

LEVERAGING EMC SOURCEONE AND EMC DATA DOMAIN FOR ENTERPRISE ARCHIVING

AUGUST 2011



Archiving is a fundamental storage process for controlling storage costs and managing long-term data for compliance and value. However, we find that many companies are either reluctant to adopt archiving or they under-utilize the archiving they have. Why is archiving so important and what are the pressures that mitigate against it?

Archiving primarily deals with unstructured data, which usually accounts for a majority of data on primary storage. Exchange, Domino, SharePoint, file systems: all of these applications generate and store massive amounts of data, which impacts networks, servers and storage. Unfortunately for the enterprise, “unstructured” all too often equals “unmanaged.” Much of the unstructured data on enterprise storage is inactive and storing it may impact application server performance. Yet some of this inactive data is subject to retention requirements for eDiscovery, compliance, risk management and more. This data needs to be stored long-term in an easily retrievable format out of primary or tier-1 storage.

Deploying archive storage addresses many of these challenges by providing long-term, protected and highly reliable data storage. However, organizations have historically not implemented archive processes; instead many simply extend retention periods with existing backup processes and call it an archive. Part of the reason many haven’t implemented archiving processes is a reluctance to add complexity to an already complicated storage environment. Although archiving actually relieves complexity, in these days of thin IT headcount and limited budgets we can understand IT being reluctant to purchase specialized archive storage technology.

However, there is an appealing middle ground that combines high performance backup storage for fast ingest and disaster recovery with long-term, cost-effective archiving storage. This solution offers IT very attractive single-platform economies by cutting the amount of data movement between disparate systems, dispensing with separate upgrade schedules that upset integration, and partnering with a single expert vendor. A single integrated storage system enables backup and archive processes to meet their common needs for performance, scalability, management simplicity, security, and storage efficiency. The system meets each process’s individual requirements: backup for high speed ingest to meet short backup windows, continuous fault detection and self healing and fast, reliable recovery; and archiving for long-term retention and reliable restores.

To this end, EMC provides integration between EMC SourceOne and EMC Data Domain for peak optimization of long-term data retention. EMC SourceOne works across critical applications to archive data for long-term data integrity and value, while EMC Data Domain deduplication storage systems protect backup and archive data with industry leading performance and scale. This paper will introduce EMC’s strategy founded on the best-in-class integration between EMC SourceOne and EMC Data Domain.

The Challenge of Unrestrained Data Growth

Due to continuous data growth, huge amounts of unmanaged data are consuming more capacity on enterprise storage systems. This results in large storage purchases, vast rack space requirements and high energy costs, high risk from being unable to find information, slow application performance, and keeping multiples copies of inactive data thanks to the traditional backup process.

Much of this data is unstructured and resistant to data management. At the same time the costs of locating specific information is growing. Litigation is a prime source of eDiscovery costs with enterprises commonly spending multiple millions of dollars per year in litigation costs – much of it on eDiscovery followed by sanctions and lost judgments for executing eDiscovery poorly. Why the high cost? The primary eDiscovery data type is email, which is notoriously difficult to search and retrieve across due to large email servers, backups and thousands of PC's and laptops containing client-side email copies.

eDiscovery would not be a big problem if more companies employed archives instead of treating backup as an archive store, but the reality is that many companies do think of their backup as long-term data retention. Backup is designed for short-term operational and disaster recovery, not for storing 5 to 10 years worth of disparate data.

Archiving software eliminates the issues associated with searching through files on backup storage. Yet, the storage demands for large archives are intense, and simply throwing traditional storage at immense data stores doesn't address the core problem. The solution is information governance software built for managing data according to business needs, combined with deduplication storage systems built for backup and archive.

INFORMATION GOVERNANCE AND THE 5 W'S

Remember learning the 5 W's in school? They apply today to information governance. Ask yourself these questions:

Who has access to what information? How do administrators apply the correct permissions to the correct people? This is a big issue with networked storage repositories including long-term archives.

What information does the company store? IT knows the answer in terms of backup targets and application storage. But they lack visibility into the content of files and therefore struggle to manage information for value, retention and recovery.

When is the information acted upon? Short of keeping everything (a terrifically bad plan) data must be acted upon lest it live indefinitely on ever-growing disk or tape. IT should be able to assign priority and policies based on the age and value of data.

Where is the information stored? "Where" can be surprisingly difficult to know outside of the data center. Storage devices proliferate throughout the enterprise ranging from SAN to NAS to DAS to computer drives to thumb drives. Information requests may require far-flung search and retrieval.

