



Research Perspective

Future-proof Data Reduction

After years of debate about data protection, the storage industry had an epiphany: It's not about backup, it's about restore—getting your data back. What is the point of backing up data if one cannot quickly and efficiently find, restore, and use it?

Several years have passed and, amid rapid data growth and rising storage costs, two new darlings of debate have emerged: deduplication and compression. Industry pundits spend much of their time comparing the speed and efficiency of compression and deduplication products and debating the merits of source, in-line, and target-based implementations.

Although important, such topics draw attention away from data reduction's raison d'être. The goal is not simply shrinking data to occupy less space today, but doing it so that future data retrieval isn't costly, cumbersome, or downright impossible. Future-proof data reduction ensures that the cost savings of compression and deduplication investments today aren't swallowed up by the expense of data access in the future. In this Data Mobility Group report, we discuss the importance of future-proof data reduction and six related factors that can influence your investment in compression and deduplication technology.

Copyright © 2002-2010 Data Mobility Group, LLC. All Rights Reserved. Reproduction of this publication without prior written permission is forbidden. Data Mobility Group believes the statements contained herein are based on accurate and reliable information. However, because information is provided to Data Mobility Group from various sources, we cannot warrant that this publication is complete and error-free. Data Mobility Group disclaims all implied warranties, including warranties of merchantability or fitness for a particular purpose. Data Mobility Group shall have no liability for any direct, incidental, special, or consequential damages or lost profits. The opinions expressed herein are those of Data Mobility Group and subject to change without notice.

Disclosure: This research perspective was sponsored by Permabit Technology Corporation. All other brands, products, or service names are or may be trademarks, or service marks of, and are used to identify, products or services of their respective owners.



Unmasking Data Reduction

Compression and deduplication can provide immediate relief for companies struggling with financially unsustainable data growth. But they can also deliver a big future hit in the form of complex and costly data retrieval.

Today's compression and deduplication products employ two dramatically different approaches to data reduction and reconstitution: one forced, the other guided. Anyone considering the use of data-reduction technology should take the time to understand both.

The Forced Approach

The forced approach relies on a form of storage middleware—software installed where a company's data is created or where it is stored or somewhere in between—that analyzes, reorganizes, and shrinks data to occupy less space. While the software may or may not be in the data write path, it is always in the read path. That is to say, once your data has been reduced, it is unreadable and unusable without some form of external assistance to decompress/reconstruct the original file(s) from otherwise unintelligible bits. You are therefore forced to use that same middleware (or a compatible application) to access your own information assets.

Consider the operational risk involved: If your data-reduction equipment, or the network that communicates with it, should fail, you could not simply recover your storage system and carry on. You would have to repair or replace that data-reduction system, or the recovered data would remain unintelligible. Should you ever wish (or need) to switch to a different data-reduction technology, all your data would have to be migrated out through your old data-reduction system and back into the new system, with all the extra resources that process would entail. While this vulnerability can be addressed with equipment redundancy—something which data-reduction vendors often recommend—we will see in the next section that the vulnerability can instead be avoided altogether.



The Guided Approach

The guided approach is that of a “hands off” storage-advisory service—software that analyzes incoming data and advises your storage systems of the most efficient way to store data in the context of its own native extent mapping and management. While the software may or may not be in the data write path, it is *never* in the read path and neither writes nor reads data itself. Read and write operations remain the responsibility of the storage system into which the intelligence is embedded and will always be within the unaided power of that storage system—your storage system. Your vulnerability to data-reduction equipment is gone. You don’t need the secret decoder ring.

This approach accomplishes what storage systems should have been doing by themselves all along: optimizing storage capacity without creating third-party dependencies. Because reduced data is stored in a format native to the storage system it inhabits, no third-party assistance is required by the system to retrieve and reconstruct information assets. Your data, however ingeniously it has been sliced and diced, is yours—in its proper form without any external devices or software—any time you want it. There are no future data-reconstitution dependencies, no data-migration issues, and no worries about data safety or the reliability and resilience of third-party data-reduction middleware or, for that matter, of the third party itself.

Estimating the Impact

As part of due diligence, companies considering data-reduction technology must assess and consider the long-term costs and risk with respect to change management and data accessibility *first*, before assessing the potential savings. If a data-reduction technology has the potential to complicate and impede future data migration and retrieval, then it would be premature to discuss reduction ratios, throughput performance, and cost-savings until the future risks are fully understood. First we want to know which parachutes work, then what they cost.

