August 2009

By Jerome M Wendt

DCIG, LLC
7511 Madison Street
Omaha NE 68127

O 402.884.9594

A Candid Examination of Data Deduplication

How Enterprise Organizations may Leverage Symantec to Deliver and Optimize Data Deduplication within their IT Infrastructures

A Candid Examination
of Data Deduplication

Table of Contents

A Candid Examination
of Data Deduplication

Executive Summary

Data and information management continue to be a major problem in enterprise organizations that no one likes to think about. So when a new technology like data deduplication comes along that promises to solve today's backup problems, it is no surprise that organizations may view data deduplication as some kind of silver bullet that will provide a quick fix to their data protection issues. Unfortunately the dilemma that organizations find themselves in is that they are unclear as to how it will benefit them, what exactly data deduplication does, what ways it is implemented and what problems it can and should solve in their storage and backup infrastructure.

To clarify some of the confusion surrounding data deduplication, organizations need to understand how it works as well as what problems it solves and what options are available. To do that, they need to understand:

- The main benefits driving the adoption of data deduplication today and what business and technical problems it addresses in organizations

- The primary role of data deduplication to reduce the size of data stores (primary, archive and backup) while improving the backup and recovery experience

- The different ways in which data deduplication is implemented and under what circumstances each form of data deduplication is the right fit

- Enterprise organizations may need to obtain and then manage the multiple ways in which data deduplication is implemented

To achieve this last objective enterprise organizations need to consider the software that is part of Symantec's information management suite. It provides solutions that address these different needs around data deduplication so organizations do not to guess what the best form of data deduplication for them but can instead focus on deploying solutions that meet the specific needs of their different applications or business units within their organization. Symantec software allows customers to put deduplication technology as close to the source of data as possible.

Symantec Corporation

A Candid Examination
of Data Deduplication

The Data Deduplication Dilemma

Whether enterprise organizations like it or not, they are being forced to think about their data storage problems. Continual data and storage growth, shrinking recovery times and flat or declining IT staff counts and budget dollars are moving peripheral issues like archiving and backup to the forefront of corporate concerns.

Therefore it is no surprise that a technology like data deduplication is viewed as some type of silver bullet that will provide a quick fix to current data storage and protection issues. Data deduplication, when used as part of a broader archiving backup and recovery strategy, can increase backup and recovery success rates, reduce backup windows, decrease the amount of storage that organizations need to buy and improve the efficiency of applications as well as backups and recoveries. These benefits contribute to lowering the overall costs of data storage and data protection which only heighten the interest in data deduplication.

That is the good news. The dilemma is that there is an over-load of conflicting information about data deduplication due to the many variations, iterations and implementations in which it is available. As a result, they are uncertain how data deduplication works, what benefits it will really deliver and under what conditions they will benefit from the implementation of data deduplication in their IT infrastructure.

So in order for enterprise organizations to make an informed decision about the best way to proceed with data deduplication so that they select a solution that meets their specific needs, they need to understand:

- The business case for data deduplication
- How data deduplication works

- The primary ways in which data deduplication is implemented
- The benefits and drawbacks that each method of data deduplication presents
- Why they may need to implement data deduplication in different forms
- How to manage each form of deduplication

The Business Drivers for Data Deduplication

Data deduplication is useful in many scenarios but there are five main benefits that are driving its adoption in businesses as part of their overall enterprise data protection strategy:

1. *Permits the cost-effective introduction of disk into backup processes.*

   Organizations find using tape as a primary backup target problematic. Using tape results in slowed or failed backups and recoveries and causes tape management headaches so the interest in using disk as a backup target in lieu of tape is on the upswing.

   Using disk as a backup target without introducing data deduplication delivers many of the same benefits as when data deduplication is used—specifically higher backup and recovery success rates and faster backup and recovery times. The issue is that without data deduplication, there is so much unchanging, redundant data across application servers that the costs for disk can quickly skyrocket.

   This is what makes data deduplication so desirable. Once a data deduplication ratio becomes greater than approximately 8:1, it becomes cheaper to deduplicate

Users Still Unclear on Data Deduplication's Role

One user explained to DCIG that the education institution at which he works is still figuring out how data deduplication will help it or if it will even help his institution out at all. At his site he is responsible for protecting application servers in production, development and testing. Some need regular backups, some just need backups occasionally and others he does not need to protect at all. Since the data protection needs for his applications are so varied, he is unsure which variation of data deduplication is right for his institution's IT infrastructure or if it will realize sufficient benefits to justify its cost should he implement it.

