COMMVAULT®

TECHNICAL GUIDE

CMI Remediation Procedure

This document outlines how protected data managed in the Commvault Content Store can be easily searched, deleted and securely sanitized to support sensitive data and Classified Message Incident actions.

Technical field guide has been published March 30, 2022 – Version-2 - Field Guide

Automating remediation of CMI incidents with the Commvault Content Store

A Classified Message Incident (CMI) or "data spill" occurs when files or data instances are inadvertently moved to an unclassified network. Classified information may also be transmitted through other forms of file transfer, including web browser downloads, files transferred through tethered connections, or backup and archived copies of the source system.

Federal government agencies, especially DOD, maintain processes to search and remove the offending data and sanitize the storage networks.

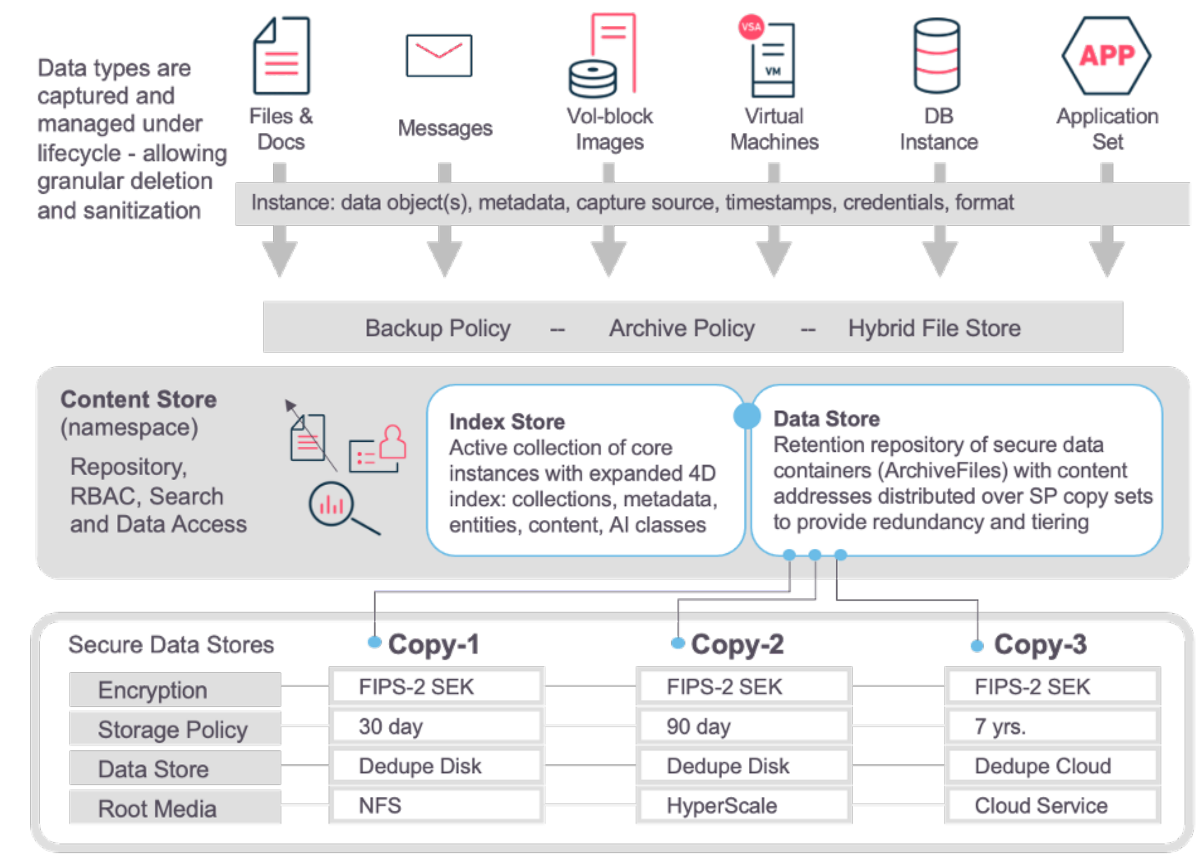
Two scenarios require investigation, search, and remediation:

1. Primary data sources or production networks where virtualization, applications, and file servers access users. This could include replicated copies and cloud in some cases.
2. Data protection copies may contain offending data, including off-site replication or long-term retention copies and any cloud data part of the data management protection.

