EMC Data Domain as an Archive Platform with Symantec Enterprise Vault

By Jason Buffington, Senior Analyst

August 2013

This ESG White Paper was commissioned by EMC and is distributed under license from ESG.

© 2013 by The Enterprise Strategy Group, Inc. All Rights Reserved.
Contents

Summary..................................................................................................................................................3

Introduction...........................................................................................................................................3
  How Archives and Backups Complement Each Other .................................................................3
  The Evolutionary Confusion and Convergence Between Long-term Backups and Archiving ..........4
  Archival Adoption Trends ..................................................................................................................4

EMC Data Domain................................................................................................................................6
  Assured Data Survivability ................................................................................................................7
  Flexible and Integratable Protection Storage ...................................................................................8

Symantec Enterprise Vault ....................................................................................................................9

Better Together: Backup and Archive Powered by Data Domain Systems ........................................9
  Data Domain Optimizations for Enterprise Vault .............................................................................10

The Bigger Truth ................................................................................................................................12

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of The Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at 508.482.0188.

© 2013 by The Enterprise Strategy Group, Inc. All Rights Reserved.
Summary

This white paper describes ESG’s research related to trends in data archiving, considers EMC Data Domain deduplication storage systems as archival repositories, and discusses the potential merits of utilizing EMC Data Domain systems in combination with Symantec Enterprise Vault archival software.

Introduction

For a while now, a certain amount of confusion has been evident in regard to “backup” versus “archiving.” The two processes are not the same; backups maintained long-term are not, were not, and will never be archives. This confusion about the processes probably stems from:

- A desire by some IT vendors to create “a different conversation” with prospective customers.
- A widespread lack of hands-on experience regarding the “nuts and bolts” differences between the two processes.
- Everyone’s general reluctance to “add” a process.

The good news is that when the distinctions between the two terms do become clear to people, they can start to enjoy the separate but complementary benefits of each activity.

In general, ESG defines digital archiving as the long-term retention and management of electronic information that has been purposefully retained to satisfy records management, data management, regulatory compliance, or litigation support requirements. “Backup data” differs from “archive information” as follows:

- Backup data is typically a temporary copy of a data set that is ultimately overwritten.
- Archived information is typically moved, not copied, from one system to another and is often a permanent record or data set stored without alteration or deletion for a specified time period.

Archives play specific, complementary roles in the larger landscape of data protection:

- **Storage efficiency by moving data out of primary storage and into an archive** optimizes the production storage by reducing the consumption of valuable capacity by stagnant data—while reducing the amount of data to be backed up, resulting in shorter backup windows and less data being transmitted and stored within the backup storage pool.
- **Retaining data within an archive** absolves organizations of some of the long-term retention requirements they were trying to satisfy using their backup toolset. By using archive software to determine what data should be retained, the organization optimizes the long-term retention pool, makes it more relevant for granular item search, and focuses the backup software on more frequent and granular short-term copies.
- **Termed “defensible deletion,”** the same kinds of archiving policies that can determine how long something should be kept can also determine when something should be destroyed—resulting in more manageable legal hold and e-discovery scenarios and further reducing storage consumption costs.

How Archives and Backups Complement Each Other

It cannot be overstated that a “long-term backup” is not an “archive”—but instead, one should really be thinking about archives as a complement to backups.

Archived data is usually acted on due to the characteristics or contents of the data, whereas backups are typically retained simply based on the age of a file according to a generic (non-data-specific) retention policy. Unfortunately, that means that even long-term backups will eventually expire and the data will be lost, without any awareness of the relevance of the content. This reality makes archiving a much better way to preserve data for longer periods because the right data can be retained longer, while the rest of the data can actually have a shorter (backup) retention window and thus maximize the storage usage for both backup and archive mechanisms.
Adding an archival capability, in conjunction with a backup solution, to one’s data protection and management strategy achieves:

- **Efficient utilization of production storage**, wherein stagnant data is removed, thereby reducing primary storage OPEX and CAPEX and reducing the amount of data to be perused during backups.
- **Efficient utilization of protection storage**, wherein both the backup and archive data sets benefit from storage reduction and ensured data integrity regardless of the retention period—generic backup data can be retained for shorter periods of time, while archival data is preserved for longer periods.

