

Deduplication Storage for Nearline Applications

Abstract

As the demand for storage capacity grows in today's enterprises, customers are continually looking for more cost-effective ways to store data while also keeping it online for accessibility. As a result, the requirements for disk capacity, data center space, power, and cooling are rising beyond sustainable levels. Finding new ways to store data with less disk footprint is becoming a critical challenge for today's IT organizations. Data Domain deduplication storage systems can be leveraged to tackle these challenges and provide a more cost-effective nearline storage solution. This paper explores the different potential uses of Data Domain systems for a broad spectrum of nearline storage applications, and how Data Domain offers a new level of savings for IT organizations as a result.

Deduplication for Nearline Applications

Table of Contents

- INTRODUCTION: CONSOLIDATED SUPPORT FOR BACKUP, ARCHIVING AND OTHER NEARLINE APPLICATIONS. 3

- APPLICATION PROTECTION UTILITIES. 4
 - ORACLE RECOVERY MANAGER 4
 - OPERATING SYSTEM IMAGES. 4

- ONLINE REFERENCE DATA STORAGE 5

- ARCHIVING 6
 - FILE ARCHIVAL 6
 - EMAIL ARCHIVAL 7
 - ARCHIVING FOR CORPORATE GOVERNANCE. 7
 - ENFORCED RETENTION 7
 - OPERATIONAL FLEXIBILITY 7

- SUMMARY 8

Consolidated Support for Backup, Archiving, and other Nearline Applications

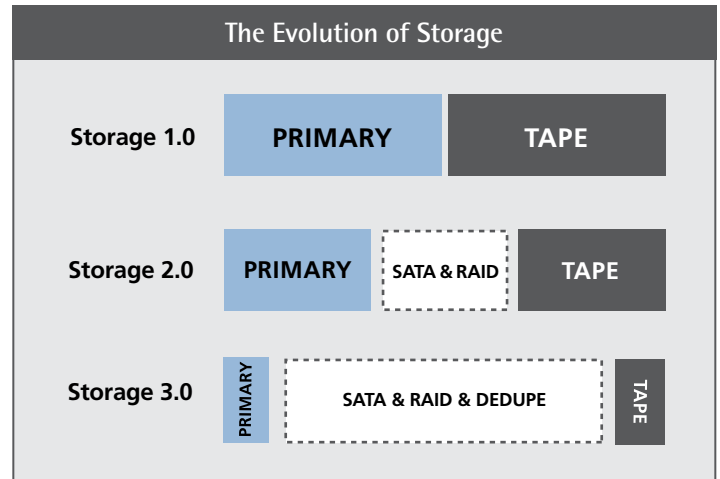
Over the recent decade, a fundamental shift has been happening in the way storage systems are deployed in enterprises. Through the emergence of a new, broadly used tier of storage based on ATA drives in RAID configurations, data stored for light use reference, backup and archiving has moved from expensive primary storage and awkward tape storage to nearline network disk storage. While it was broadly dismissed in the early part of the decade as not being sufficiently fast or safe, it turns out it was good enough, and it now accounts for a significant fraction of corporate storage deployments.

This process is about to be further expanded through the emergence of deduplication in the nearline storage tier. Data Domain has proven that deduplication can be reliable, fast and easy to manage, all while providing massively disruptive economic differentiation for backup deployments seeking greater retention, replication and recovery speed. A consensus is now emerging that for many other non-backup nearline applications, deduplication would also be extremely impactful.

Why: most data creation in the enterprise is the result of the copy-and-paste culture enabled by storing many versions of the same information. 'New' information typically includes mashups of parts of prior documents. But storing all this versioned data leads to data redundancy across documents, versions of documents and systems. While it would be useful to keep it all on disk, all the versions needed for just-in-case planning, data protection and litigation support is much larger than the data in primary storage, sometimes 5x-10x larger. Even if every document was filed uniquely, the amount of data redundancy within and across those documents can often be reduced by 80%-90% using appropriate deduplication technology. And those are not extreme factors, since most sites do not file every document uniquely; documents tend to get copied.