A common challenge with many legacy data protection products or limited backup tools that cannot manage the data at a granular instance level is their inability to support a fast and intuitive search with a matched ability to delete and purge data instances from the repository to close CMI incidents. This inadequacy can shift an escalating burden onto the administrator team to force compliance issues if the wholesale deletion of the entire backup set of millions of files is due to a small scattering of sensitive CMI files. This commonly causes the team to execute extended projects that demand the sequential restoration of extensive data collections to support indexing and search with 3rd party tools. Suppose those tools can identify the affected instances. In that case, the team will begin the intensive tasks of manual deletion followed by a re-backup of the remainder of the data set across multiple periods (backup recovery points) to ensure the "good" data is re-preserved for the required periods. These efforts expose increasing risks for mistakes as the scope and frequency of the incidents increase. Using disparate tools lacks comprehensive audit trails to ensure the process is complete, repeatable, and auditable.

Commvault architecture is built on a foundation that offers comprehensive data management services driven from a consolidated namespace/repository that centralizes search and discovery to quickly support CMI incidents and confirm if the matched instances exist in the repository. Under an authorized role-based access control process, the Administrator can permanently delete the selected data instances from the Content Store, automatically deleting the cases across all Storage Policy copies. If the CMI event demands media sanitization, the appropriate process can be applied based on the media type.

Commvault Protected Content: managed instance types


Data types are captured and managed under lifecycle - allowing granular deletion and sanitization
Files & Docs
Messages
Vol-block Images
VSA
VM
Virtual Machines
DB Instance
APP
Application Set
Instance: data object(s), metadata, capture source, timestamps, credentials, format
Backup Policy -- Archive Policy -- Hybrid File Store
Content Store (namespace)
Repository, RBAC, Search and Data Access
Index Store
Active collection of core instances with expanded 4D index: collections, metadata, entities, content, AI classes
Data Store
Retention repository of secure data containers (ArchiveFiles) with content addresses distributed over SP copy sets to provide redundancy and tiering

Secure Data Stores	Copy-1	Copy-2	Copy-3
Encryption	FIPS-2 SEK	FIPS-2 SEK	FIPS-2 SEK
Storage Policy	30 day	90 day	7 yrs.
Data Store	Dedupe Disk	Dedupe Disk	Dedupe Cloud
Root Media	NFS	HyperScale	Cloud Service

The critical elements of Commvault Content Store (repository) that aid in the CMI remediation process are:

1. **Index Store** – As data is copied and secured into the Content Store repository, the data instances are indexed and registered into the Commvault namespace. The namespace is powered by a multi-variant indexing service that powers the authorized search and browsing features to select relevant data instances for action. Actions can include restoring the data to the source application, downloading the selected data for sharing or export purposes, applying an extended retention hold of a data set, selecting critical data for a CMI event, and deleting all instances from the repository. An authorized user with the appropriate data ownership/rights can quickly find the sensitive data, cull the instances for specialized review, and permanently delete those instances from the Content Store.

2. **Data Store** – The backup job will create a logical Archive File (AF) to contain the copied indexed data instances. The AF acts as a storage container that manages the data instances and source metadata under copy lifecycle retention rules. The index store maintains individual content addresses for each data instance to allow rapid granular retrieval. A Storage Policy (SP) manages the replicated copies of the Archive Files stored in different media locations (datastores). The Content Store repository contains all the policies, controls, data movements, access, and lifecycle actions.

   - Typically, users will configure the storage policy to compress, deduplicate, and encrypt the data stream sent to the repository via the backup procedure. The deduplication process creates segmented block signatures of the data instance to identify and eliminate redundant segments and reduce the storage footprint as the data instance is written to a deduplicated data store. The format of the dedupe store obfuscates the stored cases in a secure and compact design, compared to a non-deduplicated backup format which will compact and concatenate the native instance file streams in a large container file. A dedupe store is logical, secure data collection that is physically maintained on a disk library/file system or cloud library/object-store.

