WHITEPAPER

datalink
An Insight company

# Using Deduplication: 5 Steps to Backup Efficiencies

BY MICHAEL SPINDLER, PRACTICE MANAGER, DATALINK

## ABSTRACT

Deduplication technologies are providing dramatic benefits as organizations seek to optimize their backup environments. This whitepaper provides an overview of some of the key factors to assess in selecting and deploying a deduplication solution. Specifically, the paper details five steps for enabling organizations to sort through the myriad of deduplication options and effectively implement a solution.

# Table of Contents

# The Challenge

## THE CHALLENGE

The rapid and endless growth of business data is a fact of enterprise life. In the past, fully protecting that data meant that backup storage costs increased in lockstep with primary storage. The only alternative to paying ever-higher backup costs was to protect only subsets of corporate data stores. For most companies, trading high costs for high risk was not a good option. Over time, technology advances—faster tape drives, lower-cost disk and network solutions—have helped slow the escalation of backup and replication costs. But none of these options improve on the basic 1:1 ratio of primary to backup capacity—that is, backing up 1TB still requires 1TB of disk or tape storage, even if it's a backup of an email system with dozens of instances of an identical attachment or a virtual desktop environment with hundreds of identical operating system instances. The ratio only worsens with repeated full backups.

Compression of backup data is only a partial answer to this problem. Tape devices or backup software at the client can compress some types of data well but can actually expand the final size stored for other types of data. This process uses different algorithms. Lempel Ziv (LZ), for example, is a commonly-used algorithm. Since there is no such thing as "typical data," one can see a reduction of 10-60 percent, or a 1.1:1 to 3:1 net space savings.

Deduplication technologies change the equation. Deduplication identifies redundant information and stores it efficiently while maintaining the integrity of the original content. Data is stored once, no matter how many copies are made.

Deduplication helps break the lockstep connection between primary storage and backup costs. By greatly reducing the storage capacity required for backups, deduplication can help businesses retain more data on disk (and enable faster recoverability than from tape), affordably protect more data sets, and simplify and reduce the costs of disaster recovery (DR) via optimized replication. It also provides the ability to use less bandwidth, reduce power consumption, and simplify administration.

With deduplication technologies proven in real-world applications, businesses face minimal risk in applying a well-architected solution. Experience suggests that by taking the five steps outlined in this paper—1) know the lingo; 2) know the environment; 3) choose the right product; 4) know your team; and 5) use deduplication effectively—IT managers can successfully sort through the myriad of vendor offerings and ideally leverage deduplication to achieve backup efficiencies with maximum protection of corporate data assets.

**A word on deduplication for primary storage**

Most deduplication solutions originally targeted backup devices—an obvious target because of the amount of duplicate data typically stored in backups. But vendors now offer deduplication solutions for primary or Tier 1 storage as well. There are many variables to consider when evaluating the functionality and the potential benefits of primary storage deduplication solutions. However, the focus of this whitepaper is successful utilization of deduplication technologies in the backup process managed by the backup appliance or software.

**Deduplication helps break the lockstep connection between primary storage and backup costs.**

# Step 1: Know the Lingo

## STEP 1: KNOW THE LINGO

Vendors refer to deduplication functionality using a variety of terms, including dedupe, data reduction, single instance storage, global data single instance storage, capacity optimized storage, and even molecular sequence reduction. The basic objective of each implementation is the same: find duplicate data sets and store just one copy. But it can be helpful to understand some of the differences in vendor implementations. Although not meant to be an exhaustive study of methodologies, the following discussion of how, where, and when deduplication is accomplished can be helpful in your evaluation of vendor offerings.

### How they do it

The most common techniques for performing deduplication are hashing and delta differencing. Backup appliance vendors use one or the other, or in some cases, a hybrid of the two approaches. Here are the basics of each:

***Hashing –*** In this methodology, deduplication engines view data at either a block (subfile) or file level. Data is broken down into smaller blocks or segments that are given unique identifiers created by hashing algorithms like MD5 and SHA-1. Some vendors also use content-aware logic that considers the source of the data (for example, which backup software is sending the data stream) to determine block sizes and the boundaries of the resulting blocks. Hashing is more widely used than delta differencing and has been proven over time. By comparing the hash id(s) of each block of data regardless of what backup job is sending it, hashing typically deduplicates more data in dissimilar datasets (for example, test and production) than delta differencing, but the process requires greater CPU performance. There is also the mathematical possibility of a hashing collision—that is, when the hash value is the same for two different blocks of data. But with accepted estimates putting the odds at 1 in $10^{15}$, a collision event is unlikely.