Data Domain has always offered a NAS (network file sharing) interface to its systems, but until recently, other behavioral optimizations had been focused narrowly on the backup use case only. This has changed. For example:

- In a backup store, files are typically very large, 'tar'-style aggregates of changed files. In other nearline application stores there might be several orders of magnitude more files, even before snapshots. Data Domain systems have now been tuned to support both workloads.
- In backup, point-in-time filesystem snapshots are not generally expected; while they might offer an 'undelete' capability, file versioning is already handled by the backup software and catalog. In other nearline applications, snapshots are the best practice for version storage and recovery. Since the introduction DD OS 4.3, Data Domain includes snapshots. Unlike all but a



Combining data deduplication with traditional nearline storage features like SATA and RAID will accelerate the use of nearline storage in the enterprise

few vendors' snapshot implementations, Data Domain snapshots are very lightweight in performance terms; more snapshots does not slow the system down. And unlike any other vendor, all data written to a Data Domain system benefits from deduplication; Snapshots are a simple, additional feature so it is very easy to have both deduplication and snapshots without having to carefully plan how they interact.

Data Domain offers the industry's leading inline, scalable, general purpose implementation of deduplication. Supporting both the Data Involvement Architecture and network-efficient replication, it enables Data Domain systems to be the first deduplication storage to support both the broad range of backup applications and the even broader range of other styles of nearline use.

What is a nearline application? Broadly, it is anything that does not require the high I/O per second (IOPS) and high availability (HA) focus of primary storage. While primary storage is judged on \$/IOPS and redundant/failover technology, nearline storage is judged on \$/GB and data resilience. It includes applications that usually read or write a whole file at a time, for protection, reference archiving, or even simple access, like a /Home directory share. While backup applications are discussed in many other pieces of Data Domain collateral and by its technology partners, this paper is an introduction to well-known examples of other nearline applications and the benefits of Data Domain. In particular, this paper will discuss:

- Application protection utilities (e.g. Oracle RMAN, VMware OS images)
- Online reference data storage (e.g. project data, mothball filesystems)
- Archiving storage (e.g. email, file and database archiving applications)

Application Protection Utilities

Many applications come with utilities that allow you to create versions of application-specific data. Data Domain deduplication storage systems offer a superior solution for hosting this type of data.

Oracle Recovery Manager

Oracle Recovery Manager (RMAN) is Oracle's preferred method for efficiently backing up and recovering Oracle databases. RMAN optimizes performance and capacity consumption during backup, and it takes care of all underlying database procedures before and after backup or restore operations. It provides a common interface for backup tasks across different host operating systems and offers features not available through user-managed methods, such as parallelization of backup/recovery data streams, retention policies, and a detailed history of all protected images.

The Challenge

In order to fully-benefit from using Oracle's RMAN utility to dump and maintain multiple versions of a complex database, database administrators require the ability to retain many copies on disk. Disk-based snapshots provide an efficient storage technology. However, even with this technology, traditional disk-based storage becomes too costly to deliver the required retention of this data, which administrators need in order to make an RMAN-based solution effective.

The Data Domain Solution

Users can gain several advantages by combining the use of RMAN with the Data Domain systems to store RMAN data. The first benefit is that the amount of physical storage needed for a copy of data is significantly reduced due to an average of 20x to 60x reduction of RMAN data. This also allows more versions of data to be stored on a fixed amount of physical disk or enables longer retention periods on disk for RMAN data. Next, backup and recovery times are reduced significantly by backing up directly to disk rather than tape. Many examples exist where Oracle users are economically maintaining 30 days of nightly full backups on a single Data Domain system as a result. Other databases, such as DB2 and SQL Server, have similar capabilities that can leverage Data Domain systems for efficient storage of database versions.

Operating System Images

The economics of server infrastructure virtualization with products like VMware are well understood. It is not uncommon for a given deployment to include hundreds of virtual machines. Consequently, IT departments need to develop a strategy for storing and managing Operating System (OS) images. As virtualization deployments increase and mission critical applications are built on top of a virtualized architecture, data protection of the OS images becomes a critical requirement.