3. **Media Sanitization**– When data instances are permanently deleted, the deletion process removes the critical index records (i.e., metadata, keys, and AF content addresses) to erase the instance from the logical namespace. When the data instances are stored inside encrypted datastores, this produces a "crypto-shredding" outcome as the keys and indexed addresses are

erased, rendering the encrypted data segments permanently orphaned with no means of recovery. When the AF contents are FIPS encrypted, it will eliminate the potential of any forensic recovery methods to reconstruct the data instance into the comprehendible native form.

- The data aging/pruning process is used to expire and permanently purge the relevant data segments from the data store/root media. The is action carries different dependencies based on the root media type (disk, cloud, tape) combined with WORM-file locks or FS storage snapshots which may delay the actual purging of the deleted data segments. A disk data store consists of mount path(s) to a root file service. Commvault will delete physical file objects using the data pruning or data compaction (data validation) Content Store process.
- The root media file service will need to execute any FS clean-up processes to flush the root media blocks from the deleted file objects. After that process is completed, the Content Store Administrator can run the CV disk sanitizer tool against the data store/mount path to digitally shred the root media free-block pool using one of the most common industry scrubbing methodologies.

When data is sent to the Content Store, how is it registered in the index?

1 Data is captured and secured in the Commvault Content Store via the backup or archive process; the data is indexed based on the type/instance. File-based backup policies will capture the changed files and document instances and index the source metadata and version into the central catalog. VM backup policies manage the instance at a VM level. Email captured from mailbox-based policies will align the managed instance at the message object.

2 DB backup policies generally orient the instance as the DB collection, including multiple small DBs managed as a single backup set. While CV offers restore-based granular table and record-level selection options, data instance is based on the set for retention and deletion actions.

How is the data instance secured into the storage copies?

1 When new data is copied by the backup job ("streaming"), it will create a logical data container known as an Archive File (AF) for each backup job. This container is used to pack the data instances collected by the job into a compact, secure format preserving the native data elements and metadata under the copy lifecycle retention policy.

2 The backup process will employ the Commvault deduplication methods to compress the stream's contents into block segments with corresponding SHA-512 digital identity signatures. That digital signature is checked into the dedupe database that manages the destination dedupe store. The dedupe process is embedded into the Commvault storage services to reduce the amount of data transferred and the actual storage footprint of the dedupe store library.

3 When the process finds an identical segment already managed on the store, the process will create an indexed link to the active component as part of the Archive File content address. This eliminates the transmission of the actual data segment payload, typically 128KB. As additional standard segments are found, when new data versions, backup instances, or similar data types are written in the store, the deduplication process yields faster backups and more efficient storage capacity consumption.

4 The indexed links and unique segment payloads (a.k.a. shards) are physically stored on the dedupe store library in a containerized format (SFILE, 64MB) which concatenates and compacts many segments/shards into a physical file to optimize file operations and performance. The segments are randomized by the concurrency of the streams and processes, providing a foundation layer of secure data obfuscation relative to the file instances protected in the backup job. Data encryption is also applied to the backup process and retained across all stored copies; all data segments exist in an encrypted state. The data encryption generates dynamic keys inline during the streaming process within each ArchiveFile (job). This results in securing the datastore from all forensic recovery methods.

- For example, if file ABC was segmented into three shards A-B-C, and the identical file was backed up ten times, then the collection will reduce and link to the three unique data shards [A][B][C], which reside on the dedupe store; the first backup job will store the individual shards. The following nine backup jobs will contain the repeated signatures and create new links to the same stored bits avoiding new transfers and storing the redundant data shards in the same store.

