The Business Challenge

Business and IT are both struggling with how to protect, monitor, maintain and recover business-critical data when it continues to grow at 30-50 percent per year. Traditional approaches are breaking down as backup windows are unable to accommodate the increased time needed to handle growing datasets. In fact, as applications and data become ever more critical, backup windows are actually shrinking, which puts added pressure on your overloaded staff and infrastructure.

To complicate matters, businesses are now required to retain this data for significantly longer periods of time, either for business continuity or to meet compliance requirements. As businesses strive to meet their SLAs, rapid recovery is becoming the driving motivator for data management solutions. For this reason, disk-based backup solutions are gaining wide acceptance. However, customers have to increasingly balance the need for rapid recovery against the cost of keeping protection copies on disk for extended periods. Data deduplication provides an opportunity to reduce the amount of disk required for storage of multiple protection copies of long-term data.

Deduplication appliances, including Virtual Tape Libraries (VTLs), offer the ability to eliminate redundant data at the backend storage device, leading to better storage utilization. Yet this is a limited approach that does not address the larger challenges arising from rapid data growth. Not surprisingly, these appliances offer remedies for the short term. Unfortunately, problems tend to resurface within 6-12 months, requiring more hardware, leading to vendor lock-in and compounding storage management complexity.

These dedupe appliances take a “tape is the problem” approach, yet try to force the same inefficient data management policies as in a tape environment. For example, they require more frequent full backups, which lead to a tremendous load on production systems, network bandwidth and longer backup windows. While this may solve the problem of eliminating data redundancy on the storage device and lead to high “dedupe ratio,” it aggravates the load on the infrastructure as you try to manage larger data sets in smaller windows. Also, by imposing a reassembly tax on any data read operations, these devices cause a significant degradation in restore speed as compared to regular disk during critical data recovery moments. The only way to create a distinct disaster recovery copy of the deduplicated data is by replicating it to another device of the same type, doubling the cost of the dedupe solution.

Finally, while disk is cheap and getting cheaper, native tape still offers cost advantages when it comes to truly long-term archive periods due to savings in logistics, power, cooling, etc. Alternative approaches to data deduplication rehydrate the data when it is moved from a dedupe appliance to a tape device and all benefits of deduplication are completely lost for that storage tier.

Smarter Deduplication with CommVault® Simpana® Software

Extend and Embed Global Deduplication Across All Backup and Archive Data

Key Benefits

- Global deduplication across all data, regardless of data type, source or platform with a single integrated solution.
- Increase number of recovery points stored on existing disk to allow faster direct recovery of data.
- Hardware agnostic deduplication allows the use of any disk type as backup or archive target.
- Persistent deduplication from disk to tape-based archival storage tiers retains deduped data footprint, reducing the number of tapes and tape drives required.
- Global, policy-based deduplication eliminates the need to move or rehydrate existing data when adding incremental storage.
- Minimal impact on recovery speeds as “undeduplication” of data blocks is not required.
- Recover deduplicated data quickly and easily to alternate systems and locations for test, staging and DR.
- Content-aware deduplication enables better data alignment across multiple backups increasing deduplication ratios.
- Client-site deduplication only sends unique data to disk, reducing bandwidth and processing requirements on the disk target.
- Ability to deduplicate compressed or encrypted data.
- Ability to deduplicate compressed or encrypted data.

BACKUP & RECOVERY       ARCHIVE       REPLICATION       RESOURCE MANAGEMENT       SEARCH
CommVault® Simpana® Software Approach to Data Deduplication

CommVault® Simpana® software provides a holistic approach to data deduplication. Data deduplication should not be deployed as a band-aid or “quick-fix” point product. By embedding deduplication within an end-to-end backup and archive solution you not only derive better back-end storage efficiencies, but you can leverage the full capabilities of an enterprise data management solution to deal with rapid data growth, meet extended data retention requirements and allow fast and aggressive data recovery without overburdening the infrastructure.