Why is the information stored? Unending storage purchases are unsustainable for capital and operating budgets. Justify long-term data storage by its relation to eDiscovery, regulatory and industry compliance, internal governance, or business value. Within the "why" framework, IT makes the hard decisions to retain, move or delete.

INFORMATION GOVERNANCE TO THE RESCUE

Information governance is the concept of retaining data long-term for business value, compliance and eDiscovery. When done well, information governance accomplishes three critical aspects of managing information in the business. These include 1) controlling storage infrastructure costs, 2) ensuring data protection and integrity, and 3) managing data for eDiscovery and compliance. Effectively managing data will deeply impact the storage environment and the processes that depend on it.

- **Critical Aspect #1: Control storage infrastructure costs.** These costs hit a wide variety of areas including hardware, software, connectivity, data center real estate, energy costs, management overhead and more. Investing money and time into building an efficient infrastructure may require an upfront investment, but will payback quickly by greatly easing management burdens and dramatically lowering ongoing costs. For example, a well-managed storage infrastructure will provide high capacity in a small footprint. Deduplication and policy-driven data movement will lower CAPEX and OPEX for purchasing, infrastructure, energy usage and storage overhead management.
- **Critical Aspect #2: Cost-effectively ensure data protection and integrity.** This category covers data recoverability for applications and users and securing data. Backup is the first line of defense here but note the “cost-effectively” part of this critical aspect. Traditional tape-based backup processes overrun backup windows and threaten recovery and service level agreements. Uncontrolled backup – and poorly managed replication – are equally bad over time as they consume valuable capacity and bandwidth with duplicate content. And searching backups for business processes can be enormously time-consuming. In addition, data integrity measures are required to preserve data for long-term retention including verifying data recoverability at write as well as continuous fault detection and self-healing.
- **Critical Aspect #3: Efficiently manage information for eDiscovery and Governance, Risk and Compliance (GRC).** eDiscovery and GRC both require effective search and retrieval methods, which in turn require a well-managed storage infrastructure. (It’s no accident that the first stage on the Electronic Discovery Reference Model (EDRM) is “Information Management.”) Managing information for these business processes requires: appropriately retaining data for long-term retention and enabling efficient search and retrieval on a granular level. For example, if a company retains data long-term but has not managed it well, then they will fail eDiscovery deadlines when searching unwieldy data stores. Maintain compliance and litigation readiness with centralized archives for unstructured data.

EMC’s Integrated Strategy for Information Governance

EMC provides solid information governance that meets the above three critical aspects. The solution is built from EMC SourceOne and EMC Data Domain deduplication storage systems. EMC SourceOne and EMC Data Domain systems operate in multi-vendor environments but when integrated, offer optimal efficiency for information governance. Let’s start by understanding EMC SourceOne and EMC Data Domain systems as individual technologies and then discuss the benefits of integrating the two systems for effective information governance.

EMC SourceOne

Organizations need to lower TCO of the expensive production environment while also providing data integrity and retention controls. EMC SourceOne archiving makes storage management

efficient, reduces storage loads on production servers, and slashes storage costs by providing cost-efficient tiered archive storage. EMC SourceOne works on SharePoint, email and file systems by archiving to a centralized storage repository with a single administrative console.

EMC SourceOne moves inactive data from production servers to archives and deduplicates data with single instancing using a unique object ID. The data is indexed for fast search, management and retrieval. Additional eDiscovery capabilities are available with EMC SourceOne Discovery Manager and EMC SourceOne eDiscovery - Kazeon.

The content is put into archive folders based on organizational policies. These policies are flexible enough to target any specific set of data and to place it in the appropriate archive folders. Archived content remains available to users and searchers via a simple web interface and additionally through stubs or placeholders in the source application. Incoming objects are assigned retention periods. EMC SourceOne's modular architecture offers three areas of information governance and control: email management, SharePoint and file systems management.

- EMC SourceOne Email Management.** EMC SourceOne removes messages and attachments for storage reduction on Exchange and Domino servers. EMC SourceOne retains pointers to archived messages and attachments so the action is transparent to end-users. Users may also employ EMC SourceOne Discovery Manager to manage archived email for eDiscovery and compliance actions. EMC SourceOne's benefits also extend to migration projects. For example, Exchange 2003 and 2007 users are looking to upgrade to Exchange 2010. Over time, they have also collected large volumes of email and PST files that will have a serious impact on migration. Administrators can employ EMC SourceOne to archive these messages, attachments and calendar entries and remove them from the Exchange server. This action alone dramatically improves the migration process without threatening data.

