal holds will obviate the need to keep separate copies of preservation-intense data, thus keeping the archives at a more manageable level. Also, a consistently enforced tiered storage strategy may provide some level of protection against unreasonable requests to produce data from inaccessible storage media *(Rule 26[b][2][B] of the United States Federal Rules of Civil Procedure [FRCP] for Electronic Discovery).*

- **Enhanced security.** Keeping multiple versions of full backup copies increases the risks of data loss. Data protection regulations not only create obligations to ensure that data is not kept for far longer than mandated but also oblige the organization to take precautions that specific data classes are not destroyed before their mandated expiry date. Utilizing archiving in conjunction with backup centralizes the disposition and security of static data. Security postures are assessed along the dimensions of confidentiality, integrity, and availability (CIA). Archiving in conjunction with backup addresses these security requirements. Backup and archiving technologies are designed to ensure the availability of the data in the event of an outage or a data loss. But what about confidentiality and integrity? Policy-based archiving includes a granular security model so that authorized users are able to search, retrieve, and act on the data based on their roles and entitlements profile. In addition, archiving allows the organization to capture a full range of metadata information, including the file path, custom tags, and other referential relationships, thus documenting the integrity and provenance of the data.

The eDiscovery-Ready Storage Infrastructure

The information management obligations implied by the 2006 amendments to the FRCP for eDiscovery pose potential eDiscovery liabilities for organizations that do not have solid information management programs. Organizations that use archiving and tape backup operations interchangeably have been compelled by the U.S. courts to search and produce data in backup tapes in many instances, producing a significant volume of redundant data. When one takes into account the average costs of attorney reviews plus the infrastructure and manpower expenses incurred to search, collect, process, and produce the data, the total costs for eDiscovery alone can become prohibitively staggering. The lessons from high-profile litigations such as *Zubulake v. UBS Warburg* and *Broadcom v. Qualcomm* demonstrate what happens when organizations do not have sound information management programs and use archiving and backup operations interchangeably.

Industry research concludes that technology infrastructure costs account for 10% to 20% of the total eDiscovery costs per matter; while eDiscovery review and production services make up 80% to 90% of the total. However, an organization's information management program and storage operations have the ability to impact critical value levers that can significantly lower eDiscovery review and production costs.

Using a combination of archiving, deduplication, and backup technologies to programmatically enforce retention and disposition policies reduces the data footprint in a legally defensible fashion. The strategy cancels the need for the "keep everything" approach and also allows for more targeted search and collection, thus reducing the volume of potentially responsive data that is sent for attorney review. An organization that uses these technologies as part of its tiered storage infrastructure may also be better able to demonstrate compliance with the Rule 37(e) safe harbor requirements. It may also allow the organization to argue against a requesting party's demands to produce data from the "least accessible" backup tapes. With the right process and technology, the organization may avoid unnecessary tape restoration and data migration costs.

Disk-based archiving would enable an organization to centrally manage the legal hold process and search for and retrieve information quickly, allowing legal counsel to conduct early case assessment. By gaining early visibility into the case facts and historical information, legal counsel is able to more accurately estimate the true cost of the litigation (including eDiscovery costs), identify and assess the potential risks, and evaluate options to determine the appropriate legal response to a dispute or complaint. Disk-based archiving also enables the organization to avoid the unnecessary costs associated with tape eDiscovery, restoration, and migration.

Deduplication with Archiving

The aggressive growth in data volumes is creating management challenges in archival environments, in some cases slowing down the performance of the archives and taxing the network bandwidths. Data volume growth in the archives is slowing down the ingest rate as well as the creation of indices, which are critical to facilitating fast and targeted search and retrieval.

Deduplication can reduce data volumes by up to 90% of the archive and backup data across both remote offices and centralized configurations. Using deduplication in conjunction with archiving provides the following operational benefits:

- When done at the source, deduplication can provide significant reduction in primary storage capacity. For remote offices, deduplication at the source also reduces the strain on the network bandwidth by reducing the volume of data for backup and archiving. The reduction in data footprint will result in shorter backup windows and faster ingestion of data into the archives.

- When done at the target (such as the archival environment), deduplication also provides storage capacity optimization benefits. But more importantly, deduplication could improve the performance of the archiving application, including faster search and retrieval.

Data Management Reporting

As previously mentioned, backup and archiving are different processes best done with separate tools to maximize their efficiency. Further improvements can be gained by using solutions to monitor and manage the backup and archiving processes, servers, storage devices, and information to be protected. Separate solutions exist for managing backup environments and archive environments, but increasingly, integrated management solutions are available to allow IT organizations to monitor and manage their complete data protection environment.