Policy-based automation enables you to do more with less, so that your investments and headcount are focused on supporting strategic business initiatives, not just maintenance.

CommVault® Simpana® software’s embedded approach optimizes data during data capture and data transfer, thus reducing backup windows and network bandwidth requirements. By distributing the workload across clients and Media Agents, CommVault ensures more data can be processed faster. Powerful indexing and random disk access (vs. sequential access) ensures that data blocks can be recovered back to the recovery target in the correct order as quickly as possible. CommVault’s deduplication solution is hardware independent, enabling you to use any disk type, leveraging your existing storage investments. These factors also ensure that the solution can scale as your data grows and commodity hardware can be incrementally added to the dedupe store vs. purchasing dedicated and expensive appliances. Content awareness results in superior deduplication versus a blind backend data stream-based approach. CommVault software deduplicates across 100% of the backup and archive data managed through Simpana® software, delivering “global” dedupe regardless of data type, source or platform. CommVault deduplication is a multi-tier approach that extends to tape-based back-ups, drastically reducing the cost and timeframe of your archived data.

Figure 1: Embedded Global Data Deduplication
Smart Content Aware Deduplication

Smart deduplication begins with smart data selection. CommVault software’s data selection policies allow for the extension of backup cycles to reduce the frequency of full backups and increase the number of incremental backups. This not only reduces the amount of data stored in the backend, but also the amount of data sent over the network infrastructure. Compare this with a device-based approach that requires more frequent full backups in order to drive higher dedupe ratios, overwhelming the production system and the network. The CommVault indexing model ensures that each backup, whether full or incremental, provides a full recovery image of the system, allowing for rapid and direct recovery of data.

CommVault® iDataAgents (iDAs) reside on the clients and provide optimal data selection options to deal with rapidly growing data without overwhelming production systems. File System iDAs, Application iDAs with Snap-Backup or Data Classification Enabler and our Universal Virtual Server agents are examples of agents that do not overwhelm the production system while determining data that needs to be backed up. The iDAs optionally generate digital signatures for blocks of data, an important step in the deduplication process, thus reducing the workload downstream. Clients can compress and encrypt the data as well. Compression further reduces the amount of data transferred and stored. Encryption at the client enables encrypted data to be deduplicated with minimal overhead on the production system. This unique integrated capability delivers a secure base for enterprise-wide deduplication.

Clients can package a data-stream into smaller segments after stripping out all metadata associated with the files (like ACLs, attributes, etc.) and transfer the data over to the Media Agent using our proprietary Data Pipe mechanism. Since the iDAs are content aware, they can clearly designate the beginning and end of each object, as well as the backup metadata in the data pipe stream. When the Media Agent receives the data-stream, it can clearly identify object boundaries so as to achieve perfect alignment with previous versions of the same object. The Media Agent also knows to skip the metadata and only process the data segments for deduplication.

This provides superior deduplication compared to blind device-based approaches, which implement complicated algorithms to detect small shifts in the data-streams to achieve data alignment. Furthermore, the inability of such a device-based process to handle metadata (which is always unique) separately from the actual data also hinders efficient deduplication. The Media Agent compares the digital signature of each segment against a database of signatures in the global Deduplication Database that keeps track of all unique segments processed. Redundant data segments are discarded and appropriate references are updated in the dedupe store. Since the uniqueness check is made in memory, the Media Agent sends only unique data segments over the network to disk, reducing network bandwidth requirements. This also reduces the I/O overhead of writing data to disk, further speeding up the backup or archive process.

Every segment that is written to disk is also catalogued as part of CommVault software’s self-protected, distributed index. The index provides a point-in-time image of a client system as it existed at the time of backup and the offset into dedupe store where the related data segments are stored. During a data restore or read, the Media Agent performing the restore refers to the index to determine which data segments comprise the file (or files) required. It then uses random-access, disk data readers to read the segments from the disk in the correct order. These are sent back to the restore target over the data pipe. No reassembly of the data segments is required at the storage level.