**The Evolutionary Confusion and Convergence Between Long-term Backups and Archiving**

With those easily tangible benefits in mind, it’s easy to see why archives are appealing—and thus, why some backup vendors were so quick to muddy the waters and claim that their products offered archival capabilities. Of course, another reason “long-term backup” and “archiving” were misconstrued in the past is that the components involved in the processes actually are similar. Both are driven by software on a server, with agents or other connections to the primary storage that create data transmissions to a pool of (preferably) optimized storage. In both cases, the server-based software uses rules, policies, or schedules to determine the frequency and the target data to be acted upon. The result is a secondary tier of information on protection storage that is outside of primary storage.

With the commonality of architectures in mind, and understanding that backup and archive are complementary, it is not surprising to see convergence between the technologies and messaging related to backup and archive:

- **In the early days**, product marketers, salespeople, IT journalists, and other “guilty parties” incorrectly melded long-term backups with the term “archives,” creating confusion in the industry and producing an underwhelming adoption of real archiving tools by encouraging the inaccurate assumption that a backup’s retention mechanisms were adequate for long-term retention requirements.
- **Later**, as more advanced IT organizations began understanding the differences and benefits of each activity, some began deploying archival solutions that were completely oblivious to the backup solution—using separate storage, retention rules, and even administrative personnel. The result was less effective, more expensive silos of information with gaps in coverage and redundancies in equipment and processes.
- **Today and going into the future**, the industry is beginning to consolidate with best-of-breed capacity-optimized solutions that allow organizations to meet their complete data protection needs across backup, archive, and disaster recovery (DR).

**Archival Adoption Trends**

ESG believes we are in the midst of a significant increase in the use of digital archiving across all media types (with a 56% CAGR from 2010 to 2015). Today, mature economies such as North America and Europe accounted for the great majority of worldwide digital archive capacity—with economies in Asia, Latin America, and elsewhere expected to catch up. Still, the increase in share represented by those emerging economies is muted by North America’s and Europe’s increasingly stringent regulatory environments, which continue to drive retention requirements and digital archive capacity in those regions. In a similar vein, ESG believes that archived data stored at cloud storage or archive software-as-a-service (SaaS) providers will increase dramatically. File-based content already dominates cloud archives and should continue to account for the vast majority of all cloud-archived data.
ESG also believes that data associated with clinical applications will grow at a much faster rate than administrative data. For example, e-mail, administrative applications, and general unstructured data were all expected to grow at less than a 30% CAGR between 2010 and 2015. In contrast, data associated with clinical imaging and electronic health record systems is expected by ESG to grow at 42% and 61%, respectively.¹

Companies of all sizes, across all geographies, continue to struggle with data growth—including unstructured file data, databases, and e-mail (see Figure 1).²

Figure 1. Digital Archive Forecast, 2010-2015

When considering the upward trends in archiving and the high prioritization of improved backups within IT environments, it is important to understand the two key initiatives that will affect overall IT success as data continues to grow:

- Efficient data management includes moving stagnant yet still relevant data from primary storage to protection storage.
- Efficient data protection requires preserving data (i.e., archiving) for governance and regulatory compliance through long-term retention.

The convergence of those two IT mandates, along with evolving data and storage technologies and enablers, has created a growing market for advanced archival solutions.

---

EMC Data Domain

From a storage perspective, it is impossible to consider a modern data protection infrastructure without deduplication appliances. In that market, EMC Data Domain deduplication storage systems are often considered the innovation and mindshare leader, with protection storage solutions offered for SMEs via the Data Domain DD160 device on one end, all the way through support for large enterprises provided via the DD990 (see Figure 2).