***Delta differencing –*** This approach focuses on post-backup data deduplication and uses a higher level of abstraction in the backup data analysis. In contrast to hashing comparisons that look for redundancies in byte streams, delta differencing compares objects to objects—for example, Microsoft® Word document to Word document or Oracle® database instance to Oracle database instance. Delta changes are stored in the meta database of the deduplication appliance. This method is more efficient than hashing but is dependent on awareness of the specific backup application, backup client, and backup data set.

Delta differencing can deliver better deduplication; there are no hashing collisions and the process utilizes less CPU. The drawback is that deduplication appliance vendors must deliver solutions for each of the different data types across backup software products. Additionally, the delta differencing of "file123" on client1 is not compared to "file123" on client2, which can negatively impact the efficiencies of dedupe. For example, a delta differencing process might compare and store the changes between the current and previous night's RMAN DBSRV01 backup of an Oracle Database instance, but it will not compare another backup of the test version of this database backup on another server.

## STEP 1: KNOW THE LINGO (CONTINUED)

**Where they do it**

***Target deduplication solutions*** take an existing backup created by any backup application and deduplicate the datasets. The downside of target deduplication is that although it reduces the amount of capacity required to store the backup, it does not reduce the bandwidth required to copy the original data to the backup server.

***Source deduplication*** requires the use of deduplication-aware backup software—that is, the backup product works in conjunction with the deduplication software or appliance to identify duplicate sets and prevent transmission of redundant data to the backup target. The downside of source deduplication is that backups can consume more CPU cycles and take longer to complete than traditional backups.

**When they do it**

***Inline deduplication*** processes deduplicate backup data in real time as it's received at the front end of the virtual tape library (VTL) or disk-to-disk (D2D) device. Because the process is highly CPU- and I/O-intensive, solutions are typically built on dual, quad-core processors and/or high-speed disk components.

***Post-process*** methods deduplicate after the backup has completed. Since backups occur before deduplication, there is less "at risk" time during which a backup has not yet completed. These solutions, however, requires additional disk space to hold the backup before it is deduplicated. Implementing these solutions require sizing the landing space to accommodate not just the space required for one backup set, but potentially the next one as well, because if the ingest or backup speed is very high, the deduplication process might not complete before the next backup starts.

Bear in mind that most deduplication solutions do not fit neatly into these categories and are often hybrid solutions. Understanding basic functions and terminology can help you better consider the benefits and tradeoffs of each solution.

**How do they provide disaster recovery?**

Often overlooked, replication of deduplicated data offers significant cost, labor, and time efficiencies/savings. For critical applications, organizations may already be using replication to ensure the data is available in a remote site in case of problems. Deduplication appliances offer similar abilities for backup data of non-mission critical applications where physical tape is utilized.

After that first backup, the net amount of data stored in an appliance is small since it is only meta data and new blocks of data. By replicating this new information from your data center (the source) to a remote site (the target) the data is available for disk based restores in your DR site. Another common use is for remote sites that are smaller. Leveraging replication of that data to the main site is an efficient use of communication lines (only new blocks and meta data reduce the amount of traffic). It also simplifies operation in these remote sites where you may have relied on non-IT personnel to manage tapes in the past.

**Often overlooked, replication of deduplicated data offers significant cost, labor, and time efficiencies/savings.**

# Step 2: Know Your Environment

## STEP 2: KNOW YOUR ENVIRONMENT

Equally important to understanding the technology is knowing your own data and storage environment. Early in the process of choosing a deduplication solution, you should take information-gathering and objective-setting steps to:

1) Assess, audit, and discover what you have by answering questions such as:
    a. How much and what kind of data does my organization store?
    b. How much do we need to back up?
    c. How much does our data change?
    d. Where are we storing backups now, how much capacity is required, and is it disk or tape?
    e. What are our tape backup processes?
    f. What are our archiving and information lifecycle management requirements?
    g. What are the shortcomings and issues with our current processes and/or products?
    h. How long do we retain data? Do we have different retention periods?
    i. What is our access to day-to-day restores? What is a typical timeframe for those "accidentally" deleted file restores?