The Challenge

Disk-based backup and rapid recovery of VMware virtual infrastructure environments is preferred over using tape. However, a disk-based solution is not always practical, primarily from a cost perspective, without massive data reduction.

There are some challenges particular to backing up a virtual infrastructure. It is very hard to optimize space when backing up versions of OS images. The images change slightly on a daily basis and therefore need to be backed up to tape or disk, resulting in a high demand for space. As a result, disk-based backup of OS images becomes too costly. This in turn results in slower recoveries due to having to use less-expensive media, like tape, to recover the images from.

Attempts to deduplicate VMware backup have come to market where the data analysis happens within the VMware system, but this creates the wrong dynamic. VMware succeeds by increasing the number of applications managed within a single physical server; deduplication within the same server diverts those CPU cycles away from applications. It pulls in the wrong direction.

The Data Domain Solution

Data Domain deduplication storage delivers the level of data reduction required to make disk-based backup and DR for VMware environments cost-effective. VMware images tend to have a lot of redundant data given that differences in the OS images are typically related to a small subset of the data dealing with hardware dependencies or application software variances across images.

VMware users that leverage Data Domain for their disk-based backup are able to save multiple copies of vmdk files and traditional backups for months at minimal cost. Disaster recovery in these environments is also more reliable since the data can be replicated to a DR site through the use of Data Domain's network-efficient, bandwidth-saving replication capability. The Data Domain Data Involvement Architecture, which continuously validates the integrity of the system data to ensure that it can always be restored, adds another level of data integrity to the overall solution.

Many options exist to backup VMware environments. VMware includes consolidated backup support and integration with leading backup applications from Symantec, IBM, EMC and others, which allows users to avoid running backup agents within running virtual machines in order to back them up. In addition, VMware-focused utilities like vRangerPro from Vizioncore are also available to provide data protection for VMware environments. All of these offer an integrated and efficient VMware backup solution that can leverage Data Domain systems as disk-based targets for VMware backups.

As a result, VMware users can now deploy a cost-effective disk-based backup solution for their virtual infrastructure by leveraging Data Domain as a platform to store the backup data, yielding longer disk-based retention, improved recoverability, reduced backup times, as well as the ability to replicate to a DR site.

Online Reference Data Storage

Nearline storage with deduplication provides an ideal platform for hosting a wide variety of nearline content data that can deliver many operational and productivity benefits to an organization when stored on disk, even though it may not be accessed very frequently. In many cases this is data that would not have been a candidate to store on traditional nearline storage due to cost. Home directories, particularly archived directories of inactive employees, or other administrative data, are candidates to store on Data Domain nearline storage systems. Construction and design firms have many historical projects that need to be maintained on disk for easy access on an ongoing basis which can also be stored most economically on Data Domain nearline storage.

Another example includes multiple versions of engineering file libraries for a product under development, which can deliver productivity and efficiency benefits for the engineering staff when many historical versions are stored on disk for reference.

Example Challenge: Software Build Versions

Disk-based data stores of this type of nearline data can be costly to an enterprise due to the amount of capacity required for long retention periods and the management overhead incurred by storing large numbers of files on disk. For example, a software development firm may find it useful to create on-disk libraries of daily versions of the source files used to generate its product (or the builds themselves), which are typically stored in the databases of software configuration management tools. Without maintaining on-disk libraries of these intermediate versions, engineers need to copy out files to local storage attached to a workstation or home directories on expensive networked storage whenever an older version is needed for debugging or research purposes. However, these data sets can quickly grow to hundreds of millions of files over a short amount of time, making it very difficult to maintain full copies on disk. This can be both costly in terms of storage capacity

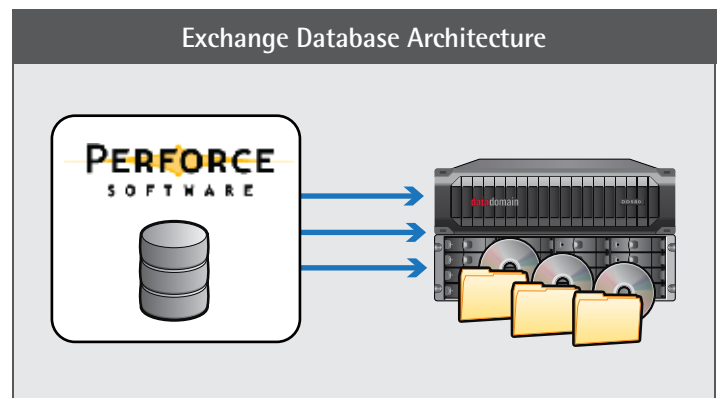
and the management overhead required. However, not having this data readily available on disk can negatively impact the productivity of an engineering and development organization.