Other Important Considerations

While Data Mobility Group believes that future-proof data reduction should always be a company’s first consideration, we acknowledge that it isn’t the only factor. Compression and deduplication



technology vendors make trade-offs to balance the performance and efficiency of their products and it is up to you to select the product(s) best suited to your specific needs. Other considerations include:

Scope

Data Mobility Group defines scope as the amount of data which is analyzed all at once in search of commonality. Intrafile methods (compression) identify and reduce commonality inside individual files, one by one, while interfile methods (deduplication) analyze and reduce commonality across entire repositories at once. Large amounts of duplicate or similar content can benefit from either approach, but interfile data reduction will always come out on top in terms of capacity optimization and, depending on the types of data, the difference can be substantial. Take, for example, 50 identical copies of an office document or multimedia file, each 5 megabytes in size. Compression might reduce each file down to 1MB, for a total of 50MB of compressed data. Deduplication would identify the commonality across all 50 copies, store the deduped equivalent of one, and consume as little as 1/50th the capacity required by compression—250 MB compressed to a single megabyte.

Scale

For data-reduction methods, scale is the point at which further use is impractical or impossible. Compression and deduplication solutions deal with scale at two levels: individual file size (the maximum file size they can effectively process) and aggregate storage size (the maximum number of files or amount of data a single product configuration can manage). Many of today's data-reduction product configurations max out at well under 100 terabytes. Ten years ago that might not have been an issue except for the largest corporations. Today, as businesses target ever larger markets, collect more data, ramp up analytics, and embrace social media—and when multi-terabyte external drives and networked storage devices are available even for home use—Data Mobility Group believes that petabyte-scalable data reduction is the only choice for forward-thinking businesses.

“Lossiness”

Lossless data-reduction methods allow the exact original data to be reconstructed. “Lossy” techniques, which literally throw out some data, allow an approximation of the original data to

be reconstructed in exchange for higher reduction ratios. Non-multimedia assets such as office documents and text files require lossless data reduction, but lossy methods can be advantageous in minimizing the storage and bandwidth footprint of audio, video, and image files such as JPGs, MP3s, AACs, and MP4s.

Compression technology can be either “lossy” or “lossless”; deduplication is exclusively lossless. For some companies, a mixture might be appropriate. While deduplication cannot match the extreme reduction ratios of lossy compression for individual multimedia files, it can eliminate unnecessary whole-file copies, which compression cannot do.

Block and File Storage

Data Mobility Group believes that data-reduction support for both file-based and block-based storage and protocols is essential, especially in mixed environments. Most data-reduction applications and appliances support one type or the other out of the box. You can implement two separate solutions, but that makes your storage infrastructure more costly and complex. We advise companies to insist on data-reduction technology that supports both file-and block-based storage out of the box, whether that technology comes pre-embedded in primary, secondary, backup, and archive storage systems or is purchased and integrated separately.

Holistic Data Reduction

When data deduplication was first introduced by the storage industry, the idea was to minimize both the amount of time it took to perform backups and the amount of storage those backup copies consumed. During the first few years of deduplication, Data Mobility Group questioned why the technology could not also be applied to expensive primary storage. Vendors voiced concerns about the potential performance penalties of in-line (i.e., in the write path) data deduplication on primary storage, but eventually developed ways to avoid those penalties (e.g., parallel and post-process data reduction). Voilà! Primary-storage deduplication was born. Around the same time, so-called “real-time” data compression was introduced and perfected for primary storage.



What this means is that data reduction can be applied across the entire information lifecycle. Most data-reduction applications and appliances still support either primary/secondary or backup data reduction and, while it is possible to implement multiple solutions, the result is a more complex and costly storage system. Data Mobility Group believes therefore that companies should insist on a single future-proof data-reduction technology that supports the entire information lifecycle end-to-end out of the box, whether it comes pre-embedded in primary, secondary, backup, and archive storage systems or is purchased and integrated separately.

Performance


The previous five factors have an impact on performance in two areas: write and read operations. In the past, Data Mobility Group believed that compression, due to its intrafile scope, was faster than deduplication. However, we no longer believe that to be universally true. The difference in performance now depends largely on implementation and recent innovations have rendered that difference negligible or nonexistent. Furthermore, the innovations underpinning future-proof data reduction have zero performance impact on storage read operations.

Data-reduction intelligence does not need to be implemented in the storage system's write or read path, so companies concerned about performance should insist on flexible future-proof data-reduction technology that need not be implemented in either path, ensuring that the performance penalty is negligible or nonexistent.

Summary

The future of compression and deduplication is clear: They will be embedded into next-generation storage systems. Since the introduction of data-reduction SDKs, storage system vendors have been able to rapidly incorporate data-reduction technology directly into their products—exactly where it belongs. Data Mobility Group believes that technologies such as compression and deduplication will continue to be embedded into storage systems and will eventually cease to exist as standalone applications and appliances. In fact, the process is well underway for a handful of OEM projects, with many more expected over the next few years. The situation is evolving rapidly and you should keep aware of the market and vendor dynamics.

