WHITE PAPER

Don't Be Duped By DEDUPE

By Jon Toigo

arcserve®

TABLE OF CONTENTS

arcserve®

Summary

While de-duplication has become a fixture in many IT shops, especially as part of the data backup process, its broader application – to primary storage – remains a problematic one. It can be reasonably characterized as a tactical and short term solution to a larger issue of rising storage capacity demand due to unmanaged data growth. Still, de-duplication is being touted by many vendors as a panacea technology that can bend the cost-curve in storage generally and in networking more broadly. Common sense should guide the application of de-dupe technology to avoid being duped by exaggerated claims of vendor evangelists. This paper provides some business savvy criteria for evaluating the de-duplication option and its fit with data storage and data protection requirements.

Introduction

De-duplication was initially harnessed to reduce the storage capacity requirements of disk-based backups.

De-duplication – a set of technologies for reducing the space occupied by data by identifying common block patterns, file names or other content in electronically-stored data and replacing them with "shorthand" hashes and/or stubs – was initially harnessed to reduce the storage capacity requirements of disk-based backups. Early on, industry insiders claimed that de-duplication was a waste management system for backup, enabling nightly full backups to be rationalized so that only changed data was retained. Simply, performing a week of full backups of a 10TB storage environment would result in 70TB of backup data without any sort of de-duplication at week's end. Applying de-duplication to eliminate from backup sets taken on days 2 through 7 any data replicating the full backup set recorded on day 1 would yield a much smaller aggregated amount of data by day 7 consisting of a full backup and change data from each of the subsequent days.

This approach was virtually identical, in concept, to the idea of incremental or changed data backup, which isolated and recorded only files that had changed since the prior backup. However, as backup processes evolved from file system centric to machine image centric, such approaches became more difficult to implement. De-duplication seemed like an answer to a new challenge.

Of course, the reason for the concern about the storage capacity consumed by backups was not simply about optimizing the backup process. Rather, it went hand in hand with an evolving preference for using disk media as

backup media, rather than tape. Tape ran afoul of capacity and resiliency issues in the late 1990s, while backup software also had its "growing pains." More importantly, a concerted marketing effort by disk array manufacturers to characterize tape as a legacy technology and to extol disk-based backup as an alternative to tape backup that was much easier to implement and automate yielded the desired results and pushed disk-based backup – even with its significantly greater costs – to the forefront. It was the increasing amount of data to be backed up, plus the use of expensive disk-based targets as backup data repositories, that ultimately propelled de-duplication into the forefront.

The one operational constraint of de-duplicated disk-based backup was the lack of a solution for mobilizing backup data and moving it efficiently out of the environment where it was created.

The one operational constraint of de-duplicated disk-based backup (other than storage capacity cost) was the lack of a solution for mobilizing backup data and moving it efficiently out of the environment where it was created. Tape-based backup data used portable media (tape cartridges) that could copied and located away from the production facility readily – insulating the backup data from the same disaster that might consume the original data assets. With disk-based backup, data needed to be transferred across a fast network interconnect and replicated on an alternate stand of disks. Doing this over distance required the use of a Wide Area Network (WAN), which introduced the problems of propagation delay and jitter into the data protection equation.

De-duplication vendors introduced the idea of faster replication across WANs and metropolitan area networks (MANs) as a selling point for their wares. While the volume of data to be transferred was certainly less if the data had been de-duplicated prior to sending, the act of de-duplicating data did nothing to accelerate the movement of data across network facilities. In point of fact, moving 10TB of data across a T-3/DS-3 link would still require more than a year, whether the data was de-duplicated or not. This did not stop the marketecture around faster network-based replication from becoming part of the de-duplication narrative.