The Data Domain Solution

Data Domain's efficient deduplication storage systems offer an economical option for storing this type of project data. Given that deduplication happens inline and automatically, before data is stored on disk, administrators do not have to worry about scheduling resource-consuming post-processing deduplication cycles, allowing for this data to be stored at random, whenever the need arises.

For example, an engineering organization can dump the source files out of a configuration management database maintained by applications like Perforce, Microsoft Visual SourceSafe, or IBM Rational ClearCase, to the Data Domain system at regular intervals to create online source file and build version archives, which engineers can reference for quick debugging, builds, or research. Data Domain customers are generating this data multiple times a day in many cases and storing it, given the superior economics that a deduplicated nearline storage system offers. In one instance, a wireless technology company stored ten million files to a Data Domain system over the course of a month, achieving 13x data reduction.

Easy access to disk-based archives of this data results in increased productivity and time-to-solution for the engineering organization, all at minimal incremental storage costs delivered by Data Domain nearline storage systems.



Software configuration management tools can copy out source file libraries on a regular basis to a Data Domain system

Archiving

Disk-based file, email, and database archiving is a growing requirement across a wide spectrum of enterprises. By storing the data online using disk-based deduplication storage as opposed to traditional methods (tape, CD, DVD), the data can be more easily retrieved and searched. As a result, disk-based archiving systems are being deployed in widespread fashion for many types of heterogeneous application data.

The Archiving Challenge

Customers often classify data based on its business value in order to more-effectively match the value of the data with the tier of storage on which it resides. However, archived data, as opposed to backup data, tends to be retained on disk for years, not just weeks or months. This increases the demand for storing this data in a cost-effective manner while also maintaining its integrity over time. If a storage platform is chosen that is cost-effective for storing this data, but does not offer effective data integrity features, the long-term usability and integrity of the data is put at risk.

The Data Domain Solution

Consolidating heterogeneous application data on a single archive platform increases the return-on-investment of the solution, and Data Domain inline deduplication storage can significantly increase the return even further.

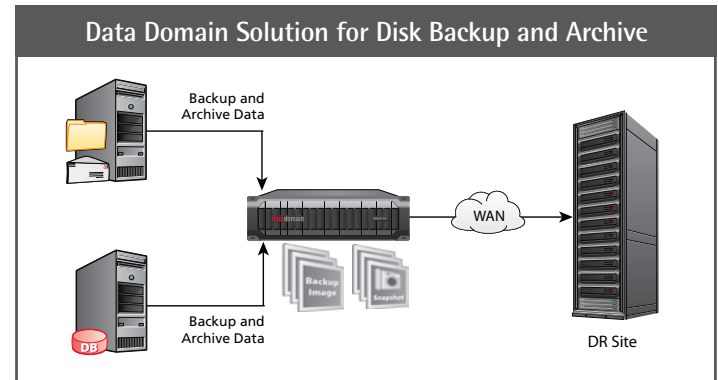
Because Data Domain offers deduplication in a general purpose storage platform, it can deliver benefits to a wide range of archiving applications. By storing the archived data on a Data Domain nearline storage device, data can be reduced significantly, depending upon the information being stored and the versioning policy.

For example, while it is common for only single-instance copies of archive files or emails to be copied in to the archive store, deduplicating this kind of data can represent significant cost savings in storage and replication bandwidth. Because deduplication looks for small, redundant sequences across files, this is much more aggressive approach to data reduction than, for example, most CAS systems, which typically single-instance whole files only. Depending on the data, this can represent a 60%-90% improvement.

