Software

Evaluating Deduplication Solutions: What You Really Should Consider

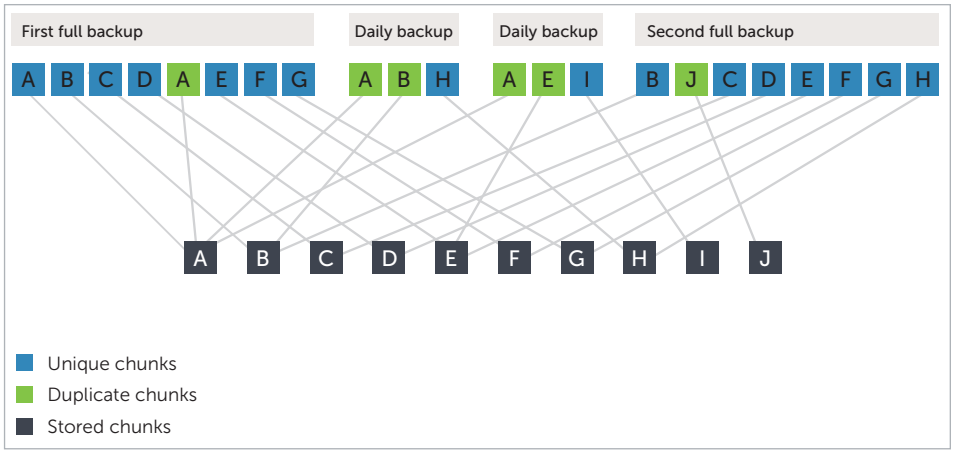
Introduction

Reducing backup and recovery costs, enhancing data backup retention and protection, and improving disaster recovery can be easily achieved through data deduplication. To make the best decision among the many deduplication solutions available, you need to not only understand what deduplication is, but more importantly, the various ways that it is implemented and what each of those might mean for your environment. This paper explains what elements you should consider and how the right decision will make your life easier by reducing the amount of storage you need to manage, maximizing your flexibility, and shrinking your backup windows.

What is deduplication?

The Storage Networking Industry Association (SNIA) defines deduplication as the process of examining a data set or byte stream at the sub-file level, and storing and/or sending only unique data. With deduplication you are only storing the unique data because duplicate data is replaced with a pointer to the first occurrence of the data.

Deduplication carves data into chunks. These chunks are hashed and the resulting unique identifier is compared to an index of all unique identifiers. If the identifier already exists, the data is a duplicate and is replaced with a pointer to the correct unique identifier in the index. If your backup is 100GB and 20GB is redundant or duplicate data, for example, you would only be storing the unique 80GB on disk. When the unique data is then optionally compressed, you can experience even more substantial disk space savings.



The key difference between SIS and deduplication is that SIS evaluates data at the file level, while deduplication evaluates data at the block or byte level.

Single instance store vs. deduplication

SNIA defines single instance store (SIS) as the replacement of duplicate files or objects with a reference to a shared copy. While SIS is not defined as a deduplication technology by SNIA, many people confuse it with deduplication.

The key difference between SIS and deduplication is that SIS evaluates the data stream at the file level by looking for duplicate files, while deduplication evaluates the data stream at the block or byte level. With SIS, even a small change to a file will cause it to be seen as a new and different file to be stored as a separate instance. This means, that if you have a document and a user copies or renames the file, it will be seen as an entirely different file. With deduplication, the entire file would be detected as duplicate. As a result SIS delivers less space savings versus deduplication.

What difference does the deduplication block size make?

Deduplication can occur at two different levels: block and byte. Block-level deduplication looks at entire blocks of data, while byte-level deduplication performs a byte-by-byte comparison of the data to the index to detect duplicates. For this reason, byte-level deduplication is the only method that can guarantee full elimination of all redundant data. Nevertheless, block-level deduplication is far more common because it delivers acceptable results with far less overhead.