In another circumstance, end-user confusion around deduplication is extending the life of legacy technologies like tape. While many analysts and industry pundits believe that data deduplication will lead to tape's ultimate demise, one consultant explained to DCIG that the users he supports are also still unsure as to which form of deduplication to implement. Because of the ambiguity surrounding data deduplication, they are opting to stick with tape and tape solutions for now until there is more clarity as to which data deduplication solution, if any, is the right one for them.

the data than to store the backup data on disk in a non-deduplicated state.

2. *Provides disk with some of tape's intangible properties.*

One of the primary ways that organizations determine whether or not to proceed with data deduplication within their backup processes is by comparing it to tape. In terms of cost, as data deduplication ratios reach and exceed approximately 15:1, the cost of data deduplication becomes on par with tape.

However tape has other features like infinite capacity and reduced power consumption. Since tape cartridges consume no power and provide infinite storage capacity (as a tape cartridge fills up, it can be replaced with a new cartridge creating the illusion of infinite capacity), these are other intangible properties of tape that some organizations still find desirable.

Data deduplication provides a reasonable facsimile of these intangible tape properties. It enables organizations to store large amounts of data on a fraction of the disk drives that would normally be possible without it and keeps the number of disk drives to a minimum which lower power and cooling costs. While not an exact one-to-one match, using data deduplication organizations receive the benefits of both disk and tape without their respective drawbacks.

3. *Facilitates the cost-effective off site replication and recovery of backup data.*

Off site replication is one of the best arguments for proceeding with deduplicating data stored to disk versus storing backup data to disk in its native format. Data deduplication makes it more affordable and practical to replicate data off site. Data deduplication reduces the amount of data to transmit over network links and organizations may be able to continue using their existing WAN links.

This also gives data deduplication another attribute of tape: data portability. Tape is often used for data movement because it provides a cost-effective means to store and move data off site. However by first deduplicating backup data stores, organizations can minimize or even eliminate the need to use tape as part of their data protection strategy while improving their off site recovery times.

4. *Controls the size and growth of archived data stores while improving application performance.*

The initials benefits that organizations expect to derive from archiving are reductions primary data storage, improved application performance and lower ongoing capital expenditures on primary storage. However archived email and file data stores can and often do contain large amounts of redundant data. Implementing data deduplication as part of the archive solution eliminates redundant emails, emails attachments and files in the archived data store. While it can potentially slow application performance when retrieving data from the archive, this data is rarely or never accessed and the storage savings more than offset whatever degradation in performance may occur.

5. *Well suited for the protection and recovery of virtualized server infrastructures.*

Virtualized server infrastructures contain high levels of data redundancy coupled with relatively low data change rates. These attributes of server virtualization play right into the strengths of data deduplication which make introducing data deduplication as part of a virtualized data protection strategy a natural fit in these environments.

A Simple Illustration of Data Deduplication in the Backup Process

Each solution deduplicates data differently and different solutions even deduplicate it from different points in the backup infrastructure. However the objective in every data deduplication solution used in conjunction with data protection is the same: Reduce the large amounts of redundant data found in backups and store it in a smaller footprint. The steps below provide a simple representation of how data deduplication works as part of the backup process.

First Backup

In the first backup using data deduplication, the majority of the data backed up and readied for data deduplication is new data so minimal data deduplication occurs in this example. Figure 1 shows the data in three different backup streams from three different application servers. Notice how there is some redundant data in each application server's backup stream which is identified during the deduplication process so that only unique data is stored.
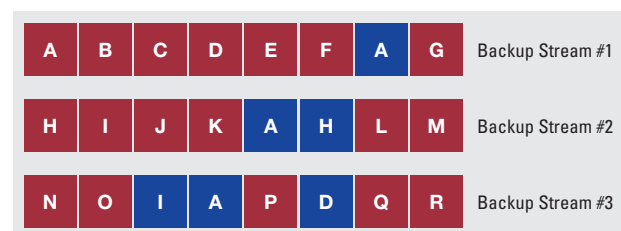

Figure 1   Unique data blocks are shown in **red** while redundant occurrences of data are highlighted in **blue**. Only the data blocks in **red** are stored.

Assuming each block of data is 1 kilobyte (KB) of data, after the first night's backup, instead of storing 24 KB of data, using data deduplication an organization would only be storing 18 KB of data which represents about a 25% savings in data storage.

