



Data Deduplication

Dramatically reduce the storage space you need & speed disaster recovery

By Mike Theriault
President & CEO B2B Computer Products LLC

The prime cause of out-of-control data duplication is (ironically) the current standard backup protocol requiring numerous copies of every document *just in case*. The situation is further complicated by ever-expanding legal requirements.

The best way out of this quagmire is data deduplication - a key technology for any organization wanting to optimize the performance, efficiency, and cost-effectiveness of its data storage environment.

As business becomes increasingly paperless, everyone wants to be absolutely sure that their documents are backed-up and safe. What this means is multiple copies of everything in the data center file share, the Internet FTP server, personal folders, etc. – all contributing to a system-wide clog.

Data deduplication software removes the clog and keeps the stream of data running smoothly.

The term *Data deduplication* is commonly shortened to *data dedup*. Essentially, it's a process of identifying and removing multiple occurrences of the same data. The first time a deduplication system identifies a file, block, or bit, it flags it. From there, the system marks each subsequent identical item with a placeholder before removing it from the system. The placeholder links to the original data so that users will always bring up the original data when they try to open the removed duplicate.

This deduplication process significantly reduces the amount of storage space needed in the system. For example, a system that has 200 copies of the same 5 MB document- one in each employee's personal folder- can reduce it to a single copy of the original file plus 199 placeholders that link back to the original document. This means 200 copies of a 5 MB file will take up 5 MB of space (plus the size of the index file) instead of 1,000 MB.

You're probably already compressing files to save storage space. Compression reduces the size of the file by eliminating redundant bits. Compression is better than doing nothing, but it doesn't eliminate redundant files. In fact, it just compresses multiple copies of the same files.

Data deduplication goes a step further by eliminating those redundant copies, storing only one. Storage-wise, this makes a big difference. Simply compressing files reduces storage space by about 50 percent, but data deduplication reduces storage space by a much greater percentage, as the 5 MB document example above illustrates.

A number of characteristics differentiate deduplication processes. Some approaches are inline, others are postprocess. Inline deduplication means that data is deduplicated before it's committed to a disk drive. The postprocess approach first writes data to disk, and once a job or a dataset has been completed, deduplication follows

313 S. Rohlwing Road
Addison, IL 60101
p 630.396.6300
tf 877.222.8857
f 630.396.6322
www.B2BComp.com

Benefits

With data deduplication, users can streamline backup, facilitate emergency data restoration, and reduce costs.

Efficiency - Streamline Backup

As system back-ups become quicker and easier, users will be able to create and maintain more backup sets that stretch further back in time. This lets users keep a complete set of document versions without straining the system.

Performance - Facilitate Emergency Data Recovery

Until the advent of reliable data deduplication software, data compression was the only way to reduce the file size of the data stored offsite. With data deduplication, the backup set is both compressed and reduced – this means only the data that changed that day is backed up at the end of the day. This significantly reduces the transfer time in the event of a data restoration. Users no longer have to compromise on performance to take advantage of extended retention capability.

Cost-Reduction

By eliminating duplicate data and ensuring that data archives are as compact as possible, companies can keep more data online longer – at significantly lower costs. It's common to see a backup appliance with data deduplication technology holding 10 to 50 times more backup data than a conventional disk storage product. The advantage depends on the data being backed up, the backup methodology, and the length of time data is retained.

In addition, organizations of all sizes have seen a range of measurable benefits from deduplication that includes the ability to:

1. Realize a quick return on investment - storage capacity requirements are reduced immediately
2. Reduce tape costs by up to 80 percent and eliminate the need to invest in virtual tape libraries
3. Lower overall storage costs - storage capacity is used more efficiently, this means fewer storage system purchases.
4. Reduce backup storage by up to 95 percent
5. Reduce virtualized data by up to 90 percent
6. Minimize backup windows and reduce network utilization by up to 90 percent
7. Reduce network bandwidth - the amount of data being transferred over a network to a target location can be reduced by up to 20 times.
8. Help the environment - requirements for power, cooling and equipment are reduced.
9. Faster backup and increased data retention capabilities allow efficient response to legal and corporate compliance requirements.
10. Leverage data deduplication at the source or target to best meet specific requirements.
11. Lower shared server resource impact — achieve up to 50 percent increase in server consolidation through more efficient VMware backup.

Drawbacks

Data loss can be an issue when backing up data using data deduplication. If the index file becomes corrupted, the data processed using data deduplication may be lost, since it will be nearly impossible to rebuild the data. However, this concern may be alleviated with leading-edge software and by making multiple copies of especially important data.