There are two prevailing methods for block-level deduplication: fixed-length and variable-length. With smaller fixed-block sizes, the same data stream will be divided into more chunks than if a larger block size is used. This leads to a higher percentage of duplicates being identified. With variable-block length, the deduplication engine has the ability to change the block size and recognize more duplicate patterns, thereby increasing the number of duplicate blocks.

Consider a file that has had only minor changes made to it, such as adding a word to a sentence or removing a word. With fixed-block deduplication, the rest of the file remains the same if the blocks do not line up exactly, so they will be seen as unique and will not be deduplicated. With variable-block deduplication, the deduplication engine can isolate only the changed data and deduplicate the rest of the blocks, resulting in more storage savings.

In the example shown in Figure 1 where some of the words in the file were changed, you can see how fixed-block deduplication will identify more blocks as being unique, whereas variable-block deduplication identifies more blocks as being duplicates.

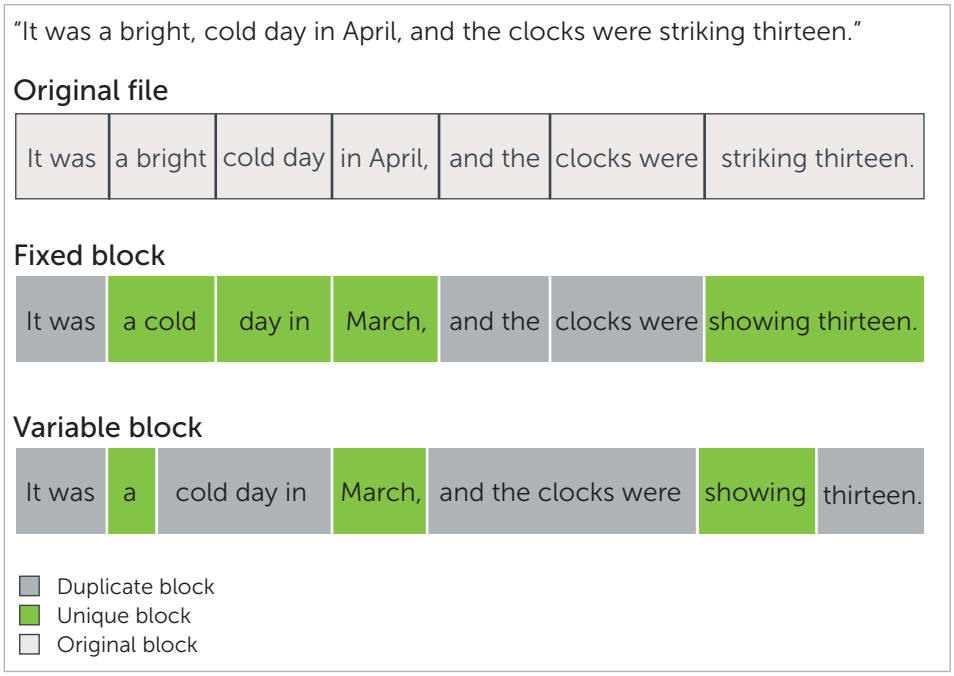


Figure 1: Comparison of fixed-block and variable block deduplication.

Inline or post-process deduplication and why it matters to you

Another technology differentiator to consider is when the deduplication processing occurs. Deduplication typically is performed either inline or post-process. Each has advantages and trade-offs.

With inline deduplication, data is deduplicated as the deduplication engine receives it. Once the data is deduplicated, it is then immediately stored on disk. The advantage of inline deduplication is that it does not require any additional disk space to store the data temporarily prior to deduplication. However, inline deduplication has the following trade-offs:

- Depending on the capabilities of the system, performing the deduplication process as part of the backup could potentially lengthen the time to complete the backup. The better systems mitigate this issue by performing deduplication in hardware.
- Restores are faster if performed soon after the backup is completed. With inline deduplication, however, if the data is immediately deduplicated, every restore might require rehydration or reassembly of the unique chunks into the original data stream.