Figure 2. EMC Data Domain Product Family

It is an underestimation to consider Data Domain appliances to simply be deduplicated storage arrays. A broader description, “protection storage,” really is more accurate because the technology does more than optimize storage consumption through deduplication:

- It also protects backup and archive data through the Data Domain Data Invulnerability Architecture.
- The Data Domain Replicator software option for efficiently replicating backup and archive data for DR.
- The Data Domain Extended Retention software option for backup data enables more efficient long-term retention.
- The Data Domain Retention Lock software option provides archival data with file-locking capabilities in order to meet internal governance or compliance regulations.

By combining an assurance of data survivability with various software options for enhanced backup, archival, and DR scenarios, EMC is further positioning its Data Domain family as an ideal “protection storage” platform upon which software vendors can build their combined solutions.
As Figure 3 shows, Data Domain systems also offer the ability to logically separate backup and archive data on the same system. This capability enables an organization to apply compliance-level retention and a separate replication schedule on only the archive data stored on Data Domain.

Figure 3. Understanding the Data Domain “Protection Storage” Capabilities and Features


As described in the ESG Market Landscape Report on Disk-based Target Systems, one should look for a few key characteristics when considering the “the protection of last resort” for archive storage needs including:

- Assured data survivability
- Optimization for data deduplication and overall performance (throughput)
- Strong integration between hardware and software for a more agile, comprehensive solution

Assured Data Survivability

Whether considering the criticality of recovering data after backup or accessing archive data that has been moved off of primary storage (and thus may be the last surviving instance of that data), the most important capability for protection storage is arguably the assurance of data survivability. In other words, you must be absolutely sure that you will be able to recover and access the data in a readable format when you need to.

- **For tape/traditional backups**, a corrupt file is inconvenient but perhaps not a catastrophe if it can be backed up again from primary storage. However, the corruption probably wouldn’t be discovered until a recovery was attempted, which means the uncorrupted file wouldn’t be on primary disk anymore. Therefore, it would have to be recovered from the DR copy (offsite tape, most likely, in this scenario).

- **For deduplicated backups**, a corrupt “piece” or segment of data has a much broader impact because an invalid iteration of even a single segment could affect the recoverability of tens to hundreds of files that reference the same single segment of data.

- **For archives**, where the data is moved off primary storage or is intentionally isolated and protected within archival storage, those files may be the only copies left. Corruption within an archival storage pool simply cannot be allowed to happen.

Data Domain systems were built from day one to address those concerns with the Data Domain Data Invulnerability Architecture (see Figure 4). The Data Invulnerability Architecture is an assurance that data is checked upon ingest after writing and deduplication, then periodically as it resides on disk, and once again upon recovery and retrieval request—to ensure that one’s “data of last resort” is there (accessible and uncorrupted) when it is needed.

---

3 Source: ESG Market Landscape Report, *Data Protection Target Devices*, May 2013. Please note that the Market Landscape Report covers backup targets; however, the same logic applies to archive targets when it comes to the key characteristics to look for when evaluating archiving systems meant to serve as “the protection of last resort.”

© 2013 by The Enterprise Strategy Group, Inc. All Rights Reserved.
Flexible and Integratable Protection Storage

With its foundation in data integrity and a laser focus on deduplication/compression for optimized protection storage, Data Domain systems are designed to support a wide range of backup and archival use cases by enabling better-together solutions with a wide variety of leading backup and archiving applications.

Data Domain systems are tightly integrated (via DD Boost integration) with EMC’s own NetWorker software and Avamar platforms. It is notable that the Data Domain family of appliances is also broadly leveraged by a significant number of non-EMC backup software providers (such as Symantec NetBackup). Likewise, Data Domain systems are optimized to support not only EMC’s own archiving software, SourceOne, but also other industry-leading products such as Symantec Enterprise Vault for file, e-mail, and SharePoint archive solutions.
Symantec Enterprise Vault

In much the same way that many people consider Data Domain systems to be the de facto standard in protection storage, Symantec Enterprise Vault is thought of as the de facto software leader in archiving for file server, e-mail, IM, and content/collaboration platforms (see Figure 5).