Including data management monitoring and reporting as part of an integrated data protection strategy can improve operational efficiencies while providing a number of benefits, including:

- **Improved efficiency.** It's difficult to improve on processes without monitoring and measuring them. The use of management tools can help increase the efficiency of the entire data protection process. Most solutions offer various types of reports that can increase data protection efficiency, including backup success rates and storage device utilization, while also providing management tools for historical trends and predictive forecasting.
- **Reduced risk.** The use of management solutions can also help minimize risk in the datacenter and ensure and then prove backup and archive process completeness and compliance with corporate governance and service-level agreements (SLAs). Many solutions will help IT organizations identify unprotected data and applications and analyze the risk of meeting recovery time and recovery point objectives.
- **Increased business awareness.** Centralized management goes hand in hand with a virtualized (and therefore centralized) infrastructure and applies across heterogeneous applications and environments. Reporting will provide various stakeholders — administrators, line-of-business owners, application owners, CXOs, etc. — with insight into the value of data protection and associated costs. Most reports can be easily tailored to the needs of the specific audience.

INTEGRATED DATA PROTECTION AND ARCHIVING FROM SYMANTEC

As a leader in the storage industry, Symantec has developed leading data protection, archiving, and management applications for many years. It has also been at the forefront of integrating data protection and archiving applications into a more cohesive solution for organizations and providing for end-to-end management of the processes and information. In particular, the following Symantec products relate to the trends and topics discussed in this paper.

- **Symantec NetBackup.** NetBackup is an industry-leading data protection platform for physical and virtual datacenters. Included in the NetBackup platform is NetBackup PureDisk, the deduplication engine for NetBackup, enabling efficient, storage-optimized data protection for datacenters, remote offices, and virtual environments. Together, these products offer:

    - ❑ Physical and virtual server support, including capabilities to protect and restore a complete server, a virtual server image, or individual files from a virtual machine. (This is all performed with a single agent on the server and fully integrated with the hypervisors.)

    - ❑ Data deduplication everywhere and wherever it makes the most sense for the application and the environment (NetBackup can support deduplication at the client, in the media server, or in conjunction with a hardware solution.)

- **Symantec Enterprise Vault.** Enterprise Vault is an industry-leading archive solution for email and content. It enables users to store, manage, and discover unstructured information across the organization. Enterprise Vault reduces the size of primary storage and applications by moving emails and files into an online archive and applying deduplication and compression. It also creates a searchable index of all archived information. Users can archive across multiple content sources and applications into a central repository and leverage a tiered storage strategy for retention and deletion. Key features of Enterprise Vault include:

    - ❑ Provides policy-based archiving that automates the migration, storage, and retention of unstructured information based on policies and is fully integrated with Symantec data protection solutions like NetBackup

    - ❑ Supports eDiscovery and searches with legal holds (Searches can be conducted and results can be reviewed based on the organization of data retained within the archive. Legal holds will suspend the deletion of archived data in response to internal investigation, litigation, or regulatory request.)

- **Symantec OpsCenter Analytics.** OpsCenter Analytics helps IT organizations enhance backup and archive operations by providing enhanced visibility and control, verifying service-level compliance, and aligning backup and archiving with the needs of the business. OpsCenter Analytics provides multiproduct and platform support, business reporting, and complete customization. Key features include:

    - ❑ Central reporting all across NetBackup, Enterprise Vault, and PureDisk operations to better understand the data and processes; improve SLA management; and monitor for compliance

    - ❑ Long-term, configurable data retention for trending and analysis to better predict disk and tape consumption for backup and archiving jobs across multiple locations based on historical backup and archive job information

CHALLENGES/OPPORTUNITIES

The integration of data protection and archiving applications provides organizations with a number of advantages and opportunities to improve data protection efficiency and performance. However, firms still must overcome organizational hurdles.

First, backup and recovery operations and infrastructure can be long established and ingrained within many IT organizations. Getting these organizations to make changes to both processes and systems can be a challenge, despite the benefits of modernization and consolidation. However, organizations will need to start moving away from siloed storage operations and take on a more unified posture. Symantec has an opportunity to work with its professional services network and its channel partners to develop a range of services to help organizations make that transition.

Second, a significant but declining number of organizations continue to view archiving and backup as interchangeable processes. The need to manage data volume growth and improve application performance by optimizing storage and the pressure to overcome eDiscovery challenges are issues that Symantec can help firms address. Symantec's integrated data protection and archiving strategy can provide material operational and compliance benefits.

Lastly, limited budgets remain a challenge for all firms, despite the slow economic recovery. Some of the backup and archiving solutions have adequate monitoring and reporting capabilities built in to provide "good enough" functionality. "Good enough" functionality may stymie adoption of deeper monitoring, reporting, and analytics tools. Other customers prioritize their limited storage budgets on additional capacity and applications ahead of management tools. As a result, suppliers need to demonstrate to IT organizations that storage management can provide the central orchestration needed for improving the efficiency and effectiveness of their data protection and archiving processes.

CONCLUSION

Symantec's data protection, archiving, and data management products are well positioned to address current market needs in today's physical and virtual infrastructures. Symantec's ability to combine these products into an integrated data management solution can further enhance the value to IT organizations. Customers can implement as many products as their requirements and budgets will allow with the knowledge that they can add remaining products as needed.

IT organizations planning for or in the midst of datacenter virtualization should evaluate all of the data protection aspects at play. In addition, they should consider centralized backup and data archiving as integral to a modernized disaster recovery strategy. When integrated around a data management and reporting solution, centralized backup and data archiving can significantly enhance overall storage efficiency, reduce risk, improve SLAs, and lower costs.

Copyright Notice