Second Backup

In the second backup, the benefits of using data deduplication become more obvious. As Figure 2 illustrates, most of the data backed up the second time is exactly the same as the first backup. So during the data deduplication process, only these new unique chunks of data are stored while the other data is indexed but not stored.

A	B	C	D	S	F	A	G	Backup Stream #1
H	I	T	K	A	H	L	M	Backup Stream #2
N	O	I	A	P	D	Q	R	Backup Stream #3

Figure 2   New unique data blocks in the second backup are shows in **red** while redundant blocks of data found in the first backup are highlighted in **blue**.

Again assuming each block of data is 1 KB in size, instead of storing 48 KB of data as would occur if backing up this data without data deduplication, using data deduplication an organization only stores 20 KB of data. This represents about a 60% reduction in data storage versus keeping all of the data from both backups without data deduplication. Each subsequent backup will continue to improve this data deduplication ratio assuming there are minimal changes in production data.

This example illustrates that the largest variables in achieving large data deduplication ratios are the rate of data change and the type of backups that are run each night. This example assumes that each backup that an organization performs is a full backup (a backup where all data is backed up) and only a

few blocks of data change between each backup. The result is a very high deduplication rate with correspondingly high level of storage utilization in the backup data store.

The Data Deduplication Decision Grid

Despite the fact that all data deduplication methods share a common objective in terms of reducing the amount of data stored in the backup storage pool, the method of data deduplication that is implemented can have a substantial impact upon its success within an organization. It is because of this that organizations need to ask and answer a few basic questions before implementing data deduplication.

1. How can I know if data deduplication is the right choice for my organization?

2. What form or forms of data deduplication are the right methods for my organization?

3. What other factors or options do I need to consider?

Answering the "To Deduplicate or Not to Deduplicate" Question

The same factors (Call-out #1) that influence data deduplication ratios are also relevant as an organization makes a decision about whether or not to deduplicate data. These factors are:

*   An "incremental forever" backup policy that cannot be changed

*   All of the data created between backups is new and completely different than prior backups

*   Backups are only retained for a few days or weeks

*   The data change rate between backups is excessive (greater than 5%)

If any of these factors are true in your environment then it is probably NOT in the organization's best interest to deduplicate its backup data. However determining if the fourth factor is

Factors That Influence Data Deduplication Ratios

Organizations should not assume they will automatically achieve high data deduplication ratios when implementing a data deduplication solution. Three factors will influence the data deduplication ratio:

▶ **Type of Backup.** If an organization does incremental backups (a type of backup that only backs up data that is new or that has changed since the last backup), data deduplication ratios may remain low for some time and may never achieve the same high rates as when full backups are done.

▶ **High Data Change Rates.** If a large percentage of the data (greater than 5%) changes between each backup or a large amount of new data is created, the data deduplication rate and resulting space savings that organizations realize may be negatively impacted.

▶ **Backup Retention Periods.** Backup data that is retained only for short periods of time never gives organizations the opportunity to achieve high deduplication ratios

applicable to an organization's environment can be difficult to answer but it is the one that most organizations need to answer to determine whether or not to proceed with data deduplication. The data change rate has a large impact on the method of data deduplication that is selected but it can be the most difficult information for organizations to quickly and economically gather.

Ideally what an organization should do is use some sort of storage resource management software or backup reporting tool that can gather this information and then analyze it. If an organization lacks these tools, it can follow this more rudimentary approach to gather the needed metrics from their current backup environment to determine its rate of data change in its production environment.

This approach does not provide exact measurements but it can help organizations gather sufficient information to make a preliminary decision as to how to proceed and size a data deduplication solution.

1. Establish how much data was backed up in the most recent full backup by checking the backup software's logs or reports.

2. Quantify how much data is backed up in each incremental since that last full backup. This should give a company a sense of how much data is changing daily in their environment.

3. Gather the amount of data backed up as part of the full backups and incremental backups over a period of the last month. This will give the organization a sense of its daily rate of data change as well as estimate longer term data growth.

4. Find out how long full backups are retained.

For example, an organization backs up 2 TBs on its first full backup for the month. Each incremental backup that occurs the week after that full backup backs up about 80 GBs of data each night. This equals a daily data change rate of about 4%. Meanwhile the size of the weekly full backup grows by 25 GB each week so that by the end of the month the size of the full backup is now 2.1 TBs in size.