With post-process deduplication, the backup is temporarily placed on a disk-based staging area prior to the deduplication process. Some deduplication technologies require the entire backup to be staged before deduplication begins, while others allow deduplication to begin after a set amount of the data stream has been staged. The latter approach reduces the sizing requirements for the staging area while also allowing the backups to complete as fast as possible. Post-process deduplication has the following advantages:

- Because deduplication is not part of the backup with post-process deduplication, backups can complete faster, which shrinks your backup window.
- Post-process deduplication allows users to provision data on existing storage systems rather than on a separate backup and recovery appliance.
- Restores are faster if performed soon after the backup is completed because the data

With post-process deduplication, the only real trade-off is that it requires additional disk space for the staging area.

has not been deduplicated yet and will not need to be rehydrated to perform the restore. Because restores typically come from the most recent backups, this enables you to speed up restores while still taking advantage of deduplication to reduce long-term storage costs.

With post-process deduplication, the only real trade-off is that it requires additional disk space for the staging area. The size of the staging area will depend upon the approach used, which determines how long the data will need to remain in the staging area awaiting deduplication, as well as upon how much data will not be deduplicated.

Source-side deduplication is better for remote sites with smaller amounts of data, while target-side deduplication fits better into environments with larger amounts of data.

What is the difference between source and target-side deduplication?

In addition to looking at when the deduplication occurs, you should consider where deduplication is performed. There are two places where deduplication occurs: either on the source/client or on the target/storage.

Source-side deduplication typically uses a deduplication engine that is located on the client that hashes the data and checks for duplicates with a centrally-located deduplication index, which is typically located on the backup server or a media server. The advantage of source-side deduplication is that it reduces network contention because no duplicate data is sent over the network.

While in some scenarios this can be beneficial, you also have to look at the drawbacks of this type of deduplication. By running source-side deduplication, you are adding a processor-intensive hashing algorithm to your clients, which could slow down the backups and lengthen the backup window.

Target-side deduplication relieves the clients from performing additional work by running the deduplication entirely at the target/storage. The advantage here is that you do not need to worry about having clients with enough processing power to handle the hashing algorithm. The trade-off is that more data must be sent over the network, so if you have a network that is already fully utilized, this method will cause periods of congestion.

The other factor to consider with source- vs. target-side deduplication is that target-side deduplication is better suited for environments with higher volumes of data. Many industry experts contend that source-side deduplication is better for remote sites with smaller amounts of data to be deduplicated, while target-side deduplication fits better into environments where there are larger amounts of data requiring deduplication.

Different vendors have created different solutions that may mix and match the when and where of deduplication. For example, you may have a solution that does inline deduplication starting at the source, while others may do post-processing deduplication starting at the target. When evaluating deduplication solutions that best meet your needs, be sure to not only consider when the deduplication process is happening, but also where it is occurring.

What do deduplication ratios really mean?

Deduplication is typically reported as a ratio or a factor; for example 12:1 or 12X, respectively, which mean the same thing. The ratio is calculated as bytes in to bytes out, or can be viewed as the data capacity of a system divided by its used storage capacity. If 500GB of data only consumes 50GB of storage, for example, the deduplication ratio is 10:1.

Because each vendor in the deduplication market has its own test environment, comparisons become problematic, as each test uses a different set of data and a different set of assumptions. When dealing with deduplication ratios you have to

Deduplication Ratio	Space Reduction Percentage
2:1	1/2 = 50%
5:1	4/5 = 80%
10:1	9/10 = 90%
12:1	11/12 = 91.67%
15:1	14/15 = 93.33%
20:1	19/20 = 95%
30:1	29/30 = 96.67%

Figure 2: The effect of various deduplication ratios on storage space reduction

understand that their significance is based on certain factors:

- Ratios are only meaningful if they are compared when using the same set of data and assumptions
- Even low deduplication ratios provide significant space savings
- Higher ratios yield marginally less space reduction

As you can see in Figure 2, once you hit the 10:1 deduplication ratio, the actual amount of disk savings does not increase significantly. If one vendor claims 20:1 ratio and the other claims 12:1 ratio, the two numbers may sound drastically different, but when you look at the actual amount of disk savings, you find that there is less than a 5% difference. This is why deduplication products must be evaluated on factors other than deduplication ratios.