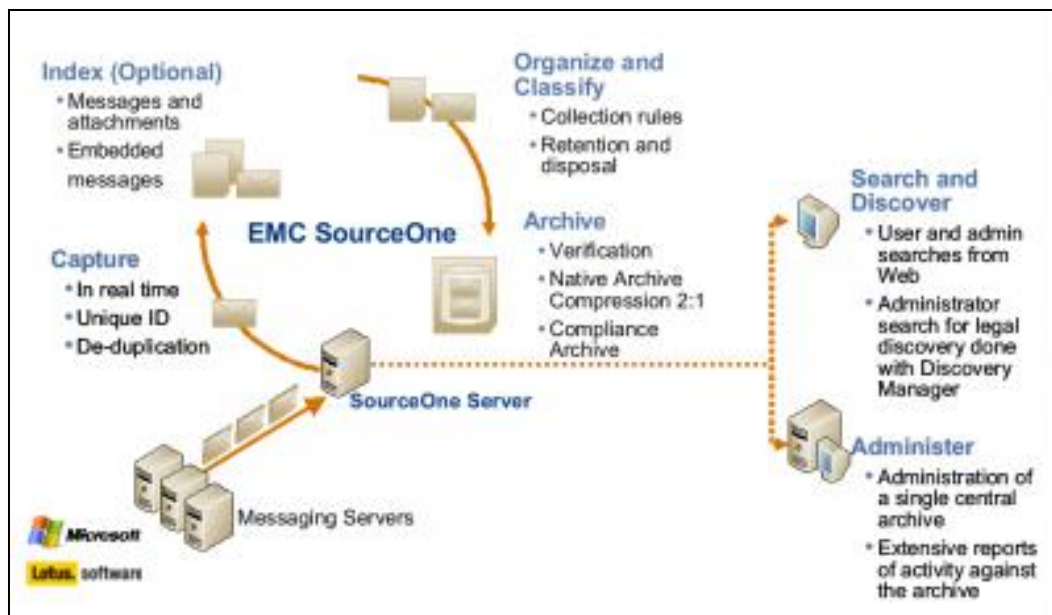


Fig. 1: EMC SourceOne Email Management (EMC)

- EMC SourceOne for Microsoft SharePoint.** The SharePoint module moves inactive content for archiving from SharePoint to EMC SourceOne and leveraging its classification and retention policies. Both capabilities are hugely important in light of SharePoint's fast data growth and

content management. Its native infrastructure stores content BLOBS in the SQL database, which quickly overwhelms performance. By externalizing active storage and archiving inactive files, administrators will significantly improve SharePoint performance and longevity.

- **EMC SourceOne for File Systems.** The File Systems module provides policy-based file archiving to the EMC SourceOne archive where it indexes file metadata for efficient search and retrieval. EMC SourceOne for File Systems is content-aware, which ensures full-text indexing of the content located within the files and also computes a unique hash value for each file so that duplicates are filtered out. EMC SourceOne can flexibly target files using a variety of options including file name/path, size, type, and age. Users can also apply actions to file servers, file server shares, and folders in a share.

EMC Data Domain Deduplication Storage Systems

EMC Data Domain deduplication storage systems are high-speed, scalable appliances built for efficient backup and recovery. For example, most organizations have data retention policies that require keeping backups online for 30 to 90 days. Data Domain systems support this requirement with high performance to meet backup windows, in-line deduplication for longer onsite retention and network-efficient replication to an off-site disaster recovery location.

Data Domain systems deduplicate data inline and average 10-30x data reduction for backup workloads, allowing for longer onsite retention of backups and network-efficient replication. Because Data Domain systems reduce data so effectively, companies can keep data onsite longer without sacrificing data center space or spending a lot on new storage capacity and energy costs. Companies can cost-effectively keep backup data immediately available for fast recovery over the LAN or WAN. Note the “WAN” part: a single Data Domain system can support replicated data from over 250 remote sites. Since only unique data is transferred over the network, datasets effectively shrink up to 99 percent. This happens without investing in a large-scale WAN acceleration initiative.

Data Domain systems work with all leading enterprise backup software and archiving applications and do not require any custom backup application. IT can easily integrate Data Domain systems into the existing storage infrastructure, always a big plus. The Data Domain Data Invulnerability Architecture provides continuous recovery verification and additional data protection and comes configured with dual disk parity RAID 6. Additional system status and collection reporting keep IT completely up-to-date.

Data Domain systems include appliances for remote offices and primary data centers including the Data Domain Global Deduplication Array and Data Domain Archiver. These systems can support enterprise data centers with up to a half a PB of data. Aggregate throughput across the product family ranges from 490GB an hour to 26.3TB an hour.