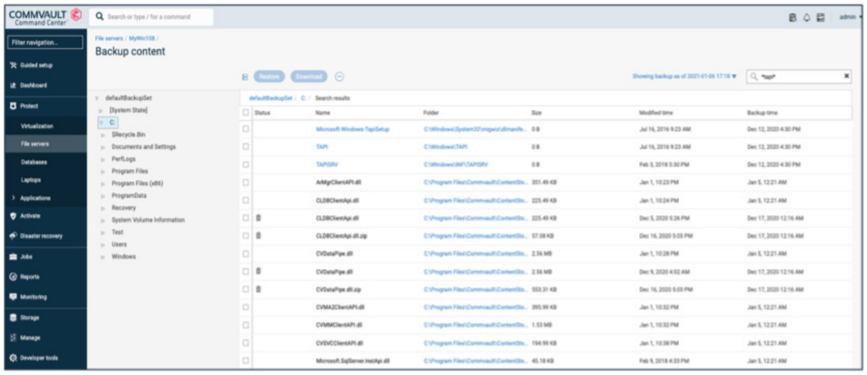
5 If the same file ABC was changed slightly to ABD, the next backup job will add links to the existing [A][B] shards, while the new unique data shard [D] is transmitted over the network and written to the store.

6 In subsequent jobs, if the file changes to AED, the common shards [A] and [D] add more links, while the new fragment [E] is stored to link to the new segment.

Now, what happens when the user permanently deletes a data instance?

1 The user will search or explore (browse) a data collection managed in the Content Store to select the instance/version they want to restore, download, or delete.[1] The user is presented with the logical source data view, which is further trimmed based on the user's RBAC Content Store access rights. The statement provides a point-in-time collection of the folder/file contents or mailbox /folder message list.[2] If the user has been authorized to use the data deletion role, they apply that action against the selected data set. This initiates the data deletion process.

Search across the file namespace on managed versions (active, hidden, deleted) which can be expanded with content (CI options)



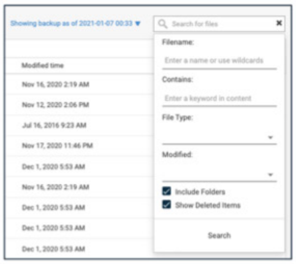
File servers / MyWin108 /

Backup content

Showing backup as of 2021-01-06 17:18 ▾ *tapi*

defaultBackupSet / C: / Search results

□	Status	Name	Folder	Size	Modified time	Backup time
□		Microsoft-Windows-TapiSetup	C:\Windows\System32\migwiz\dlmanife...	0 B	Jul 16, 2016 9:23 AM	Dec 12, 2020 4:30 PM
□		TAPI	C:\Windows\TAPI	0 B	Jul 16, 2016 9:23 AM	Dec 12, 2020 4:30 PM
□		TAPISRV	C:\Windows\INF\TAPISRV	0 B	Feb 3, 2018 5:30 PM	Dec 12, 2020 4:30 PM
□		ArMgrClientAPI.dll	C:\Program Files\Commvault\ContentSto...	351.49 KB	Jan 1, 10:23 PM	Jan 5, 12:21 AM
□		CLDBClientApi.dll	C:\Program Files\Commvault\ContentSto...	225.49 KB	Jan 1, 10:24 PM	Jan 5, 12:21 AM
□	🗑	CLDBClientApi.dll	C:\Program Files\Commvault\ContentSto...	225.49 KB	Dec 5, 2020 5:26 PM	Dec 17, 2020 12:16 AM
□	🗑	CLDBClientApi.dll.zip	C:\Program Files\Commvault\ContentSto...	57.08 KB	Dec 16, 2020 5:03 PM	Dec 17, 2020 12:16 AM
□		CVDataPipe.dll	C:\Program Files\Commvault\ContentSto...	2.56 MB	Jan 1, 10:28 PM	Jan 5, 12:21 AM
□	🗑	CVDataPipe.dll	C:\Program Files\Commvault\ContentSto...	2.56 MB	Dec 9, 2020 4:02 AM	Dec 17, 2020 12:16 AM
□	🗑	CVDataPipe.dll.zip	C:\Program Files\Commvault\ContentSto...	553.31 KB	Dec 16, 2020 5:03 PM	Dec 17, 2020 12:16 AM
□		CVMA2ClientAPI.dll	C:\Program Files\Commvault\ContentSto...	395.99 KB	Jan 1, 10:32 PM	Jan 5, 12:21 AM
□		CVMMClientAPI.dll	C:\Program Files\Commvault\ContentSto...	1.53 MB	Jan 1, 10:32 PM	Jan 5, 12:21 AM
□		CVSVCClientAPI.dll	C:\Program Files\Commvault\ContentSto...	194.99 KB	Jan 1, 10:38 PM	Jan 5, 12:21 AM
□		Microsoft.SqlServer.InstApi.dll	C:\Program Files\Commvault\ContentSto...	45.18 KB	Feb 9, 2018 4:33 PM	Jan 5, 12:21 AM