CommVault® Simpana® software’s embedded approach to deduplication allows for an efficient, scalable and hardware-agnostic solution that helps deal with rapid data growth, extended data retention and allows for fast low-impact recovery. Smart integrated deduplication spans your infrastructure, from individual clients all the way to backend storage. Other approaches take a parochial or isolated approach to deduplication. CommVault’s iDA agents, Media Agents and the CommCell® management software component all work together to leverage the core capabilities of the Common Technology Engine, the foundation of Simpana software.
Extend and Scale Data Deduplication Across the Enterprise

Global Policy-Based Deduplication
By using a global deduplication-enabled policy, you can take advantage of a robust global deduplication capability that spans 100% of Simpana® software-managed data including backup and archive data. Using a combination of the GridStor™ feature on the deduplication policy and a shared Dedupe Database (DDB), you can extend the deduplication store across multiple Media Agents and multiple disk devices. The GridStor™ feature automates load balancing and failover of data streams across several Media Agents. A shared Dedupe Database keeps track of all unique segments sent by backup and archive clients to the global policy. This ensures that data from different backup and archive agents is deduplicated across all GridStor™ Media Agents and related disk targets, regardless of which Media Agent processed the data stream.

Hardware Agnostic
CommVault’s software-based approach to data deduplication leverages your existing storage investment with the ability to work with storage systems from a wide variety of vendors. CommVault® Simpana® enables different disk types to be added as data paths to a global deduplication policy. Your choice of disk is not limited to disk from a particular vendor or to a specific interface type. These disk targets can be connected directly to the Media Agents over storage area networks using Fibre Channel or iSCSI. They can also be presented as a network share using CIFS/NFS or even a Global File System (GFS). Since deduplication is managed at the storage policy level, the deduplication store can span all disk and network types.

As the amount of backup and archive data grows, there will be a need to expand the deduplication store to accommodate the new data. With a CommVault® software solution, you can incrementally add commodity disk to the dedupe store as a new datapath, eliminating the need to buy additional deduplication appliances.

Automatic Forward Referencing
CommVault® deduplication software uses the notion of forward referencing. After a configurable time interval, the Media Agent writes a new copy of the data segment even if a copy already exists in the dedupe store. All future instances of this data segment then point to the new copy. When the forward referencing time interval is aligned with the data retention on disk, all valid backup and archive data on disk will refer to the latest copy versus the earliest one. This ensures that should a data segment become unknowingly corrupted on disk (due to hardware issues), the entire dataset will not also become corrupted. Users can continue to access and recover data from the dedupe store without worry. This process also ensures that the Dedupe Database is continuously refreshed and references to older copies of data segments are removed from the system as the data ages, enabling the space utilized by those older segments to be reused.

Fast and Scalable
All of the above factors add up to the most robust and scalable deduplication solution available today. CommVault® deduplication software enables your corporation to keep pace with ongoing rapid data growth, while comfortably meeting all your data recovery SLAs. As the amount of enterprise data grows, it is only a matter of incrementally adding disk (and additional Media Agents if needed) to the global policy. CommVault® deduplication software spans multiple pools of data, including multiple data tiers, decreasing the cost of all of your backup and archival data tiers.

CommVault® Simpana® software architecture allows for unlimited corporate data growth that is not negatively impacted by your deduplication approach. CommVault deduplication will scale with your business and data requirements when and where you need it.
Bringing Dedupe to Long-Term Tape-Based Retention (and Back)

While most enterprises are moving to disk to perform data recovery, there is still a need to move data to offline media such as tape for long-term archiving. Unfortunately, other deduplication approaches cause a complete rehydration of the data once it is copied to offline media. CommVault introduces the first solution in the industry to break down the deduplication wall between disk and tape. CommVault software allows you to complement disk-based dedupe storage with long-term, tape-based storage tiers for archived data. You can use our automated, policy-based capabilities to seamlessly shift deduplicated data from disk to a tape library without rehydrating the data. Since the deduplication form is retained, the data that is moved has a much smaller footprint, resulting in fewer tape drive requirements and a smaller number of tapes. This data transfer is transparent to the end user. When data from the tape is required, instead of first copying all the data on tape back to disk, we bring only the relevant pieces of data back to the disk-based dedupe store from where the restore request is serviced. This reverse data movement from disk to tape is also transparent to the end-user.