The Data Domain Operating System provides data integrity capabilities with the Data Domain Data Invulnerability Architecture. This provides end-to-end data verification, fault avoidance and containment, continuous fault detection and healing, and file system recoverability. It also includes dual disk parity RAID (6), unique write verification, multiple access methods and centralized administration. EMC Data Domain Retention Lock software further ensures internal IT governance over stored data by providing defensible file deletion by disallowing rewriting and erasure on individual or global files, and also provides efficient ways to administer and customize policy.

EMC Data Domain Encryption software is also available as an option and encrypts data inline during the deduplication process. Users may also use EMC Data Domain Replicator software for network-efficient asynchronous replication of deduplicated data over the WAN.

DRILLING DOWN: DATA DOMAIN ARCHIVER

Data Domain Archiver enables long-term retention of backup and archive data. As we discussed earlier in this paper, the reality of so-called “archiving” is that companies are simply storing backups on tape and calling them archives. This makes it extremely awkward to prove specific data compliance or to search and recover for eDiscovery. The best solution is to use dedicated archiving software like EMC SourceOne and we strongly suggest it. But EMC must deal with the reality of customer storage choices, and many IT organizations will continue to use backup as long-term archival.

EMC provides archiving capabilities on all Data Domain systems in order to integrate archival and backup storage in the same deduplication storage system. Since Data Domain Archiver was built for both backup and archive, users can benefit from the performance advantages of a Data Domain system for ingesting backup data as well as the long-term retention architecture required for archives.

DD Archiver is purpose-built to provide long-term data retention, both for archives *and* backups. This lets customers have the best of both worlds: 1) Data Domain Archiver ingests and deduplicates data inline directly from EMC SourceOne and other leading archive software, and 2) also ingests backup data directly from all leading backup applications for long-term retention.

Data Domain Archiver uses distinct tiers to enable long-term retention of backup and archive. The active tier is similar to a standard Data Domain system and is built for short-term retention of backup data. The archive tier is a massive secondary tier that uses the same Data Domain controller, management and namespace. The user defines the data movement between the two tiers using the familiar Data Domain Enterprise Manager.

DD Archiver achieves up to 9.8TB per hour of throughput and capacity scales up to 768TB raw. Data Domain Archiver enables IT to send multi-vendor backup and archive data to a single system. All data initially lands on the active tier and based on user-set policies, DD Archiver will move the deduplicated data into the high capacity archive tier. The tiered architecture enables Data Domain users to employ an active tier for short-term backups and to use a massive archive tier for long-term data retention. The archive tier consists of multiple archive units and scales by adding storage shelves to easily increase the number of archive units.

Integrating EMC SourceOne and EMC Data Domain

EMC Data Domain systems present great advantages for disk-based backup and long-term data retention, while EMC SourceOne enables a powerful centralized archive repository. Putting them together leverages the advantages of both into a single integrated process. Combined usage gives users effective information governance as well as duplication storage with industry leading performance, scale and reliability. With this solution, Data Domain systems serve as EMC SourceOne’s centralized repository for archived email, SharePoint and files. The Data Domain system provides fast ingest, deduplication, encryption, and dramatic scalability to EMC SourceOne archives. EMC SourceOne leverages Data Domain value by allowing users to quickly and defensibly retrieve EMC SourceOne archives for eDiscovery, investigation, audits and compliance.

Leveraging the two products results in a complete long-term retention solution that benefits from deduplication and archive software efficiencies. When customers add EMC SourceOne archiving to a Data Domain system they achieve a new level of information governance.

Integration Benefits

Let's look at three primary benefits in terms of the requirements we set out earlier: controlling the storage infrastructure, providing data protection and readying data for discovery.

- **BENEFIT: CONTROL STORAGE INFRASTRUCTURE COSTS**

Using EMC Data Domain systems with EMC SourceOne lets users leverage a single deduplication storage system for both backup and archives. One of the primary considerations for archiving data is the cost and longevity of the storage system behind it, especially given the large volumes of email with their many copies and large attachments. By using Data Domain systems as EMC SourceOne's archive and eDiscovery repository, users leverage the archiving and eDiscovery capabilities of EMC SourceOne along with the deduplication, performance, scale, centralized management and long-term retention capabilities of Data Domain systems.

The combined solution also lowers the cost of keeping inactive files on primary servers. Instead of impacting server performance with bloated storage, use EMC SourceOne to apply policy-driven data categorization for deletion or retention and store the resulting archives on Data Domain systems. This greatly improves server performance by reducing stored files and by cutting down dramatically on backup operations.

