Not all deduplication is created equal

A COMPARISON OF DIFFERENT APPROACHES, BY EXAGRID

Why should data be deduplicated?
In standard backup practice, dozens of copies of primary data are be kept so that the data is available to be restored if necessary. If an organization keeps 8 weekly, 24 monthly, and 5 yearly backups, then it needs to store over 30 copies. The cost of disk can become too expensive for backup when of all these copies, known as retention, are kept.

This is where data deduplication comes in. Data deduplication compares one backup copy to another and only keeps the changes from backup to backup. It reduces the amount of storage versus having no deduplication at all.

There are two considerations that factor into data deduplication. The first is the how much deduplication can be achieved (how high the data deduplication ratio is) and the resulting storage reduction, and the second is the impact that deduplication has on backup and recovery performance, which can vary based on the approach. Data deduplication ratios are all over the map based on what size blocks are used, what advanced techniques are employed, and other factors.

How is the data deduplicated?

Data is either deduplicated by the backup application or by a dedicated backup appliance. Examples of backup applications that include data deduplication in their software are: Veeam, IBM Spectrum Protect, Commvault, Rubrik, and Cohesity, among others. When data deduplication is performed by a backup application, the block size is typically very large and also fixed so deduplication ratios are typically low, between 2:1-5:1. This is better than no deduplication at all as it provides some storage savings, but as an organization's retention needs grow, this approach will still require a lot of storage and therefore still be costly. The reason why backup applications only provide low deduplication ratios is because data deduplication is very compute-intensive which impact backup performance. In addition, all the data is deduplicated which greatly slows down the speed that the data can be restored.

Dedicated backup appliances usually achieve higher deduplication ratios, averaging from 14:1-20:1. The higher the deduplication ratio, the greater the storage savings. The reason why the appliances can achieve this level of deduplication is because they include dedicated hardware to handle the heavy compute required. The lowest-cost storage solutions are those that offer an average 20:1 deduplication ratio. These solutions have dedicated process, memory, and networking to handle the aggressive deduplication. However, these appliances are inline which means the backups occur during the backup window creating a compute-intensive bottleneck that slows backups down and expands backup windows. In addition, since all data is deduplicated, the restores are slow due to data rehydration for each request.

The drawback of inline deduplication in a backup application or deduplication appliance

In addition to deduplication ratios, an important factor to consider is a solution's approach to data deduplication. Most deduplication is performed inline, which means data is deduplicated while the backups are occurring and before the data hits the storage, so the backup performance takes a big hit.

Writing to disk is fast but if inline deduplication is used by either a backup application or an inline deduplication appliance, it creates a bottleneck which greatly slows down the backups by as much as 3 to 4X. In addition, if the backups are performed inline, whether by a backup application or a dedicated appliance, all the stored data is deduplicated. For every restore request, the data has to be put back together (called data rehydration) which takes 10-20 times longer than simply reading undeduplicated from disk.

Key Considerations When Adding Deduplication to Backup

When looking at data deduplication, ask:

- What block size is being used?
- What deduplication ratio is being achieved?
- Is the deduplication being done inline during the backups which slows backups down?
- Is the data stored only in a deduplicated format which results in slow restores, recoveries, and VM boots?
- OR, does the solution have a front-end disk-cache for fast backups and restores as well as a second repository tier for long-term deduplicated storage?

Deduplication in the backup application will save some storage but also reduces backup and restore performance.

Dedicated inline deduplication appliances save far more storage than the deduplication in backup applications provide, but they also are slow for backups due to inline data deduplication (when data is deduplicated on its way to storage) and slow for restores due to the need to rehydrate the deduplicated data that is stored on the appliance.

There is a unique approach to deduplication called a disk-cache Landing Zone which is used in Tiered Backup Storage. With this approach, backups are written to the Landing Zone first for fast performance then and adaptively deduplicated into a repository tier for long-term retention storage for cost efficiency. The most recent data in the Landing Zone is not deduplicated, so that restores are fast and avoid the lengthy and compute-intensive data rehydration process. Tiered Backup Storage offers up to 20:1 deduplication ratio, which provides a huge savings on long-term retention storage, without the negative impact on backup and restore performance.

Deduplication is key to managing backup costs, but it is important to consider which approach will work best in your backup environment.