Contrast this with a VTL approach where the data is fully rehydrated as it is moved from virtual tapes to physical tapes. Special handling by an administrator is needed to ensure that the data on the physical media is returned to the virtual tape before a restore can be initiated. The entire physical tape then has to be copied back to VTL disk, not just the relevant data. All of these steps add crucial delays and complexity to the recovery process.

With Simpana® software, once the deduplicated data is on tape, you can optionally choose to leave a copy on disk and send the tape offsite for long term archive. You can also choose to create more than one tape copy; each copy will be in a dedupe format. One copy can be kept in-house and the other can be sent offsite. This method can also be used to create DR copies of your data in case the primary dedupe store or the entire site is lost. Data can be copied to tape and moved to a DR site. In case of a disaster, data can be read directly from the tapes in the DR site to another CommVault® dedupe store and be made available for recovery. This is a much simpler and less expensive approach to DR than other deduplication appliances where the only method is to provision another appliance of similar or larger capacity and replicate the data from one location to another. This doubles the cost of the solution, while locking you into a single vendor’s hardware.

Leveraging Deduplication to Protect Remote Office Data

An increasing portion of business data now resides in remote and branch offices. Traditional methods of protecting remote office data using tapes just do not work anymore. There are always issues with tapes at remote sites: people forget to change tapes, tapes get lost in transit and often local staff do not have the expertise to manage tape systems. As the amount of business critical data in remote offices grows, it is imperative to provide a central management framework so the data can be available for recovery at a moment’s notice. To facilitate this, some businesses are eliminating tape at remote sites and moving all remote backups to disk. But this still leaves the problem of centralizing all of the remote office data within a single site. Another challenge companies face in centralizing data via remote office backup is often the limited amount of bandwidth over the WAN. There is simply not enough bandwidth to move all the backups in the available time.

Deduplication can play a vital role in protecting and centralizing remote office data. CommVault® Simpana® software’s singular approach allows you to protect remote office data in the same way you would in the data center, from a central console, using a consistent set of policies. Deduplication helps reduce the backup footprint at the remote site by writing only changed data to disk. It is much easier to move this changed data over the WAN versus the full backup set. Simpana® software offers two solutions that leverage deduplication to protect remote offices and centralize data over the WAN.
**Scenario 1**: Remote Office Data Protection, featuring Client–Site Deduplication

In this solution a centralized, shared deduplicated secondary storage pool is presented as a network share to the Media Agent in the remote site in the upper lefthand side. Optionally, a local deduplicated disk pool can also be connected to a Remote Media Agent. Simpana® software protection and/or archive agents are deployed on remote office servers. Backup data is written to the local disk pool in deduplicated form. By enabling the inline option on the dedupe policy, a second copy of deduplicated data is concurrently sent over the network share to the central storage pool. Since the CommVault data mover-based deduplication process eliminates redundant data segments in memory and writes only unique segments to the backend, only these segments are sent over the WAN to the central site. Note that this happens whether you are performing full or incremental backups. The data in the central site is immediately available for secondary processing and for migration/archival to tape.