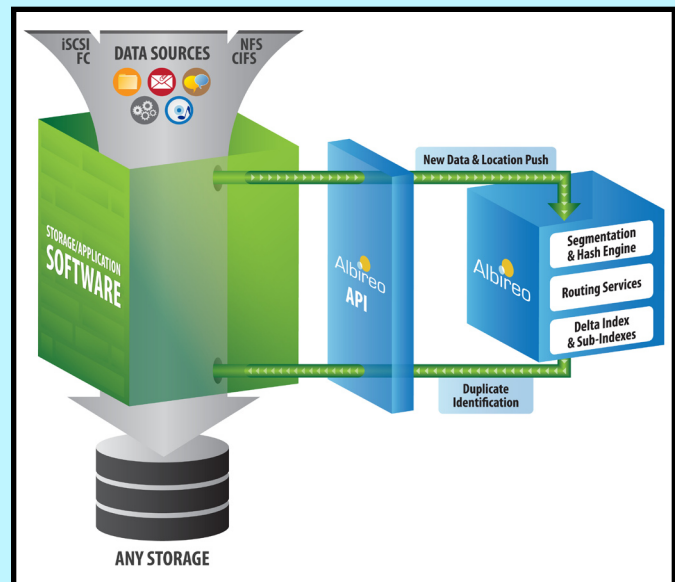


Future-proof data reduction propagated across the storage stack (end-to-end) minimizes costs (CAPEX and OPEX) and risk over the entire information lifecycle by reducing storage, resource, operating, and networking costs, by eliminating the potential performance penalties of forced data reconstitution, and by avoiding the inevitable costs of dependence on third-party data reconstitution. Future-proofing is the single most important factor in the long-term cost of data-reduction ownership and change management. As you struggle with financially unsustainable data growth and consider compression and deduplication technology for relief, one question should come first: “Is it future-proof?” From the short-list of future-proof data-reduction technologies, you can focus on the one that best fits your data world. 

for example

Permabit Albireo: Genuine Future-proof Data Reduction

Permabit Albireo High Performance OEM Data Optimization software performs data deduplication with both fixed and variable block sizes so it can be optimized for any type of data on primary storage. It uses the highly secure SHA-256 hash algorithm and stores hash keys in its High Performance Index Engine. Albireo is an embedded solution and always sits outside of the data read path (see figure on right) making it the only genuinely future-proof data reduction technology on the market today. Even if Albireo is removed or becomes unavailable the data *always* remains accessible—a benefit no other compression or deduplication technology can offer.



According to Permabit, Albireo’s High Performance Index Engine table is patented technology and returns hash queries/matches in a matter of microseconds—orders of magnitude faster than some deduplication solutions. The High Performance Index Engine is a derivative of the key index table,



and as such it contains information about the original hash key table, but is small enough to remain in memory. Each hash key requires only 4 bytes in memory vs. 64 bytes in the full hash key table on disk.

Independent tests have shown ingestion rates that exceed 140 MB/sec per processor core. Ingestion rates can exceed 1.6 GB/sec with a 4 KB chunk size and multiples higher with hardware hash acceleration. Because of its extremely small memory footprint, the High Performance Index Engine is conservative with system CPU and memory resources, a benefit when integrating with existing storage platforms. Routing algorithms minimize data movements and provide support for single or multiple name spaces in a single system.

Albireo is delivered in a Software Developer's Kit (SDK) that contains the full Albireo software library with a C-language API, extensive documentation, code samples, and application notes for both file and block integration. Permabit's Albireo Developer Support team provides full technical assistance to answer any questions and ensure rapid and seamless integration. In a recent implementation, one client commented that they had the Albireo library integrated and operational in just two days time and were taking performance measurements within four days.

Permabit Albireo is also delivered in an end-to-end storage solution—Permabit Enterprise Archive and Cloud Storage solutions. Permabit's Enterprise Archive and Cloud Storage solutions are some of the industry's most cost-effective, scalable, data-reduced storage systems capable of storing multiple petabytes of information. Both solutions offer full inline data optimization—deduplication plus compression—for maximum storage efficiency.

Any organization researching its data reduction technology options should take a closer look at Permabit's future-proof solutions. Additional product information can be obtained at www.permabit.com.

Questions?

If you have questions about future-proof data reduction or about available compression and deduplication technologies, we invite you to contact Data Mobility Group and arrange a FREE teleconference to discuss your concerns. We encourage you to invite members of your development team, IT staff, and suppliers to join the conversation.