- **BENEFIT: ENSURE DATA PROTECTION AND INTEGRITY**

EMC SourceOne exists to meet governance rules, preserve business value, and enable eDiscovery. These strong and compelling drivers depend on data availability and integrity. Data Domain systems protect both backup and archive data with the Data Domain Data Invulnerability Architecture, Data Domain Retention Lock, Data Domain Encryption and Data Domain Replicator software options.

Together this solution achieves the required level of reliability and security for backups and archives. Backup requires fast throughput, accurate verification and immediate availability for restores. The Data Domain Data Invulnerability Architecture operates to ensure this level of backup integrity and availability. Archive storage needs an additional set of security features beyond secure hardware storage. Although backups are usually restored by volume, archives are more frequently restored on a granular level. They must be highly searchable and available to multiple business processes including eDiscovery and governance. Data actions such as deletions must also be defensible, as is the integrity of the originally archived file. DD Retention Lock and EMC SourceOne's own disposition policies operate together to provide this level of security and defensibility.

- **BENEFIT: MANAGE DATA FOR EDISCOVERY AND COMPLIANCE CAPABILITIES.**

Rather than implement archiving process, today most companies use their backups for long-term retention. This is less than ideal but it is common, so EMC provides for this aspect of the real world with Data Domain systems. However, companies are under increasing pressure to search and restore granular data in response to business processes such as eDiscovery, investigations and audits. By adding EMC SourceOne archives and eDiscovery features to the

mix, Data Domain systems become a high value eDiscovery repository as well as optimized backup storage.

Having centrally managed backup and archives on the same system optimizes both Data Domain systems and EMC SourceOne's respective benefits. Companies decrease the risk to chain of custody from data movement since integrating EMC SourceOne and Data Domain systems minimizes data movement once the backup or archive data enters the Data Domain system.

Another benefit of using EMC SourceOne with Data Domain systems is extremely efficient retention policies. Data retention impacts a variety of business and storage management needs including legal retention, regulatory compliance, internal governance, and efficient capacity management. For example, EMC SourceOne can implement legal holds with Discovery Manager, which the Data Domain system enforces with Data Domain Retention Lock.

TYPICAL INTEGRATION SCENARIO

A company has purchased EMC SourceOne Email Management to archive their Exchange data **and** plans to purchase the SharePoint and File System modules over the next 12 months. An existing compliance agreement with Finance also requires them to store backup data containing SAP financials for 7 years. By combining EMC SourceOne with EMC Data Domain systems, IT can accomplish all three objectives:

1. *Send archived data from EMC SourceOne directly to EMC Data Domain.* Data Domain Archiver archive tier provides optimized data retention, scalability and management for archive data. As IT adds additional EMC SourceOne modules or other leading archive software, Data Domain Archiver will easily scale to store the archival data.
2. *Use a deduplicating storage system for fast and economical disaster recovery.* Data Domain systems deduplicate and store backup data for fast and efficient disaster recovery. The organization plans to phase out backup tape over the next few months and Data Domain systems give them an excellent solution for effective disk-based backup.
3. *Have access to an archive tier for long-term data retention.* In addition to directly supporting archived data, Data Domain Archiver lets IT set policies to automate data movement from the active to the archive tier. Administrators create a policy to move weekly full SAP backups to the archive tier and retain them for 7 years. The result is a highly effective archiving system that grows along with the company.

Taneja Group Opinion

Today's backup and archive environments are more challenging than ever. Backup operations must meet service level agreements for speed and recoverability even in the face of extreme data growth. And archiving is critical for business processes that impact aging and inactive data including eDiscovery for litigation and reporting for compliance. These processes and the groups who use them depend on the ability to search, restore and defend relevant data.

EMC SourceOne and EMC Data Domain systems combine to fulfill these requirements in a solution that offers powerful features and economies. This level of flexible integration enables organizations to keep backup and archives online for long periods of time at a fraction of the rack space and energy demands of traditional backup and archival storage, and with far more reliable and faster restore than tape.

EMC SourceOne and EMC Data Domain systems are excellent products in their own right. Putting them together into a single archiving and backup solution yields tremendous benefits for long-term data retention and security, data availability, saving space and energy costs, and real-life information governance. We urge companies to seriously consider the strong advantages of leveraging EMC SourceOne with EMC Data Domain Systems.

NOTICE: The information and product recommendations made by Taneja Group are based upon public information and sources and may also include personal opinions both of Taneja Group and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. Taneja Group, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors that may appear in this document.