Deduplication: You may know more than you think

Understanding your organization's data composition is the most important factor when making a deduplication decision

It is difficult enough that data is proliferating at an incessant pace across all IT environments yet the challenge is magnified by the ongoing need to protect critical data. Snapshots, replicas and clones all create duplicate data and increase the degree of difficulty for backup administrators. Fortunately there are many methods that can be employed to combat the proliferation of redundant data.

For example, when managing primary storage, snapshots should serve as temporary protection and be deleted after the first backup or another designated timeframe. Another best practice is to deploy an archiving solution for electronic documents that need to be stored for legal/regulatory purposes.

When discussing the reduction of backup data, deduplication has become the technology of choice. Deduplication delivers exceptional results including improved backup windows, increased productivity, and reduced costs. A 2012 IDC study of 155 customers shows that 68.7% of organizations are now using deduplication, a huge increase from 47% in 2011.[1] It's clear that deduplication helps organizations with their data reduction goals. If you haven't already started investigating this technology, the time is now.

If you have been researching deduplication, you may have observed that vendor claims are all over the map. This can lead any IT professional to ask a number of questions:

- Are vendors that advertise 50:1 dedupe ratios for real? Is 30:1 or 10:1 more realistic?
- Will the dedupe solution interfere with processor speed or network integrity?
- Is the sales person from the post-process dedupe vendor correct when he says that competing inline dedupe products will bring my network to a crawl?

Let's stop, take a deep breath, and try to put some of these dedupe claims into perspective.

Before you decide on a deduplication vendor you need to carefully evaluate your data composition and infrastructure

It's all about your data

Before you can reap the benefits of deduplication, a serious analysis of your data and infrastructure needs to be undertaken. Certain data types such as text files contain more duplicate data and therefore can be deduplicated more effectively than other data types such as images, videos or zip files. Other factors that have a significant effect include:

- Data growth rate—measures data being added to the data set
- Change rate—addition, deletion and modification to the data set
- Frequency of backups—for example, full weekly and daily incremental backups
- Retention period—length of time that backup data is retained before it is archived

Each of the above factors affects deduplication and associated compression ratios and helps zero in on the appropriate choice of data reduction solutions for your organization.

What about grandiose claims from deduplication vendors?

Why not buy the product that claims a 50:1 dedupe ratio instead of a product with a "measly" 10:1 ratio. The reason is that to achieve a 50:1 dedupe ratio, your data, your backup policies and possibly Halley's Comet would need to be perfectly aligned. You would probably be running full backups daily (frequent backups increase the dedupe rate), your data would be primarily text files (more redundant data than images or videos), and you would be retaining data for 60-90 days (the dedupe ratio improves with every backup).

If you met these ideal criteria and achieved a 50:1 ratio, you would be saving an eye-popping 98% on disk space. Yet that is only an 8% increase in savings compared to the 90% savings afforded by a 10:1 ratio. Even a low 2-1 dedupe ratio will achieve 50% in storage savings.

The point is that any dedupe solution will provide significant storage savings. Rather than obsessing over ratios, let's look at other essential criteria when evaluating deduplication. The primary areas to consider are the "how, where and when options" of deduplication.

1 IDC "Data deduplication gaining adoption and enabling disk-based data protection and recovery," April, 2011

How to deduplicate: There are three primary methods of deduplicating data:

- File-level deduplication (AKA single-instance storage)—examines files in their entirety and looks for duplicates.

- Block-level deduplication—breaks the data into blocks and looks for duplicates. This method typically provides greater storage reduction than file-level dedupe.

- Variable block deduplication—varies the size of the blocks to help locate additional commonalities. Variable block is usually considered to be the most effective method.

Where to deduplicate: There are two locations where deduplication can occur:

- Source deduplication—reduces the data on the host that is being backed up, decreasing the amount of data that must be transmitted over the network. This method appeals to customers with constrained network resources.

- Target deduplication—occurs on the backup server and/or target dedupe appliance. While this method requires more networking resources than source-based dedupe, it does not require overhead on the clients or servers being backed up.

When to deduplicate: There are two choices when using target deduplication:

- In-line deduplication—processes the data before it is written to disk. Data is deduplicated as it arrives to disk and can be replicated immediately. Some customers worry that in-line processing may slow down backup operations.

- Post-process deduplication—writes the backup data in its entirety to disk and then later dedupes it as a batch process. While this method doesn't affect the speed of the backup network, it can add delays of minutes or even hours to the replication process.

Deduplication appliance versus deduplication software

Another key decision point is whether to deploy a backup software solution application that includes deduplication or purchase a purpose-built backup appliance (PBBA).

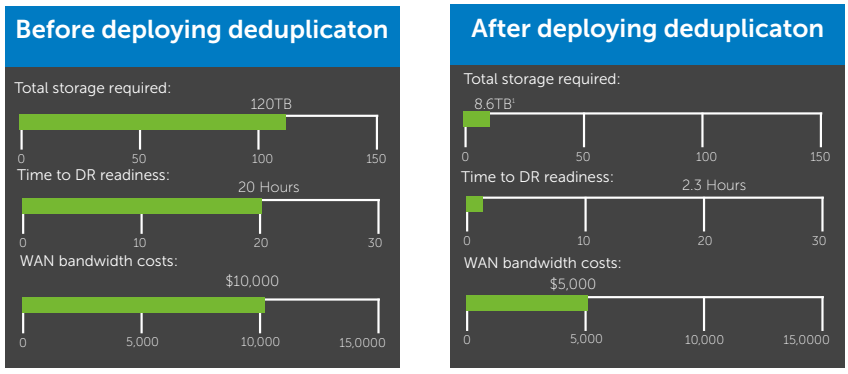
Organizations that are satisfied with their current backup software may just want to add deduplication as an integrated software option. Be aware that there may be added hardware costs because dedupe becomes the responsibility of the media/backup server which is already running its core functions. There may also be additional charges for dedupe licenses.

There are two types of PBBA. A PBBA backup target is an appliance whose primary function is to offload the memory- and processor-intensive deduplication function from third-party backup software. The other category, PBBA integrated systems, not only deduplicates but also includes the software that coordinates backup operations. Many times these appliances include built-in backup, replication and archiving, eliminating the need for multiple point products.

Accessing your options

As the amount of strategic data continues to grow and organizations continue to view data as a strategic asset, deduplication will become even more of an essential component in the backup infrastructure. Taking the time to evaluate deduplication options should result in beneficial results for your organization, but finding the most appropriate solution requires a detailed analysis of your organization's environment and data set. Although the analysis can often be performed internally, customers should look to solutions-focused deduplication providers that can provide the analytical tools and subject matter expertise to help you choose the best possible data reduction solution.

How much can deduplication save?



Customer environment

Primary storage: 9TB
Data type: User home shares
Weekly data change rate: 5%
IP LAB/SAN-based backup: Yes
Replication: Yes

Full backup: Friday
Incremental backup: Monday-Thursday
Full backup window: 5-7 hours
On-disk retention period: 3 months

This is an illustrative example. Actual results will vary by customer.
Example assumes 15:1 dedupe ratio.
Bandwidth savings will vary by line speed.

Simplify your storage at dell.com/deduplication

CB142_Selecting deduplication solution requires homework_041713