Showing backup as of 2021-01-07 00:33 ▾

Modified time
Nov 16, 2020 2:19 AM
Nov 12, 2020 2:06 PM
Jul 16, 2016 9:23 AM
Nov 17, 2020 11:46 PM
Dec 1, 2020 5:53 AM
Nov 16, 2020 2:19 AM
Dec 1, 2020 5:53 AM
Dec 1, 2020 5:53 AM
Dec 1, 2020 5:53 AM

Filename: Enter a name or use wildcards

Contains: Enter a keyword in content

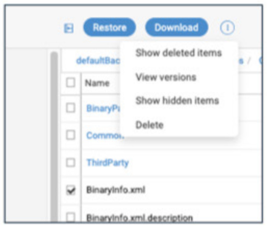
File Type:

Modified:

☑ Include Folders

☑ Show Deleted Items

Selected items can be restored back to active clients, downloaded to export for further investigation by the CISO/InfoSec team or partner or deleted (surgical delete) from the backup collections to completely purge the "bad" data

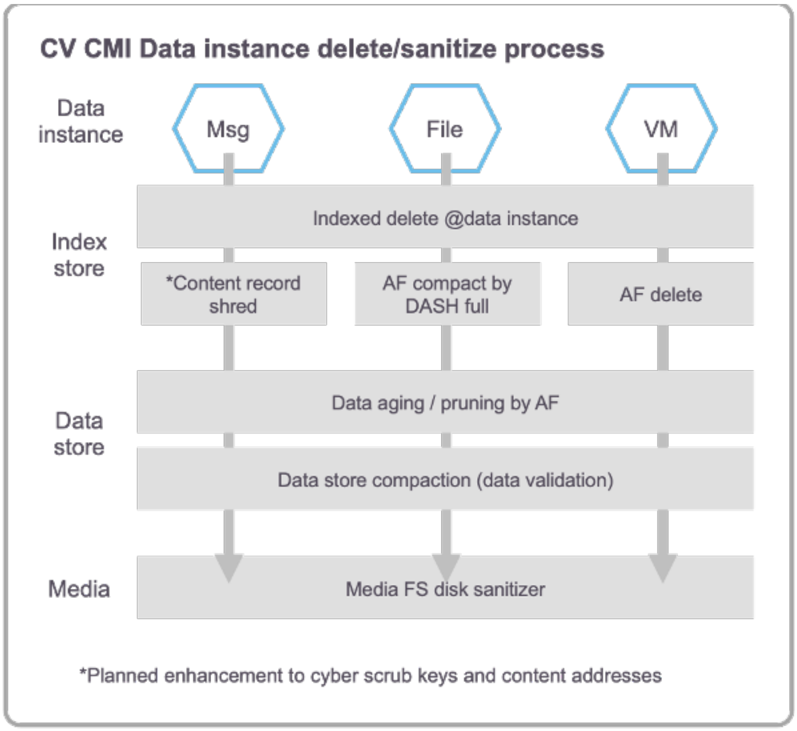


□	Name
□	BinaryP
□	Commo
□	ThirdParty
☑	BinaryInfo.xml
□	BinaryInfo.xml.description

- Data deletion action will immediately erase the data instance from the active namespace as the Content Store. The deletion process is permanent, and there is no roll-back or undo- actions. The data deletion role should only be assigned to authorized users who need to execute deletion tasks against sensitive data instances.

