



Data Deduplication Best Practices

As you might guess, data redundancy is a primary contributor to explosive data growth. Studies estimate that multiple copies of data require organizations to buy, use, and administer two to fifty times more storage than they'd need with data deduplication.

Initially, data deduplication eliminated data redundancy in specific cases like full backups, email attachments, and VMware images. However, you'd soon notice the pervasiveness of duplicated data. That's because test and development data multiplies across an organization over time. Replication, backup, and archiving create multiple data copies scattered across the enterprise, and users often copy data to multiple locations for their own convenience.

Organizations now recognize that—far from being a niche technology—deduplication should be an integrated and mandatory element in their overall IT strategies.

There are essentially two ways to reduce the cost of your data storage. First, you can try to leverage a lower-cost storage platform, which results in an additional set of problems. Your other option is to leverage data deduplication to reduce your data growth and total required storage.

Data deduplication can lower the cost of your data storage by reducing the amount of disk needed to store your data, whether it's backups or online primary production volumes. Here are five best practices to help you select and implement the optimum deduplication solution for your environment.

Top Five Deduplication Best Practices:

1. Consider the broader implications of deduplication.
2. Learn what data does not dedupe well.
3. Don't obsess over space reduction ratios.
4. Don't use multiplexing if you're backing up to a Virtual Tape Library.
5. Pilot multiple deduplication systems before selecting your solution.

1 Consider the Broader Implications of Deduplication

Like disk-to-disk backup or server virtualization, you don't want to evaluate deduplication as an isolated product or feature. You must consider the broader implications of deduplication within the context of your entire data management and storage strategy.

For example, deduplication can be performed at the file, block, and byte levels. You'll have to consider the tradeoffs for each method, which include computational time, accuracy, level of duplication detected, index size, and in some cases, the scalability of the solution.

Also, consider how you can use deduplication to eliminate tape where it makes sense in your environment. That might be remote offices or any locations where your company doesn't have trained IT personnel.

Consider how you can use deduplication to eliminate tape where it makes sense in your environment. That might be remote offices or any locations where your company doesn't have trained IT personnel.

2 Learn What Data Does Not Dedupe Well

In the simplest terms, data created by humans—documents, transactions, and email for example—dedupes well in most dedupe systems. Photos, audio, video, imaging, or data created by computers generally don't dedupe well, so you should store these sets of data on non-deduped storage. Learn what data does not dedupe well in your particular environment, and consider not deduplicating it. For some situations, you might consider a deduplication solution that can selectively avoid certain sets of data.

3 Don't Obsess Over Space Reduction Ratios

The length of time that data is retained affects data deduplication ratios in two ways: If more data is examined when deduplicating new data, you're more likely to find duplicate data and increase space savings.

While you should closely examine this number when you're comparing multiple products, try not to overanalyze this number once your system is up and running. Rather than performing more frequent full backups just to get a better data deduplication ratio, consider increasing your backup retention period for your on-disk data store. Once you have your first set of backups on disk, adding additional backups to that same deduped system will take up less space than sending them to tape.

4 Don't use multiplexing if you're backing up to a VTL

If you're backing up to a virtual tape library (VTL), don't use multiplexing. Even if your deduplication solution can de-multiplex data, consider turning this feature off. Often a carryover practice from writing to physical tapes, multiplexing data merely wastes computing cycles—cycles that could otherwise be used to dedupe your data faster. For example, instead of multiplexing ten backups to two virtual tape drives, create twenty virtual tape drives and turn off multiplexing.

5 Pilot multiple systems before selecting your solution

Before selecting your deduplication solution, try to pilot several deduplication systems in your environment. While current vendors offer many good solutions and various deduplication approaches, you may also find some products with real limitations. Only by comparing multiple products can you best determine the optimum approach for deduping your data, whether it's inline, post-process, target-side, client-side, via backup software, etc.

Avoid Unnecessary Complications

Common challenges of deploying a deduplication solution involve problems related to performance, increased complexity of management, and proliferation of deduplicated data silos. To avoid unnecessary complications, first ensure ease of integration into your existing environment and get customer references in your industry. Take time to understand the vendor's roadmap, but *test everything*. Once you've selected your data deduplication solution, make sure you follow the best practices suggested by your deduplication solution vendor.

Rather than performing more frequent full backups just to get a better data deduplication ratio, consider increasing your backup retention period for your on-disk data store. Once you have your first set of backups on disk, adding additional backups to that same deduped system will take up less space than sending them to tape.

To avoid unnecessary complications, first ensure ease of integration into your existing environment and get customer references in your industry. Take time to understand the vendor's roadmap, but *test everything*.

Essential Features of Deduplication Solutions

When evaluating deduplication solutions, look for the following essential features:

- Ability to scale without expensive hardware upgrades.
- More recovery points and with shorter recovery times.
- Point-and-click deduplication management.
- Built-in reporting of deduplication across vendors, data types, sources and platforms.
- Tight integration with all necessary applications to minimize end-user downtime.
- Single solution simplicity for ease of deployment and administration.
- Ability to rapidly and securely recover business-critical data across all locations, applications, storage media and points-in-time.
- D2D2T-optimized for backup performance and reliable data recovery.
- Fast, comprehensive search to aid in recovery.
- Data integrity and security features.
- Built-in Disaster Recovery capabilities.
- Data classification.
- Cost-effective and timely eDiscovery.
- A common technology platform.
- Single point of management.

About the Author

Mark Teter is the Chief Technology Officer at Advanced Systems Group. He is an internationally recognized authority on information technology who regularly advises IT organizations, vendors, and government agencies on a broad range of information management issues. Each year, Mark conducts dozens of seminars and training programs for corporate and government institutions. He sits on several financial industry advisory boards and has recently published *Paradigm Shift: Seven Keys of Highly Successful Linux and Open Source Adoptions*.

About Advanced Systems Group

Advanced Systems Group (ASG) is a Denver-based IT consulting, integration, and project management firm—fully equipped with a high-end computing facility that provides testing, benchmarks, demonstrations and an executive briefing center. Acknowledged by *Computer Reseller News* as one of the Top Ten Storage Solution Providers, ASG pursues active involvement in the industry, maintaining the highest level of engineering certifications with partners and the vendor community.



ADVANCED SYSTEMS GROUP
Technology for the Speed of Change®

Call us at 800.894.3619
www.virtual.com

Denver . Colorado Springs . Salt Lake City . Phoenix . San Diego . Los Angeles . Orange County
Portland . Boise . Houston . Seattle . New Orleans . Baton Rouge . Oklahoma City . Tulsa