**Figure 2**: Remote Office Data Protection Featuring Client–Site Deduplication
Scenario 2: Remote Office Data Protection—Featuring Replicated Disk Library with Deduplication

In this scenario, CommVault® backup or archive agents are installed on remote office servers that are configured to backup/archive data to a remote Media Agent with local deduplicated disk store. In the central site another Media Agent is configured with its own deduplicated disk store. A replicated disk library is configured between the two disk stores. CommVault Data Replicator agents are deployed on each of the Media Agents. When backup/archive operations are executed on the remote servers, the remote Media Agent writes only new or changed data segments to disk. Once the data is committed, the data replicator node transfers only the changed data segments over the WAN to the disk store in the central site. Once replicated, the data at the central site is instantly available for secondary processing at the central site via the central Media Agent. This is similar to using replication between two deduplication appliances. However, CommVault’s approach has the advantage that data is readily available at the central site. There is no need to transfer or restore catalogs on the central site to access the data. In the time it takes to replicate the changed data over the WAN, a DR copy is created and ready to use without any additional steps.

Figure 3: Remote Office Data Protection Featuring Replicated Disk Library with Deduplication
Embedded Global Data Deduplication for Enhanced Storage Cost-Efficiency

CommVault extends the power of Simpana® with hardware-independent, content-aware deduplication. CommVault deduplication software eliminates all duplicate copies of backup and archive data managed within the guidelines of your Global Storage Policy, regardless of source, data type or platform. We do this using our content-aware and adjustable segment-based approach that delivers better deduplication ratios compared to other block-based approaches. We also compare the digital signature of each segment against all the other segments of data in your global storage policy, maximizing deduplication across 100 percent of your backup and archive data. Combining client-based generation of hash signatures with integrated data compression and data encryption capabilities minimizes network bandwidth requirements and maximizes end-to-end network speed and security that goes from clients all the way to disk and tape-based storage tiers.

With Simpana, you can expand the use of data deduplication across all tiers of storage, including tape, without the need to “rehydrate” the data back into full form. Using our automated, policy-based approach, you can implement a hands-free approach to migrate and manage aging data quickly and efficiently. The migration is invisible to end-users, and ensures that the deduplicated data is brought back seamlessly directly from any storage tier. CommVault also combines our Shared Index/Catalog with our unique SILO storage to track and retrieve just the data required, without the need for lengthy restores and searching through entire volumes of data.

Deduplication With Simpana is Especially Helpful for:

- End-to-end elimination of duplicate data across multiple backup cycles, and across backup and archive data stores → Greater deduplication ratios.
- Direct data restores from any storage tier and the ability to create and retain dramatically more recovery points → Faster recovery of data covering longer retention periods.
- Automated policy-based migration/management of long-term data on tape-based backup media → Better manageability with optimized storage/tiering of data.

CommVault Simpana integrates deduplication with our broader platform capabilities around Backup and Recovery, Archiving, Replication, SRM and Search to deliver a holistic integrated deduplication approach that maximizes efficiency and minimizes barriers to implementation throughout the enterprise.
## Key Features and Benefits