In addition, further gains in reduced capital and operating expense can be realized when consolidating backup and archive data on the same Data Domain system, since these applications often store multiple versions of the data to enable different types of recovery. Storage and indexing for litigation support often require different techniques than storage for high speed system recovery.

Data Domain also offers a Retention Lock software feature to allow IT departments to easily implement data retention rules dictated

by corporate governance policies or government regulations. As a result, users gain the benefits offered by Data Domain, the leader in deduplication storage. Data Domain delivers efficient, nearline storage systems with deduplication, along with built-in data retention capabilities that also provide the operational flexibility that administrators require to efficiently run the enterprise on a day to day basis.



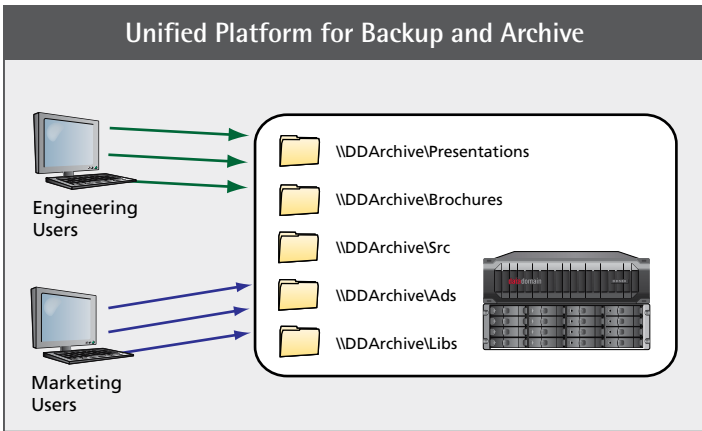
Leverage a single platform for backup and archive using Data Domain

File Archival

Many enterprises are interested in archiving unstructured file data in order to reduce the capacity it consumes on expensive primary storage. By moving older or less-frequently-used files to archive storage, IT managers can lower their storage costs. Archiving can also reduce the amount of data that needs to be backed up frequently. The primary data set being backed up can be reduced in size by moving some of it to archive storage, where it can be backed up less, or not at all since efficient snapshots and/or replication can also be utilized to protect this data.

Many customers employ their own scripts and utilities to facilitate this as well, including simple drag-n-drop data movement to migrate data to nearline storage, where it can be stored more economically.

File archiving can also be completely driven using purpose-built applications that implement automated policies and orchestrate both the archival and optionally the recall of files from primary storage when they are accessed. Examples of these types of full-featured file archiving applications include CommVault Data Archiver, Arkivio Auto-Stor, EMC Disk Xtender, AXS-One, and Symantec Enterprise Vault.



Users can access archived data directly from a Data Domain system in a Windows environment

Email Archival

Email is a rapidly expanding data set in most enterprises, and presents a growing data management problem to IT staffs. Email archiving systems are being deployed to reduce the overall cost of storage by moving older email data from the primary storage tier to an archive tier, lowering the demand for the most costly storage platforms. In addition, compliance requirements also drive users to implement email archival in their enterprise in order for IT managers to gain more control over how, where, and how long email is retained. By archiving email data to cost-effective storage automatically, strict per-mailbox size limits do not need to be imposed on users' data, which tend to drive more email data into distributed PST files (Exchange). This practice creates 'islands of data' that are difficult to back up and manage for compliance and legal discovery needs.

Many purpose-built applications exist that focus on the email archiving problem, including Symantec Enterprise Vault, EMC Email Xtender, CommVault Email Archiver, AXS-One Compliance Platform, and Mimosa NearPoint. Combining these products with deduplicated nearline storage offers a solution that delivers maximum benefits in the areas of storage efficiency and overall cost savings.

By coupling backup and email archiving to the same deduplicated store, additional data reduction benefits can aggregate. For example, a file that is stored in an archive may be emailed around to many recipients. In addition to storage in the Exchange database, it may be separately copied to many .PST files on desktops or servers. Email archives may store unique message instances. If all of these copies are stored to the same deduplicating storage, surprising amounts of redundancy can be identified and eliminated.