Since the organization's data change rate is under 5% and retaining full backups for a period of three years, an organization would likely benefit from implementing some form of data deduplication. Having made that decision, an organization can leverage its annual storage growth rate and rate of data change to select an appropriate data deduplication solution for their environment.

Deduplicating at the Source, Target, Media Server or in Archiving Software

Once an organization has the supporting evidence that it needs to justify its decision to move ahead with data deduplication, it now needs to decide which method or methods of data deduplication to implement. Each data deduplication method has distinctive advantages and drawbacks associated with its implementation that will influence when a specific solution should be implemented. However the benefits that each of these data deduplication methods share is faster backups and recoveries, more successful backups and restores

Source (or Client) Side Data Deduplication

How it works: Data deduplication software is installed on each application server. The software then analyzes and deduplicates data on the client or application server before sending unique blocks of deduplicated data to a central backup repository maintained by the backup software.

Source Side Data Deduplication Key Advantages and Considerations	
Key Advantages	**Key Considerations**
Can scale out data deduplication on other application servers by installing a backup agent	May need to install new backup software agent on application server
Can compliment and/or replace existing backup software	May require application server disruption (reboot) to install backup software agent
Reduces network traffic	Can be time consuming to implement
Heightened awareness of state of production data to support application consistent backups	May incur a substantial performance hit on application server on initial backup
Extremely fast backups in application servers with low rates of data change	
Can use any vendor's storage system to store deduplicated data	

Organizations that are good candidates for source side data deduplication will find one or more of the following as applicable to them:

• Application servers with ample CPU and memory and minimal data changes or small amounts of data to backup

• Large numbers of LAN attached application servers

• A virtualized server environment

• Remote and branch offices that require centralized data protection

- Implemented or are planning to implement a cloud-based backup scheme

How to Proceed: Organizations should view source side data deduplication as the ideal and preferred way to deduplicate data. Currently organizations should give preference to this approach when protecting application servers at remote and branch offices, LAN-attached servers and virtualized servers. The requirement to install and use the same backup software across all application servers coupled with the potential performance hit created by the initial backup may preclude some organizations from adopting this approach.

Target-Side Data Deduplication

How it works: Data deduplication is included with a storage system appliance. The existing backup software is reconfigured to send backup data to the appliance. The appliance deduplicates data in the backup stream and stores only unique blocks of deduplicated data.

Target Side Data Deduplication Key Advantages and Considerations	
Key Advantages	**Key Considerations**
Can keep existing backup software in place	No reduction in backup traffic sent over network between application server and target appliance
Data deduplication processing is performed by the target appliance	Ability of target side appliances to handle enterprise backup workloads still varies
Easy and quick to deploy since no need to touch or change agents on application servers	Variations in how data is deduplicated on the appliance (ingest, post-processing, hybrid)
Offer file system and/or virtual tape library (VTL) interfaces for LAN and SAN attached application servers	Need to appropriately size the capacity and performance of the appliance to match the backup environment into which it is deployed

Organizations that are good candidates for target side data deduplication will find one or more of the following as applicable to them:

- Application servers cannot handle the overhead associated with data deduplication
- Backup administrators are restricted from installing agents on backup servers
- Backup software that they cannot change
- Need to support multiple backup software products
- Backup situation that needs immediate resolution

How to Proceed: Organizations should view target-side data deduplication as the fastest and easiest fix to their

current backup problems. It minimizes the need for organizations to change and/or reconfigure their existing backup software and provides near-immediate relief for their current backup problems. However organizations should not assume that they have solved their backup problems long term as continuing or unexpected data growth can result in growing backup windows and capacity and/or performance shortfalls on the target appliance.

Media Server Data Deduplication

How it works: Data deduplication is included with the backup software and performed on the backup's software media server. The backup software agent on the application server sends backup data to the media server which deduplicates the data and then stores the unique blocks of deduplicated data to storage assigned to it

Media Server Data Deduplication Key Advantages and Considerations	
Key Advantages	**Key Considerations**
May be able to keep existing backup software	No reduction in network backup traffic between application server and media server
Data deduplication processing is performed by the backup media server	May need to upgrade media server's performance to handle data deduplication
May be no need to touch or change agents on application servers	To deduplicate data application servers must use that specific backup software
Can use any storage device to store deduplicated data	

Organizations that are good candidates for media server data deduplication will find one or more of the following as applicable to them:

- Application servers cannot handle the overhead associated with data deduplication
- Can standardize on a single backup software
- Have sufficient processing overhead on backup media server
- Want to have the option to store data on any vendor's storage device

How to Proceed: Organizations should use media server data deduplication when they want to use software to deduplicate data but are concerned about the impact that data deduplication would have on application servers. Organizations can introduce data deduplication using their existing backup software without putting the burden of data deduplication on

application servers. Organizations should not assume that they have solved their backup problems long term as continuing or unexpected data growth can result in growing backup windows and capacity and/or performance shortfalls on the media server.

