# 5.5.3 Data Deduplication Facts

Data deduplication in Windows Server is designed to store data in less physical space using subfile variable-size chunking and compression, which can deliver optimization ratios of 2:1 for typical user files and up to 20:1 for virtualization data. Data deduplication finds and removes duplicate information across files without compromising data integrity. The data deduplication optimization process is as follows:

1. Files are segmented into small variable-sized chunks that are 32–128 KB in size.
2. Duplicate chunks are identified.
3. A *single* copy of each chunk is then maintained. Redundant copies of the chunk are replaced with a reference to the single copy.
4. The chunks are compressed and then organized into special container files in the System Volume Information folder.

When determining whether a volume should have data deduplication enabled, consider the following questions:

- Does duplicated data exist on the volume? Good candidates for data deduplication include:
  - Shared folders with user data
  - Software deployment shares
  - Virtual hard disk (VHD) file storage for provisioning to hypervisors
- How frequently does the duplicated data change?
  - User documents, virtual files, or software deployment files that contain data that is modified infrequently and read frequently are good candidates for deduplication.
  - Files that change often and are constantly accessed by users or applications are not good candidates.

The following storage configurations do not support data deduplication:

- Boot volumes
- System volumes
- Volumes with FSRM hard quotas configured
- Removable drives

Data deduplication includes checksum validation and metadata consistency-checking features to protect data integrity. It also provides built-in redundancy for critical metadata and frequently-used data chunks. If corruption is detected, it is recorded in a corruption log file. *Scrubbing* jobs analyze the chunk store corruption logs and attempt to make repairs. Three sources of redundant data are maintained:

- Backup copies are maintained for chunks that are referenced over 100 times in an area called the *hotspot*. If the working copy is corrupted, data deduplication will use a backup copy.
- If implemented on a mirrored Storage Space, data deduplication can use the mirror of the chunk to fix corrupted chunks.
- If a file is processed with a chunk that is corrupted, the corrupted chunk is eliminated, and the incoming new chunk is used to fix the corruption.

Data Deduplication has three schedules configured by default. Optimization runs every hour, while Garbage Collection and Scrubbing run once per week.

To enable data deduplication, complete the following:

- Add the Data Deduplication role service with Server Manager.
- Use the DDPEval.exe utility to analyze server volumes for data deduplication. The DDPEval.exe utility evaluates space savings that could be gained from implementing data deduplication.
- Use Server Manager to configure data deduplication on a server data volume:
  - Select **Enable data deduplication**.
  - Specify the number of days that should elapse from the date of file creation until files are deduplicated.
  - Specify the extensions of any file types that should not be deduplicated.
  - If necessary, manually specify any folders with files that should not be deduplicated.
  - Configure the deduplication schedule.
  - Apply the changes.

Data deduplication can also be managed using the following PowerShell cmdlets:

- **Enable-DedupVolume [volume]** enables data deduplication on a volume.
- **Set-DedupVolume [volume] -MinimumFileAgeDays [days]** sets the minimum number of days that must pass before a file is deduplicated.
- **Get-DedupVolume** displays a list of the volumes that have been enabled for data deduplication.
- **Start-DedupJob -Volume [volume] -Type Optimization** immediately starts an optimization job. If a job should start at a later time, use the **-wait** parameter with this command.
- **Get-DedupJob** displays the progress of a running optimization job.
- **Get-DedupStatus** displays key optimization statistics, including free space, space saved, and optimized files.
- **Start-DedupJob [volume] -Type Scrubbing** runs a scrubbing job that attempts to repair all issues identified in the corruption log. To check the integrity of all the deduplicated data on the volume, use the **-full** parameter.
- **Get-DedupSchedule** displays data deduplication schedules for deduplication jobs.
- **Set-DedupSchedule [schedule_name] [properties]** configures the schedule for a deduplication job.

You can also take advantage of data deduplication optimization on CSV volumes.

---