Figure 5. Solution Overview—Symantec Enterprise Vault

Enterprise Vault solves multiple application platform challenges related to preserving multiple content types: In particular, it has had much success in archiving content generated by Microsoft Exchange environments, file systems, Microsoft SharePoint, instant messaging platforms, Bloomberg terminals, and social media platforms.

Symantec Enterprise Vault uses various classification technologies to analyze, categorize, index, and store data according to corporate and regulatory mandates. Specifically, it can support faster backups, improved BC/DR, and an increased level of regulatory compliance. Using Symantec Enterprise Vault archiving with Microsoft Exchange reduces the stress put on users through mailbox quotas and various policies related to offline PSTs. Similar improved user experiences and operational benefits also can be seen in Windows file services and Microsoft SharePoint platforms.

It is the storing activity that can take real advantage of the optimized storage capabilities of a Data Domain system—by leveraging EMC’s Data Invulnerability Architecture (for assured data validity) and the Data Domain retention software options—in order to drive further efficiencies into an IT infrastructure.

Better Together: Backup and Archive Powered by Data Domain Systems

Because of the dichotomy in years past—when immature or archaic archival approaches were being confused with long-term backup approaches—many environments began deploying parallel solution architectures. The parallel architectures had separate backup/archive software tools, separate IT admins, and separate repository silos. And as each architecture evolved, it did so at its own pace—often without benefiting its “sibling’s” data protection efforts.

Although most backup architectures are recognized as optimal when they combine intelligent software with optimized protection storage, not everyone sees the parallels and implements a similar technology combination for archival use cases. And many of the archive architectures that do leverage optimized storage miss the opportunity to be a smarter architecture that leverages the same protection storage for both backup and archive repositories. But that consolidation of architectures and efficiencies only works if the storage is designed and optimized to
benefit archival use cases. Data Domain systems are designed to serve both use cases, leveraging built-in data integrity protection, highly efficient inline deduplication, and compliance capabilities.

Backup and archive are complementary data protection approaches, and they often depend on each other to be best implemented (for example, not having to back up as much data because some of it is archived). Moreover, in the case of both backups and archives, the whole secret to an effective implementation is to have protection storage that is optimized by deduplication and integration with the software that will be driving it.

And, while all of the earlier archival-plus-backup synergies apply between EMC Data Domain systems and Symantec Enterprise Vault, it is equally interesting to note the two companies’ leadership positions in respect to deduplication and archiving (respectively). They have ensured that their specific technologies not only work well in conjunction, but in fact are also optimized to deliver an even better experience together (see Figure 6).

**Figure 6. File, E-mail, and Microsoft SharePoint Archiving with Enterprise Vault and Data Domain**

![Data Domain and Enterprise Vault Diagram](image)

**Data Domain Optimizations for Enterprise Vault**

EMC recently announced enhancements to its Data Domain OS (version 5.3) that provided significantly optimized handling of small files within its deduplicated storage system. Although traditional backup solutions have natively taken advantage of Data Domain deduplication methods, archival solutions such as Enterprise Vault weren’t previously able to utilize the Data Domain systems in their native configuration, instead having to store EV’s numerous small files within compressed “CAB” containers on the Data Domain system.

With DD OS 5.3 or later, Data Domain systems are able to efficiently store the smaller files while providing the performance and deduplication benefits that software providers count on from EMC. In particular, this newly supported scenario benefits customers using Data Domain and Enterprise Vault by enabling them to leverage **Enterprise Vault savesets** (natively storing archival data such as files, messages, or objects in their more optimal and granular formats). This further enables:

- **Data Domain as primary vault store for Enterprise Vault configurations.** In comparison, most traditional storage devices must be used as secondary vault stores for Enterprise Vault, forcing Enterprise Vault to consume space on the production servers’ storage for prepackaging of savesets into numerous 10MB (or bigger) “CAB” files before transmission to the storage device hosting the Enterprise Vault secondary vault store. More than just utilizing the space on primary storage, such a practice can lead to sizeable OPEX investments related to the sizing of temporary space on production storage, etc. Because Data Domain systems can now be accessed as a primary vault store, users can eliminate those additional OPEX investments that many traditional storage providers/approaches impose.