How long should I retain deduplicated data?

A final issue to be considered when evaluating deduplication technologies is deciding how long you wish to retain data. The more data that is examined, the greater the likelihood that duplicate data will be found, which will increase your disk space savings.

For example, as shown in Figure 3, the initial full backup that you deduplicate will only be deduplicated against itself and will result in a relatively small amount of reduction in the storage footprint, depending upon the type of data that is being deduplicated. When the full backup for week two is performed, only the unique data that has been updated or added since week one will be stored. When the full backup for week four is performed, the chunks will be compared against all the unique data for weeks one, two and three, which increases the chances that duplicate chunks will be found.

When you deduplicate your backups, each additional week of backups can be retained for a decreasing amount of additional disk space. This allows you to store even more backups on the same amount of disk-based storage for a longer period, and can virtually eliminate the need to restore from offsite storage unless there is complete site failure.

What should I consider in my deduplication decision?

Your goals for your deduplication solution will influence which deduplication technologies you should evaluate. Here are some typical deduplication solution goals and considerations.

Maximum disk space savings

- Deduplication offers more disk space savings than SIS.
- Variable-block deduplication typically provides better deduplication ratios than fixed-block deduplication.

After you hit the 10:1 deduplication ratio, the actual amount of disk savings does not increase significantly.

NetVault SmartDisk is hardware-agnostic, enabling it to work on most file systems without specific drives or expensive appliances.

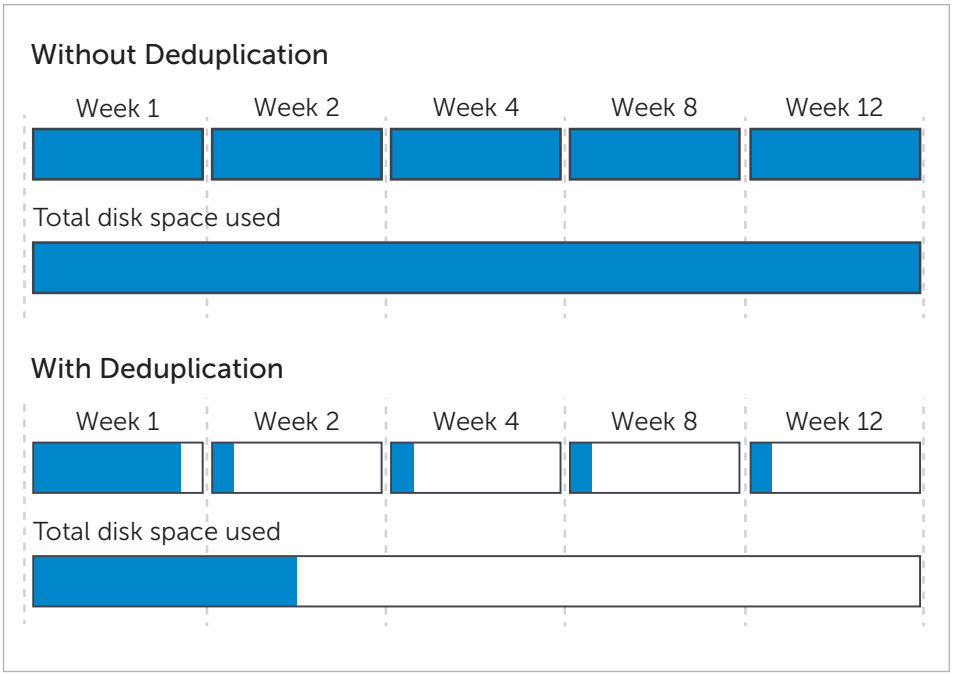


Figure 3: The accumulative effect of deduplication over time.

