Manager's Guide to Data Deduplication

— By SearchCIO.in

SearchCIO.in

Data deduplication is the process of preventing duplicate data from being stored and archived. This reduces the rate of data growth in an organization thereby significantly easing its management and lowering overall TCO.

In this guide:

- **Data deduplication basics**
- **Business benefits of data deduplication**
- **Data Deduplication disadvantages**
- **Key product evaluation considerations**
- **Data deduplication vendors and their offerings**

Data deduplication basics

Duplicate data results in increased storage, backup and retrieval time and costs. Multiple copies of data in secondary storage systems are the biggest source of this type of data. The greater the amount of data in an organization, the more difficult it is to meet RTO (recovery time objectives) and RPO (recovery point objectives).

Data deduplication helps address this by preventing duplicate data from being part of the equation.

Backup applications benefit from deduplication as most of the data in a given backup doesn't change from the previous backup of an existing file system.

Virtual environments also benefit quite a bit as marginally separate system files of different virtual machines can easily be combined into a single storage space.

Effectiveness of deduplication solutions is

measured in terms of the de-dupe ratio as follows:

De-dupe ratio *= Total data sent by the application(s) to the storage system/Total 'raw' storage capacity of the storage system*

Deduplication solutions have been applied mainly to secondary storage systems in the past. Vendors are now augmenting their existing solutions or offering new ones to apply deduplication to primary storage systems as well.

Business benefits of data deduplication

- Substantial reduction in data center capital and operating costs by lowering the requirements for equipment, power, cooling and floor space.

- Recently, tapes have been replaced by disks as secondary storage media. Deduplication slows the need to purchase new disks.

- Enabling more data to be backed up.

- Increasing disk efficiency, thereby enabling longer disk retention periods.

- Helping optimize use of network bandwidth by reducing the data sent on the wire for backup, replication and DR.

Data Deduplication disadvantages

- **Performance impact:** Data deduplication employing a fixed-block approach on primary storage may adversely impact performance.

  This is because of the inherently random nature of primary data leading to multiple chunks written at different locations, thus increasing data re-assembly times.

- **Loss of data integrity:** Block-level deduplication solutions utilizing hashes create the possibility of hash collisions (identical hashes for different data blocks).

  This can cause loss of data integrity due to false positives, in the absence of additional in-built verification.

- **Backup appliance issues:** Data deduplication requires a separate hardware device, often referred to as a “backup appliance”.

This has a cost, and can lead to performance degradation if data is required to be transferred from backup server to the hardware device.

- **Impact due to shared resources:** If a deduplication solution uses the hashing technique and shares computational resources on a device provisioning other services, overall performance is affected.

Key product evaluation considerations

Do you really require a deduplication solution? This decision would be based on internal diligence about type and nature of data redundancy, data change frequency, full backup frequency, RTO and RPO objectives. Once you have determined that a data deduplication solution is required, consider the following aspects:

- **What levels of de-dupe ratios** does the deduplication solution vendor claim? These claims must take into account the data set size as well as the point at which data deduplication is initiated, with linked assumptions about available processing power and network bandwidth.

- **Which of the three deduplication methods** would suit your organization's requirements?

  a) **Block-level** - This method operates by creating treating data as "blocks/chunks" of

data. Only one unique instance of the data is retained on secondary storage media. Hashes for each block are computed and stored in a hash lookup table. When new data enters the system, block hashes are computed and compared with entries in the hash lookup table. In case of a match, block data is replaced with reference to the existing block. An entry is made in the hash table for all other cases. Smaller chunks increase the probability of finding redundant data.

b) **File-level** - Works best for data that doesn't change often, such as employee profile. Each duplicate instance is referred back to the single, unique file copy.

c) **Content-aware deduplication** - Content is semantics of the data being duplicated. This helps determine resources (processor/memory/IO) required for data deduplication.

- **Inline or post-storage?** - To reduce storage requirements, the inline mode may

be used, wherein the deduplication method is applied on incoming data. This mode is better suited for block-level methods. Alternatively, in the post-storage mode, incoming data is stored first and analyzed later for duplicates. Prefer this method in case the deduplication appliance also doubles up as an NFS or CIFS share. This will ensure zero impact on application response times or access of shares by users.

- **Conduct systematic capacity planning of storage** required for data recovery. This is essential when considering recovery of de-duplicated data created from primary data snapshots, as the deduplication process expands the data to its original size and not the snapshot size.

- **Determine the type of device** your existing backup application expects the deduplication solution to be: VTL (Virtual Tape Library), tape or disk.

- Does the deduplication solution support

both NAS and SAN server connections?

- Does the deduplication solution support deduplication of data from any type of application?

- Does the solution move de-duplicated data or duplicate data for replication across the wire?

Data deduplication vendors and their offerings

Data deduplication tool vendor	Solution name and features
IBM	IBM ProtecTIER - Inline deduplication method ensures backup windows are adhered to without leading to any disruptions. Use of non-hash-based approach preserves data integrity.
EMC	Data Domain DD140 - Used for data deduplication in remote locations with replication for meeting RTOs. Data Domain Global Deduplication Array - Protection and storage of large datasets in a single deduplication system.
CA	CA ARCserve Backup - Employs target-based deduplication method to compare new backups with stored data at the block-level.
Symantec	*NetBackup 7* offers single console for multi-site monitoring, analytics and reporting. *NetBackup 5000 series appliances* offer source or target-based deduplication using telecom grade hardware. *NetBackup real-time* provides continuous data protection and replication for critical applications.
NetApp	Data ONTAP - Deduplication and compression are core components of NetApp's operating architecture, implementable across multiple applications and storage systems.
HP	HPStoreOnce - Deduplication technology used speeds up backups up to 4TB/hr and cuts storage capacity requirements by 95%. Can be used to centrally manage data center and remote office backups.

Further reading

Definition: Data deduplication definition

Tutorial: Data deduplication tutorial

TIP: Test data deduplication tools with these five guidelines

TIP: Deduplication of data: 10 tips for effective solution evaluation

TIP: Five data deduplication technology considerations

TIP: Nine data deduplication technology implementation considerations

TIP: A quiz on data dedupe

TIP: Data Center Research Library