- **Faster ingest for mailbox/journal archives and PST file migrations** to enhance the support of the fundamental use cases of mailbox archive, journal archive, and e-discovery for Symantec Enterprise Vault. Instead of offline e-mail containers (e.g., PSTs), which are harder to protect and catalog and which can create uncertainty during e-discovery efforts, Symantec Enterprise Vault can import PSTs into its archive, optionally deleting the PST afterwards. The result is a singularly managed e-mail system. Users are
unencumbered by mailbox quotas or self-managed data grooming, and the company has what it needs for
governance and defensible deletion of what it doesn’t need. With the recent enhancements, EMC is
claiming more than a ~3X increase in both ingest and recall rates.

Note: In regard to those small files—and depending on the total number of objects to be archived within the
Data Domain system—different configurations of Enterprise Vault’s single-instance-storage (SIS) data
reduction in combination with Data Domain compression/deduplication may be more appropriate than
turning SIS off to take full advantage of Data Domain’s deduplication optimizations.

Perhaps the most compelling aspect of the combined solution is the simplicity with which Enterprise Vault
customers can take advantage of the benefits of Data Domain’s protection storage, which is simply a matter of
declaring a new Vault location within the Enterprise Vault interface (see Figure 7). Following that step, all of the
integration between the Symantec and EMC solutions is transparently configured for the combined solution.

Figure 7. Enabling Data Domain Storage Within Enterprise Vault

In a time when many IT environments are looking to reduce complexity, some are running to the cloud in hopes of
avoiding infrastructure challenges. But with topics such as “the data of last resort” and threats of e-discovery that
are bound to how quickly one can identify and gather the pertinent data, on-premises archival solutions will likely
continue to be the norm with large enterprise customers. Moreover, in consideration of the market presence of
Enterprise Vault as archive application software and Data Domain for protection storage, the complexity-reduction
solution that IT managers seek may already exist on their data center floor. They simply have to start using the two
technologies together.
The Bigger Truth

Backup and archiving are not the same. But they are complementary in a comprehensive data protection strategy, and you need to be doing both.

For too long, various vendors confused the public by referring to any “long-term backup” as an “archive,” which did a disservice to both of those complementary data protection tasks. A backup is for restoration, while an archive is for retention. Archives also help manage storage growth (and improve backup processes) by moving unused data or enabling policy-driven, defensible deletion. Archives are suited for long-term retention; backups are most useful in a much more timely solution.

But as much as the two initiatives are separate, they should not be deployed and managed in isolated storage silos. Not only should they interoperate enough that they benefit from each other (e.g., by reduced backup windows from smaller production data sets), but they also should be thought of as complementary and mutually necessary focal areas for data management. By the same broader understanding, data protection includes not only backup and archiving, but also snapshots and replication.

With so much architectural commonality and so many better-together implications, it should not be a surprise that some of the leading data protection companies also are leading the way in archiving, including spearheading the synergies with backup. EMC provides optimized storage through its Data Domain systems for all leading backup and archiving applications. EMC’s innovations to optimize Data Domain for archive data enable its market-leading Data Domain appliances to be the “protection storage of choice” for backups of all sizes and, now, archives as well. This should be welcome news for Symantec customers who are using Enterprise Vault as their archival solutions, particularly if a Data Domain system is serving as the protection storage for backup.

The industry is finally beginning to understand the distinction between backup and archiving. With two market giants bringing their formidable technologies together in a complementary way, IT organizations that recognize the power of backup plus archiving—and the potential of Data Domain systems plus Symantec Enterprise Vault as a comprehensive data protection strategy—will benefit.