- Retaining deduplicated data longer will allow you to store even more backups on the same amount of disk-based storage for a longer period.
- Near inline or inline deduplication reduces disk space requirements because it does not require a large staging area or a staging area at all.
- Source-side deduplication can reduce network traffic.

Maximum flexibility

- Variable-block deduplication enables the technology to adjust to the data stream to find more duplicates.
- The ability to leave data that does not deduplicate well in a non-deduplicated state ensures that you are not using valuable time and processing power on data that will not benefit from deduplication.
- With post-process deduplication, restores are faster when performed soon after the backup is completed because the data has not been deduplicated yet and will not need to be rehydrated to perform the restore. Because restores typically come from the most recent backups, this provides the flexibility to speed up restores while still taking advantage of deduplication to reduce long-term storage costs.
- Post-process deduplication allows users to provision data on existing storage systems rather than on a separate backup and recovery appliance.

Shrink backup windows

- Post-process deduplication is not part of the actual backup and can be scheduled to occur outside of the backup window.
- With some target-side deduplication solutions, deduplication can also be performed independently of the backup job to ensure deduplication does not elongate backup windows.

NetVault™ SmartDisk

While there are many deduplication vendors in the market, Dell was the first vendor to offer an Open Data Protection Platform (ODP) that extends data protection environments with an affordable and long-term, disk-based storage solution that can remain in place regardless of who or where the data is coming from. This gives you the opportunity to consolidate your data protection solutions using the same deduplication solution, thereby reducing

costs and providing more flexibility with regard to vendors.

NetVault SmartDisk seamlessly integrates with NetVault Backup and vRanger®, allowing you to use NetVault SmartDisk for disk-based backup and deduplication needs within an enterprise-class, easy-to-use data protection suite. NetVault Backup and NetVault SmartDisk are compatible with SAN and NAS systems, some of which are supported with custom extensions, such as the DD Boost agent for seamless integration with EMC's Data Domain storage system.

NetVault SmartDisk's deduplication uses byte-level deduplication with a variable block size. The deduplication occurs post-process so that it does not interfere with the backup window and allows you to schedule the deduplication process to occur when you need it. NetVault SmartDisk also facilitates near-line deduplication by allowing the deduplication process to start before the entire backup is completed.

Because not all data deduplicates well, NetVault SmartDisk's integration with NetVault Backup offers job-level deduplication allowing you to create different selection sets. You can select which data you do and do not want to deduplicate to ensure you are not deduplicating data that does not lend itself to deduplication, such as encrypted files. The ability to create these different selection sets allows you to create the best data protection scenario for your environment and get the most benefit from the deduplication solution without having to waste valuable space and processing time on data that benefits little from deduplication.

To provide you with a solution that can be integrated into your environment, NetVault SmartDisk is hardware-agnostic, enabling it to work on most file systems. This means there is no need to acquire specific drives or expensive appliances in order to create a data protection strategy that includes deduplication.

NetVault SmartDisk also offers the ability to easily add additional file system paths to NetVault SmartDisk Storage Pools, reducing costs by deferring storage expenditures into new budget periods and ensuring that storage does not sit unused. In addition NetVault SmartDisk supports the most popular operating system platforms in the market.

NetVault SmartDisk was designed from the ground up to give you more choices and maximize your investment. With NetVault SmartDisk, you can deploy multiple NetVault SmartDisk instances to improve load balancing and performance, as well as place the deduplication solution where you will see the greatest benefit, such as in remote sites or at DR sites. To further facilitate disaster recovery preparedness, you can configure NetVault Backup to replicate deduplicated backups among SmartDisk instances, and optionally encrypt secondary copies used for off-site storage, or as required for regulatory compliance.