Archiving Software Data Deduplication

How it works: Data deduplication functionality is included with email and file system archiving software. The archiving software agent on the application server moves aging or selected emails and files from the production data stores to the archived software's managed data stores. All emails and files in the archived data stores are deduplicated.

Archiving Software Data Deduplication Key Advantages and Considerations	
Key Advantages	**Key Considerations**
Infrequently accessed emails and files moved off of production data stores—reduces capital costs	Need to determine what policy should be implemented to invoke the archive—age based, quota based or both.
Data deduplication processing handled by archiving software after email/file archived	Need to set up retention and expiration policies for emails and files
Shortens backup windows and reduces backup data stores since fewer emails/files to backup	Need to synchronize backup policies with retention policies to ensure they are adhering to established corporate policies
Can improve application server performance	User experience when retrieving archived data is different
Policy-based archiving rules reduces overhead for application server and archiving system	Creating policies requires the formation of cross-functional business and technical teams

Organizations that are good candidates for archiving software data deduplication will find one or more of the following as applicable to them:

- Large numbers of aging emails and files (have hundreds of thousands or millions)

- Have someone who can set up and manage archiving software

How to Proceed: Organizations should use archiving software data deduplication in environments where large amounts of email and file system data exist. Using archiving software data deduplication organizations can attack the source of where backup problems often begin—email, file and Microsoft SharePoint data stores are the most obvious places for organizations to start. The benefits can be and are substantial but organizations should only proceed with archiving after they have carefully considered how they will configure and set up the policies.

The Need for an Agnostic Approach to Managing Data Deduplication

These summaries of the four general approaches currently available for implementing data deduplication illustrate why enterprise organizations have an interest in each one. It also explains why these same organizations may find any one approach insufficient for all of their data deduplication requirements and may elect to use one or more of them within their enterprise.

The problem that enterprises therefore encounter is not a shortage of data deduplication techniques. The larger issue is, "How to manage these various data deduplication techniques?"

Enterprise organizations have a need for all of these data deduplication techniques because rarely will all of their data deduplication requirements fit neatly into one bucket. Instead they need to keep their options open for introducing these different data deduplication techniques into their environment. But to do so, they also need to have a mechanism to accommodate and manage these different forms of data deduplication.

Leveraging Symantec for Optimal Data Deduplication Results

Enterprise organizations recognize that they may have the need to implement and support data deduplication across their IT infrastructure in any of the ways described here. Specific application constraints, budgetary limitations and evolving data protection requirements create the need for these size organizations to have multiple data deduplication techniques at their fingertips for these different scenarios.

Leveraging Symantec's portfolio of information management products enterprise organizations do not need to bet on only one data deduplication approach but can select from any of these data deduplication method and choose a technique that bests fits their needs. For example:

- Symantec NetBackup PureDisk supports source side data deduplication so organizations address their data protection needs in their remote and branch offices, LAN-attached servers and in virtualized server environments.

- Both Symantec NetBackup and Backup Exec natively interact and support target based data deduplication appliances. However using Symantec's OpenStorage API (OST) available shortly in both products, organizations can manage replication between different target based data deduplication appliances, keep their backup software catalogs updated and improve backup and recovery speeds.

- Both Symantec NetBackup and Backup Exec support media server data deduplication on their respective

media servers so organizations can offload the data deduplication to the media server and store dedupli-cated data to any storage device.

- Symantec Enterprise Vault, its email and file archiving software, natively supports data deduplication and integrates with NetBackup to provide a centralized control panel.

Understanding the different forms of data deduplication and best methods in which to implement it is already confusing. Using Symantec's suite of information management products that provides solutions to address each of these needs, organizations can stop worrying about betting the farm on one form of data deduplication. Instead they can turn their attention to identifying the form or forms of data deduplica-tion that best meets the specific needs of their different appli-cations or business units within their organization and leave Symantec with the responsibility of making all of these differ-ent forms of data deduplication work nicely together.