Finally, questions began to arise regarding the impact of de-duplication (1) on the timeframe for the restoral of data to a useable form following an interruption event and (2) on the conformance of data storage with legal and regulatory mandates around data governance and retention. On the first matter, concerns began to surface about the "overhead" imposed on data

restore by the need to re-hydrate de-duplicated data in order to return it to a useful format. If backup software was used to copy the data from the production environment, the data was already placed into a backup "container" or format preferred by the backup software vendor. Add to this the possibility that the data in the backup container might be subjected to an encryption algorithm for security reasons. Finally, add in the de-duplication process, designed to de-hydrate the encrypted backup set. The operator now confronted a need to unlock their data from multiple software-imposed storage services before it could be restored for use by applications and decision-makers. The impact of the overhead might be important in cases where the criticality of an application required the expedited restoral of its data.

Another issue that began to gain importance involved the acceptability of data that had been subjected to de-duplication to regulatory or legal mandates around data preservation and protection. The SEC for example required publicly-traded companies to file quarterly and annual economic reports using a full and unaltered copy of financial data. Was de-duplication technology (particularly the block level processing variety) "materially altering" the required data? What company wanted to pay the legal fees that might accrue to testing the compliance of the technology?

Today, it comes as little surprise that many companies have huge swaths of data that are excluded from de-duplication processes to avoid legal exposure.

Today, it comes as little surprise that many companies have huge swaths of data that are excluded from de-duplication processes to avoid legal exposure. This has increased the functionality requirements for de-duplication products (they must be instrumented to enable the exclusion of certain data) and reduced the efficacy of the overall approach by limiting the volume of stored data to which the technology may be effectively applied. The latter point may be the more important one since excluding certain data for reasons of regulatory compliance adds to the volume of data that is already excluded by virtue of technology limitations: de-duplication doesn't typically work with rich media (video, graphics, etc.), with compressed data, or with database output; its use is restricted mainly to files.

Even in that carefully-defined use case, file storage, de-duplication still raises questions. One has to do with de-duplication ratios. If the case for de-duplication is based on its role in shrinking the storage capacity requirements of the data to which the process is applied, it stands to reason that an intelligent choice over which de-duplication product to deploy should be guided by the

ratio of data reduction that the product is designed to deliver. However, reduction ratios tend to be exaggerated by vendors – sometimes greatly.

An early offering in the de-duplicating appliance space (an appliance combines de-duplication software embedded on an array controller or "head" to which one or more racks of disk storage are attached) carried a manufacturer suggested retail price of over $400K. The components of the appliance were mostly commoditized server and storage gear with a street price of about $3 - $4K. The vendor claimed that the 70:1 reduction ratio that could be realized from the use of its de-duplicating software would enable each disk drive to handle data equivalent to 70 drives of data that was not de-duplicated – thereby justifying the $400K price tag.

De-duplication technology still provides a mixture of capabilities and limitations that need to be explored by consumers prior to adoption.

Since that time, the claims about de-duplication ratios delivered by products in this category have spanned so large a range that users are best advised to try the products under their own workloads and with their own data before they buy anything. That is as good a bit of practical, business-savvy guidance to those who are considering this technology as one can provide. Following are additional criteria that should guide the selection and deployment of de-duplication technology.

Buyer's Guidelines

De-duplication technology, for all of the vendor hype, still provides a mixture of capabilities and limitations that need to be explored by consumers prior to adoption. Here are a few check list items that might help in guiding decision-makers to a beneficial choice.

1. Know what you are trying to accomplish. The effort to use de-duplication to reduce the footprint of disk-based backup sets is quite different than the use of the technology to reduce the aggregate capacity demand of file-based production data storage. These differences need to be understood.

    a. De-duplicating backup data requires an understanding of the rate of change in backup data and the types of data that are being backed up. How much new or changed data is added each day? Is a de-duplication solution a cost effective alternative to incremental backups or tape-based backup with off-site storage?

At a minimum, planners need to design a backup process before purchasing a de-duplication solution.

b. De-duplicating production data requires careful consultation with governance, risk and compliance or audit managers to ensure that the technology does not compromise any legal or regulatory requirements. If certain data needs to be excluded from the process, this should be clearly understood so that the actual value of data reduction can be assessed.