Archiving for Corporate Governance

For most businesses, much of the information stored in email and file archives is a key asset that requires careful management and protection. This business data is the foundation for many parts of a company's operations, whether it is finance, sales & marketing, or engineering. Ensuring that these assets are protected and continue

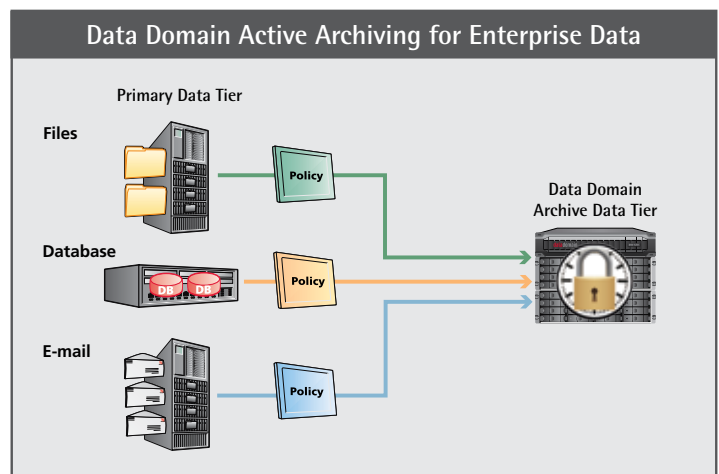
to be easily accessible is an essential requirement for IT organizations. In addition, many government regulations, as well as internal IT governance policies, require that critical business records, files, or email be retained for specific periods of time. Ensuring that these policies are being implemented and enforced is an ongoing challenge for IT staffs in many businesses. Data Domain's Retention Lock software option delivers the right level of data integrity assurance while providing the required operational flexibility to allow IT staffs to manage their archives effectively.

Enforced Retention

Data Domain Retention Lock software allows administrators to comply with both regulatory requirements and internal best practices for information retention by making files non-rewriteable and non-erasable, ensuring that critical business information is available until a specified retention date, at which time the information can be deleted if necessary. The retention parameters can be set on a file-by-file basis, and minimum and maximum retention periods can be set globally.

Operational Flexibility

Retention Lock combines the benefit of securing critical business information with other key features that ensure operational flexibility to allow IT administrators to efficiently and securely manage their storage environment. Retention Lock allows administrators to modify the security attributes of critical data if needed in order to adjust security for critical documents in the event that the security policy changes during the retention period. In addition, retention settings can also be adjusted by administrators with appropriate privileges if required in order to comply with new regulations or retention policies.



Archive different data sets to the Data Domain System using multiple policies. Use Retention Lock to securely retain the data to meet compliance and IT governance policies

Summary

Data deduplication technology can significantly lower both capital and operating costs in the data center by reducing the amount of storage required to serve nearline applications, simplifying and consolidating nearline storage and enabling much more efficient disk-based deployments.

Data Domain products now provide the key requirements that must be addressed in order for data deduplication technology to be leveraged by a wide variety of nearline applications. No other platform combines automatic, inline deduplication, snapshots, replication, data locking for compliance and corporate IT governance, and industry-leading data integrity features like Data Domain's Data Invulnerability Architecture in a single, easy-to-deploy NAS nearline storage system. As nearline data continues to multiply, Data Domain will be there to help you tame it.

Data Domain
2421 Mission College Blvd.
Santa Clara, CA 95054
866-WE-DDUPE; 408-980-4800
sales@datadomain.com
24 international offices: datadomain.com/company/contacts

Copyright © 2008 Data Domain, Inc. All rights reserved.

Data Domain, Inc. believes information in this publication is accurate as of its publication date. This publication could include technical inaccuracies or typographical errors. The information is subject to change without notice. Changes are periodically added to the information herein; these changes will be incorporated in new additions of the publication. Data Domain, Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time. Reproduction of this publication without prior written permission is forbidden.

The information in this publication is provided "as is". Data Domain, Inc. makes no representations or warranties of any kind, with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Data Domain and Global Compression are trademarks of Data Domain, Inc. All other brands, products, service names, trademarks, or registered service marks are used to identify the products or services of their respective owners.
WP-DFNA-0608