NetVault has a reputation for being both enterprise-class and easy to use. With NetVault SmartDisk, Dell has continued that tradition by providing a deduplication solution that reduces the level of storage expertise required to perform deduplicated disk-based backups, allowing you to spend more time on your many other job requirements.

Dell DR4100 backup and recovery appliance

In addition to software-based deduplication with NetVault SmartDisk, Dell offers the DR4100 backup and recovery appliance. The DR4100 supports both deduplication and compression, and is compatible with NetVault Backup, vRanger and many other backup solutions. By removing redundant data inline from the backup work stream and compressing the results, the DR4100 minimizes the storage footprint, enables backups to remain on disk and online longer,

The DR4100 supports both deduplication and compression, and is compatible with NetVault Backup, vRanger and many other backup solutions.

The Dell DR4100 delivers data reduction and protection in a single, turnkey appliance that helps you meet stringent RTO/RPO and DR objectives.

provides faster and more reliable restores, and reduces tape management complexity.

The high-performance, disk-based DR4100 backup and recovery appliance is easy to deploy and manage, and offers a low total cost of ownership. The DR4100 is a 2U-high, rackmountable appliance available in several capacity configurations—making it ideal for small enterprise, remote office and multi-site environments. The system supports both 1GbE and 10GbE interfaces, and the 12 integral disk drives are protected against individual failures in a RAID 6 configuration.

Through the use of innovative deduplication and compression technology, the DR4100 can help you achieve data-reduction levels up to 15:1. This reduction in data enables you to retain more backup data for longer periods in the same footprint. Shorter recovery time objectives (RTO) and more attainable recovery point objectives (RPO) can also be assured as critical backup data remains on disk and online longer.

As a target-based deduplication appliance, the DR4100 is specifically engineered to handle streaming backup workloads. And because all dedupe/compression operations are performed at the disk backup target, there is no adverse impact on backup and recovery performance. By replicating only deduplicated data, both network bandwidth and disaster recovery times are reduced significantly. Replication is normally configured to occur during non-peak periods, so as not to interfere with other applications, including scheduled backups.

An intuitive graphical user interface (GUI) gives you an overview of the system, including status, hardware and software alerts, storage capacity and savings, and other important information, such as system and software versions. The DR4100 automatically monitors the health of the hardware and continuously verifies the integrity of the system software, and critical alerts can be sent by email and SNMP traps for immediate notification.

Part of the Dell Fluid Data architecture, the Dell DR4100 delivers data reduction and protection in a single, turnkey appliance that changes the economics of disk-based data protection. By accelerating and streamlining the backup, replication and recovery processes, the DR4100 helps you meet stringent RTO/RPO and DR objectives. And by being easy-to-use, the DR4100 provides enhanced protection against downtime and disaster, while freeing up your time to work on more strategically important initiatives.

Conclusion

In conclusion, we saw that deduplication analyzed data at the block or byte level while SIS examined at the file level. We also learned that smaller, variable block selection size means more duplicates will be found for greater space savings. We observed that, while inline processing saves some space, it is not as fast or flexible as post-processing deduplication. Source-side deduplication reduces network loads, but significantly increases the client processing workloads when compared with target-side deduplication. We noted that, while deduplication ratios may sound dramatically different, the actual difference in space savings is marginally smaller after about 10X. Finally, we found that the longer data was retained, the greater is the likelihood that duplicate data will be found.

We saw that NetVault SmartDisk and the Dell DR4100 backup and recovery appliance offer two versatile and cost-effective solutions for you because either choice reduces the amount of storage you have to manage and shrinks your backup windows.

For More Information

About Dell

Dell Inc. (NASDAQ: DELL) listens to customers and delivers worldwide innovative technology, business solutions and services they trust and value. For more information, visit www.dell.com.

If you have any questions regarding your potential use of this material, contact:

Dell Software
5 Polaris Way
Aliso Viejo, CA 92656
www.dell.com
Refer to our Web site for regional and international office information.