2) Identify what you most need to accomplish. For example:
    a. Accelerate backups
    b. Back up more of the environment
    c. Archive data that doesn't change and remove it from the normal backup regiment
    d. Reduce tape storage and handling costs
    e. Conserve bandwidth to reduce cost of off-site replication/DR

3) Determine reasonable expectations for deduplication benefits. For example:
    a. Use x% less backup capacity
    b. Replicate y% less data to your DR site
    c. Reduce backup administration costs z% by automating processes

Results obviously vary for each IT environment and product implementation, but the better description of your data environment and business expectations that you can bring to the vendor and/or your solutions integrator, the better you'll be able to accomplish your desired outcome.

The level of deduplication or ratio will be most affected by the retention of data and the redundancy of data across the backup environment. For example, a Datalink client in the healthcare industry used deduplication appliances and achieved an 8:1 data reduction ratio retaining backups for one month. Another client achieved a 14:1 ratio with only 10 days retention. Still, another client from within the same industry achieved 11:1 data reduction space saving with two months retention.

**Equally important to understanding the technology is knowing your own data and storage environment**

# Step 3: Choose the Right Product

## STEP 3: CHOOSE THE RIGHT PRODUCT

The industry offers many choices for where to get your deduplication functionality, including: disk-based, purpose-built deduplication appliances; VTL appliances; enterprise backup software; and general-purpose network-attached storage (NAS) solutions. As one might expect, most options are not one-size-fits-all solutions. Organizations need to consider each alternative in light of their requirements for short-term backup, data protection/recoverability, and long-term backup and archive. In addition the performance of the product is a factor. It's imperative to weigh details such as whether the product will accomplish backups in the allocated backup window and whether the current infrastructure will adequately support the solution. The ideal solution for one type of backup may not be the best fit for another.

There are several factors involved in determining the capacity required:

- Post process dedupe vs inline dedupe—Post process requires additional space to temporarily hold the data before processing.
- Use of hot backup modules in backup software—Many appliances can recognize a database or mail server backup stream, which ultimately helps optimize deduplication.
- Retention of data—Generally longer retention of data yields higher deduplication
- Backup of compressed data—Usually this yields little deduplication.
- Network infrastructure—Will trunking of 1Gb connection provide the needed throughput? Will 10Gb be needed?
- Backup stream count—Will the solution be able to accommodate an adequate number of data streams to meet the backup window?
- VTL or NAS—Infrastructure and/or capacity can be affected.

NAS-based deduplication products and technologies have sweet spots for various sized companies. For example, some solutions are geared to and most applicable for small to medium-sized businesses that manage less than 5TB of data and need affordable replication and simplicity. This type of solution could also be a good fit for larger businesses deploying solutions for remote offices. Most vendors offer solutions that can start at 1 or 2TB of usable capacity and expand to about 8-10TB.

Other deduplication products better address the needs of mid-range to enterprise-size businesses that need usable capacities starting at 8-10TB and grow to about 60TB. These businesses more often place a premium on performance, scalable capacity, efficiencies, and replication. The products for this segment of the market provide higher performance compared to the previous segment and 10Gb connection is optional or standard.

At the high end are enterprise customers that require in excess of 60TB of usable capacity today or to accommodate growth in the future. They require multiple deduplication appliances, clustered solutions, global deduplication functionality, and resources for navigating complexity of this segment.

## STEP 3: CHOOSE THE RIGHT PRODUCT (CONTINUED)

The good news is that there are many vendors and products from which to choose. That variety is also the bad news. Here are some questions that can help you narrow the choices from a very large field of suppliers and products:

- What differentiates your deduplication solution, algorithms, and methodologies? How will those features and functions apply in my environment?
- What should we expect for deduplication efficiencies? And on what data sets should we expect the best return? Where should we NOT use it?
- What happens if we run out of capacity or performance?
- Do you offer global deduplication?
- How do we recover deduplicated data? Will there be a performance impact?
- How much can we automate processes? Do you support policy-based deduplication?
- What management and reporting tools do you offer?
- What training will our administrators need?
- How will implementing your deduplication solution impact my current backup processes?
- How does your solution integrate with or complement my existing backup software and devices?
- What does it take to configure your solution with my current storage and networking infrastructure?
- Do you provide a complete solution and full support for both hardware and software elements?
- Do we need installation assistance?
- How long have you offered your deduplication solution?
- Do you offer any capacity savings guarantees?
- What is your licensing structure? Does that license include all features?

If you need help at any point in the process, most vendors offer professional services to help you evaluate and size solutions from within their product offerings. Independent data center infrastructure and service providers offer the advantages of unbiased and vendor-neutral assessment processes, as well as real-world experience and depth of knowledge across multiple vendors and product lines.