<table>
<thead>
<tr>
<th>Features</th>
<th>Benefits</th>
</tr>
</thead>
</table>
| **“Open Disk” deduplication** | **Reduce hardware lock-in:**  
  - Treat “disk like disk” to unleash the full random access performance, concurrent I/O and resiliency capabilities of your disk investments—reduce barriers to entry by integrating with existing customer storage assets. Minimizes new hardware spend.  
  - Employ advanced disk sharing options to meet enterprise requirements and increase the scope of the global deduplication pool over a consolidated set of storage resources—reducing hardware spend. |
| **Global deduplication policies** | **Minimize storage resource investment:**  
  - Data is deduplicated at the global storage policy level consolidating backup and archive copies across different clients, operating systems, applications and multiple sites.  
  - Utilizing content-aware segmentation, deduplication reductions of up to 50X can be achieved as compared to traditional tape-based protection methods. |
| **Data Mover processing for in-memory deduplication** | **Maximize network efficiency:**  
  - Client-side compression and signature generation maximize network throughput performance, while in-memory processing on the data mover ensures only unique segments are written to the disk storage targets including WAN connected clients.  
  - Embedded into the backup or archive job to only process and hash items that are relative to the job (full, incremental), which greatly reduces the operational load placed on the client system. |
| **Increased number of recovery points** | **Service faster recovery objectives:**  
  - Maximize the benefits of disk pool investment by retaining more near-line recovery points while preserving the fastest recovery time performance.  
  - Distribute your deduplicated copies across sites to afford cheaper and faster recovery, while supporting secondary needs such as copy-to-tape or content indexing/search policies. |
| **Tiered migration to offline media** | **Lowest cost of ownership over the entire data lifecycle:**  
  - Extend without expanding; extend the deduplicated benefit directly to tiered media to lower cost media to contain and reduce the total cost of ownership over longer data lifecycles. Unlike traditional copy methods that expand the data to full form, the deduplication state is preserved on all the migrated instances. |
| **Redundancy and availability** | **Proactive protection against loss of business-critical data:**  
  - Employ inline dual site copies to produce two deduplicated copies with independent retention within the same backup/archive job. Reducing risk of loss from hardware or site failure. This is best aligned for remote office protection scenarios.  
  - Dual copies—in-line or replicated—can provide immediate data availability in the event of a loss in the primary copy while significantly offsetting the cost associated with traditional tape based DR. |
| **Centralized management** | **Simplified operations:**  
  - As an embedded policy capability in our data mover and library management sub-system, we are significantly reducing the complexity of configuring and using deduplication for data protection and archive retention. Our system automatically manages the data aging process and monitors media usage through the library threshold and reporting controls.  
  - Distributed resources managed through central control; employ remote site distributed protection and archiving combined with enterprise-class data center capabilities. Standard disk library features—capacity threshold policies, load-balancing across mount-paths, alerts, logging and reports—apply directly to your deduplication solution unburdening the operations team. |
Additional CommVault® Simpana® Solutions

CommVault® Simpana® software offers data management software built on the same software platform so that all capabilities work together. This makes it easy to add capabilities when and how you need them. For file system and database applications, consider:

▶ Backup
- Object-level backup and recovery of file systems, applications, databases and virtual server systems
- Reliable operation with failover resume and job completion capabilities

▶ Archive
- Reduced cost and growth of primary storage systems
- Legal discovery and compliance readiness

▶ Replication
- Remote office data centralization as a seamless extension of backup
- Disaster recovery flexibility

▶ Search
- Information Access and eDiscovery of Electronically Stored Information (ESI)
- Simplify, speed and reduce the cost of enterprise-wide discovery from a single console

▶ Resource Management
- Analysis of primary storage growth requirements
- Analysis of secondary storage growth requirements

▶ SnapBackup
- Near Instant™ snapshot backups across heterogeneous systems and storage
- Eliminates backup impact on production servers

▶ Virtual Server Data Protection
- Flexible backup and recovery options for VMware® and Microsoft® virtual platforms
- Agent-less backup of virtual machines

▶ Remote Office and Branch Office Data Protection
- Leverage multiple options for protecting remote office data based on RTO/RPO and recovery SLAs
- Optimize disk and network utilization with client-side deduplication

▶ Workstation Backup and Recovery Solution
- Centralizes data protection to data residing at the edge on end user desktops and laptops
- Ensures preservation of end-user files and data in efficient, searchable, long-term archives

System Requirements

Deduplication - Windows® and UNIX® Media Agents

Deduplication Backup Copies - Mac® File System, Microsoft® DPM, Microsoft Exchange Mailbox, Microsoft SharePoint® Documents, Microsoft Windows File System, OES File System, ProxyHost, UNIX File System

Deduplication Archive Copies - Domino® Mailbox, Microsoft Exchange Compliance, Microsoft Exchange Mailbox/Public Folder, UNIX File System, Microsoft Windows File System, NAS Files, Microsoft SharePoint

Note: System Requirements are subject to change. For up-to-date System Requirements, please refer to product documentation on the CommVault website at http://www.commvault.com