What to Look for

Here are a few considerations for choosing the best products and services:

- Determine your requirements first, then look for the technology that can fulfill them. You need to think about what deduplication functions your existing system has, whether your existing system will accommodate data deduplication software, and – if not – whether you're willing to rebuild your system from the ground up. If you've determined that your existing system can integrate data deduplication software, be prepared for the fact that you may have to compromise functionality, scalability, and ease of use. Determine the degree of probable compromise and ask yourself if you can live with it.
- If you decide to integrate data deduplication hardware and software into your current system, it should be as non-disruptive as possible. Many companies turn to virtual tape library (VTL) technology as a method of introducing a data deduplication system and improving the quality of their backups without significant changes to software. The capabilities of the VTL itself must be considered as part of the evaluation process. Determine how well the VTL can mimic your existing tape environment and consider such things as the performance, capabilities, and stability of the VTL. If you use a disk-to-disk (D2D) backup model, you'll need a data deduplication system that has a network interface with the backup application. This process improves and simplifies D2D backups, allowing deduplication to take place without disrupting ongoing operations.
- Consider how data deduplication aligns with the backup process. Some data deduplication systems run while data is being backed up, processing the backup stream as it runs through the deduplication software. This approach can slow backups and eventually degrade deduplication performance. But data deduplication systems that run after backups are finished or that run concurrently with backup processes avoid all this.
- Because a data deduplication system is often used for long-term data storage, scalability, - capacity and performance - is important. Estimate what your requirements will be in 5 years.
- You'll want data deduplication to occur across your enterprise – headquarters, regional offices, offsite data storage, etc. A data deduplication system that allows multiple levels of deduplication and ensures that only unique data across all sites will be centrally replicated is ideal.
- It is extremely important to create an available, failsafe deduplication repository. Since a very large amount of data is consolidated in one location, there should be zero risk of total data loss. This means complete access to the deduplicated data repository is critical and the repository should not be subject to failure. A good data deduplication system will protect against local storage failure as well as provide replication for extra disaster protection.
- File-size deduplication systems do not reduce storage capacity as much as those that analyze data at a block or bit level. While larger chunks of data are processed at a faster rate, the trade-off is lower duplication detection. But if the solution has the capability to look for duplication in chunks and bits within the files, the duplication detection will be much higher. Some sophisticated deduplication solutions can self-adjust chunk size to optimize deduplication – leading to a large increase in the amount of duplicate data detected.

Focus on the total solution. Before deciding on a data deduplication system, be sure that you understand the entire process, from backup to restore, and know how to manage it.

Finding a Vendor

While the benefits of data deduplication can be astounding, pay attention to facts, not hype. It's important to deal with a vendor that you know and trust. The ideal vendor will have network of experts on staff and offer an optimal combination of experience, product choice, technical support, and competitive pricing.

Among the other considerations is the vendor's ability to proactively address your data deduplication hardware and software needs. Failure to do so could mean the difference between replacing the entire system in 5 years vs. updating select components.

As the optimal way to dramatically reduce data volumes, shrink storage requirements, and minimize data protection risks and costs, data deduplication is a vital technology. Nearly all organizations will realize significant benefits from it.

About B2B Computer Products LLC

Award-winning B2B Computer Products LLC was identified by *Inc.* magazine as one of the fastest growing businesses of its type in the U.S. in 2009 and 2010, and by Crain's as one of the largest privately held companies in the Chicago metro area. B2B Computer is a single-source provider of products and manufacturer-certified services that include virtualization, VoIP systems, data deduplication, disaster recovery, SAN storage, server consolidation, energy-efficiency improvement, and testing environment implementation. B2B Computer's engineers can design, configure, install, and/or manage the products and systems they sell to their clients. As a national business-to-business reseller of computer hardware and software representing hundreds of manufacturers – B2B guarantees a best practice combination of competitively priced customized products and expert services. In addition to its Addison, Illinois headquarters and multiple distribution points, B2B Computer's offices are in Chicago; New York; Davenport, Iowa; Philadelphia; and San Francisco. To contact B2B Computer, call 1-877-222-8857 or visit www.B2BComp.com

B2B Computer Products' Data Deduplication Solutions

BACKUP SERVICES

- Architecture
- Implementation
- Support

B2B Computer has dedicated data backup specialists that offer both pre and post-sales expertise. Our sales and engineering specialists are well qualified to help architect backup solutions as well as implementing and supporting these solutions after the sale.

COMMVAULT SYSTEMS, Inc.

- Server Backup (Galaxy)
- Email and file archival (Archive)
- Host based replication (CDR)
- Deduplication
- Backup to Disk

CommVault's Simpana platform offers a single integrated data management family capable of backup, recovery, email archiving, file system archiving, host-based replication and deduplication just to name a few. All of this managed from a single backend database and a single console. CommVault offers an Enterprise solution that scales from very small to very large. It also supports all major storage and host platforms. CommVault has experienced rapid growth while competitors have slowed. CommVault's products are also highly awarded for quality and customer satisfaction.

EMC CORPORATION

- Server Backup (AVAMAR)
- Backup to Disk VTL (DL)
- Deduplication (AVAMAR)

EMC has backup solutions including backup software, backup to disk and deduplication. With the AVAMAR solution, EMC offers a unique hardware and software backup and recovery solution with built-in client-level deduplication and backend hardware redundancy for an all-disk based solution. EMC also offers Virtual Tape Library (VTL) products based on EMC CLARiiON SAN hardware.

QUANTUM CORP.

- Tape Backup (Scalar)
- Backup to Disk VTL (DX)
- Deduplication VTL (DXi)

Quantum offers industry leading tape backup products in addition to backup to disk solutions with built-in deduplication. Tape and disk solutions scale from very small to very large and support all major backup applications and host platforms.