**The good news is that there are many vendors and products from which to choose. That variety is also the bad news.**

# Step 4: Know Your Team
# Step 5: Use Deduplication Effectively

## STEP 4: KNOW YOUR TEAM

Being brutally honest about your in-house resources helps ensure the best return on your deduplication technology investment. Here are important realities to address:

- Has your IT staff been trained on both the technology and the products? Where do they need more?
- Can they apply best practices—that is, do they have the knowledge, expertise, and time?
- Where could you benefit from help? How would it accelerate your time to results?
- What education/training services are available for the technology, the product, and best practices? Can you arrange to talk to other companies that have used these services?

## STEP 5: USE DEDUPLICATION EFFECTIVELY

Once you've determined the best-fit solution and brought your staff up to speed, you can put the deduplication solution to work. But be aware that most solutions are not "set-it-and-forget-it" deployments. After you turn on functionality, you need to circle back to assess how it's working and then fine tune your processes accordingly. Following are some tips to maximize efficiency without compromising data protection and recoverability:

- Determine where deduplication is working most effectively. Results can vary depending on both the deduplication technology and the dataset, but typically you could expect to see the highest percentage capacity savings on datasets that include: Microsoft Office PowerPoint, Excel, and Word documents; email attachments; Oracle Database and Microsoft SQL Server backups; and in virtual server and virtual desktop environments. In contrast, deduplicating archive log information, images, and other rich media files will likely produce less capacity savings.
- Make sure that current processes are allowing you to meet your recovery time and point objectives (RTOs/RPOs). Are you getting backups done in available windows, are all critical applications and data adequately protected?
- Decide if you need to modify where and when deduplication is turned on and whether or not it's set to run automatically. How is it impacting your environment right now?
- Continue to do backup reporting. Not all backup and deduplication solutions offer adequate reporting functionality. However, good reporting tools are essential in helping you ensure adequate protection, control costs, accommodate IT and business requirements like chargebacks, and fully leverage backup solution and deduplication capabilities. If your solution lacks robust functionality, consider third-party products or Software-as-a-Service (SaaS) reporting capability from providers like Datalink.
- Assess the impact on your in-house IT resources. Does your team have adequate cycles to meet business demand for services? Is there value for you in managed services? Datalink, for example, offers a monitoring and alerting service for Symantec® NetBackup.™ Vendors and IT services providers can help supplement your IT resources to save time, extend technical functionality, and enhance the overall value of your deduplication solution.

# Benefits for the Taking

## BENEFITS FOR THE TAKING

Done right, deduplication can help you accelerate backup processes, conserve bandwidth, minimize risk, protect more of your information assets, and dramatically reduce your storage, maintenance, and administration costs. Although the process of choosing and deploying a deduplication solution can have its share of complexities, the viability of the technology has been proven in a wide array of real-world business environments and IT infrastructure settings. Leveraging our expertise from these real world implementations, Datalink can help you identify the best combination of products for your business environment and needs, effectively deploy the technology, and manage it for maximum efficiency. We can get you to the point where the proper preparation, product(s), and services can virtually eliminate risk and ensure your success with deduplication solutions.

Dozens of organizations, from mid-tier enterprises to Fortune 500 corporations, have trusted Datalink with their IT initiatives. We have the extensive knowledge and experience to guide you through development of a data center deduplication strategy, and then navigate the organizational and technical challenges of implementing it.

**Partnership with Datalink**

*A complete data center solutions and services provider for Fortune 500 and mid-tier enterprises, Datalink transforms data centers so they become more efficient, manageable and responsive to changing business needs. Datalink helps leverage and protect storage, server, and network investments with a focus on long-term value, offering a full lifecycle of services, from consulting and design to implementation, management and support. Datalink solutions span virtualization and consolidation, data storage and protection, advanced networks, and business continuity. Each delivers measurable performance gains and maximizes the business value of IT. To learn more about how Datalink can help your organization use deduplication technologies to improve the overall efficiency of your data center and deliver dramatic ROI to your organization, contact Datalink at (800) 448-6314 or visit www. datalink.com.*

**To receive the latest white papers and insight into data center technologies and practices, follow Datalink online at the sites below.**

**http://twitter.com/datalinkcorp**

**http://blog.datalink.com/**

**http://www.facebook.com/datalinkcorp**





datalink

An Insight company