2. Assess the alternatives for deploying de-duplication technology. Vendors are pressing the model of "in-line" de-duplication and source deduplication – two approaches that potentially reduce the storage capacity requirements for new data right from the start. With source de-duplication, a software process applies de-duplication algorithms on an on-going basis during data creation and storage. Many file systems have a de-duplication option that can be activated to apply the technology to an existing file I/O process. If file system-level de-duplication is not forthcoming given the kit that a company is currently using, in-line deduplication involves the insertion of an appliance in or adjacent to the storage I/O path where software applies de-duplication algorithms to data while it is "in flight" to the storage repository.

Target de-dupe typically involves the use of a specialized storage appliance or software-defined storage model.

Another alternative is target-based de-duplication or post-processing de-duplication. Target de-dupe typically involves the use of a specialized storage appliance or software-defined storage model in which data is written to the storage target, then subjected to a de-duplication algorithm that eliminates the redundant data. Post-processing approaches use software processes to de-duplicate data once written to storage. Often, the post-processing service is deployed on a backup server or general purpose server with access to data on whatever storage repository where it is deployed.

Post-processing and similar techniques impose no performance overhead on application since their work is performed after data is written to a target. This may be important, especially in virtual server environments that are already busy processing production workload.

Ultimately, the de-duplication solution that is chosen should be the one that best matches both the user's current configuration and the performance requirements of that configuration.

3. Measure the impact of de-duplication on backup and restore. Before purchasing any solution, it is important to test the impact of de-duplication on the time required to take a backup and the time required to restore data from a de-duplicated backup data set. Ensure that the product works with the data with which you intend to use it. Compressed data, and sometimes encrypted data and database or email system output, don't de-duplicate at all. Know what you are backing up and perform a test of the technology you are considering to determine, realistically, the impact of the product on critical time-frames and the actual reduction ratios that the solution delivers.

Software-only solutions tend to be the most cost effective.

4. Understand the actual cost of the solution. Software-only solutions tend to be the most cost effective, though integrated hardware/software solutions often have the greatest appeal in shops with limited IT support staff. Be aware that the mark-up on hardware dedicated to de-duplication (in-line appliances or de-duplicating storage targets) can be extraordinary and may not be merited by how effectively the product reduces storage capacity demand. Moreover, seek a solution that avoids the "gotcha" in many de-duplication products: the lack of scalability of the solution. With many products, once the capacity of the rig has been used, the customer must deploy another copy of the appliance or array – with a completely new de-duplication process to track and manage. The more appliances or arrays with more de-duplication algorithms and indexes, the more complex the management requirements. Ensure that you know what the requirements will be.

5. Make sure that the technology that you are deploying is consistent with the knowledge and capabilities of your IT staff. De-duplication processes are comparatively easy to start up, but only after a careful assessment has been made of the type, location and accessibility of the data that is being de-duplicated. If the technology is to be leveraged with an existing backup process and is post-processing based, the disruption of existing processes will be minimal. However, deploying in-line or source-centered de-duplication is nearly always

fraught with a requirement for upfront research and analysis (to identify data assets to include and exclude) and training for server and storage administrators.

Conclusion

De-duplication can be a great adjunct to data protection and data preservation projects, as well as a helpful technology for curbing storage capacity demand in both backup and production storage. Ultimately, however, the efficacy of the de-duplication technology that you deploy is contingent upon how well the need for the technology has been defined and how well the technology options themselves have been tested against actual workload and data at your shop.

The efficacy of the de-duplication technology that you deploy is contingent upon how well the need for the technology has been defined.

Look for a vendor who is willing to support pre-purchase testing or a trial period during which software can be deployed and used with certain data processes so that realistic expectations can be formulated regarding the impact of the technology. De-duplication does not replace tape backup, which is also steadily improving in terms of performance, capacity and cost. Be sure to include a tape trial with any data protection strategy you are considering.

Jon Toigo is a 30-year veteran of IT, and the Managing Partner of Toigo Partners International, an IT industry watchdog and consumer advocacy. He is also the chairman of the Data Management Institute, which focuses on the development of data management as a professional discipline. Toigo has written 15 books on business and IT and published more than 3,000 articles in the technology trade press. He is currently working on several book projects, including The Infrastruggle (for which this blog is named) which he is developing as a blook.