[1] https://documentation.commvault.com/11.25/expert/146291_deleting_backup_data_and_archive_data_from_command_center.html

[2] https://documentation.commvault.com/11.25/essential/104085_deleting_mailbox_items_folders_or_messages.html


CV CMI Data instance delete/sanitize process
Data instance
Msg
File
VM
Index store
Indexed delete @data instance
*Content record shred
AF compact by DASH full
AF delete
Data store
Data aging / pruning by AF
Data store compaction (data validation)
Media
Media FS disk sanitizer

*Planned enhancement to cyber scrub keys and content addresses

- When a set of email messages are deleted, we will apply a new automatic record-shredding process (Targeted -Near-term roadmap, 2H 2022) to digitally shred the index keys and content addresses in the index store to permanently remove any option to retrieve the message instance from the data stores.
- If the data instance is securely stored in an encrypted data store, then the destruction of the keys produces a crypto-shredding outcome for the data instance. This case would apply to email message data instances or digital records combined with encrypted data stores. Based on the retention policy, when the AF containing the message collection expires, the data pruning process will delete the stored segments and trigger the physical flush/delete of the media stores, permanently removing the contents from the data stores.
- When a set of files is deleted, the user will need to run the synthetic full or DASH full job[3] after the deletion task to create a new replacement AF containing the remaining active file data instances. Collection-oriented data instance types (i.e., files, documents) require this additional step before running the data pruning operations to ensure that deleted content stored segments and index records are physically flushed/deleted from the media stores. This permanently removes the contents from the data stores.
- When an individual VM is deleted, the related AFs will be marked for deletion and purged with the data aging/pruning process to remove the contents from the data stores permanently.

2 The data deletion process will automatically apply the aging process to mark and expire the active links to the data segments that comprise the data instance within the data store copies. Data aging typically runs on a scheduled basis, or the user can run it immediately from the console.[4]

- The data aging process determines which jobs have completed the mandatory retention lifecycle rules and should be expired to trigger the data pruning service to purge the stored instances from the root media physically. The process will resolve and expire the relevant linkage into the storage data stores — Archive Files, instances, and copies to remove the data instance reservations. When the data resides in a dedupe store, the aging/pruning will automatically remove the links to any relevant segments. If all active links have expired on a data segment, then that data segment will be marked as deleted from the data store.
- In our previous example, if the ten instances of the ABC file were deleted from the Content Store, but the ABD and AED instances remain on retention, then the system will retain the [A] [B] segments as they are actively linked. The related dedupe shard will be marked for the purge operation when all links are expired.

3 The Data Pruning service runs every hour to continuously discover expired elements to purge the raw media store. The service runs across all copies. This operation will process the physical or logical deletion of the dead data segments from the root media type. The actual execution method can differ based on the media type.

- The data pruning process will delete the entire expired container media file or mark expired data ranges inside the container file relative to deleted content. Progressive pruning operations will expire other fields until the whole file container is expired, which triggers the flush or physical deletion of the container file on the media store.

[3] https://documentation.commvault.com/11.25/expert/11694_synthetic_full_backups.html
[4] https://documentation.commvault.com/11.25/expert/11917_running_data_aging_job.html

- A subset of file system media stores, such as Windows NTFS, supports sparse files, which allow a prescribed block range to be directly deleted inside an existing physical file. The Commvault storage system can automatically use that feature to delete segments and immediately trigger the release of the deleted blocks into the media available block pool. This is commonly referred to as "drilling holes" in Commvault operations. Due to the limited number of sparse-supported file systems that can be used as media stores – we recommend users employ the data compaction jobs as the immediate next step in a CMI incident.
- A scheduled administrative feature for disk libraries will run the Data Validation-Space Reclamation job[5] to remove the deleted segments or orphaned files from the dedupe store. This method also applies to object-based cloud libraries. We recommend running the task after the data pruning operations have been completed. Use the level (4) setting for the aggressive reclamation level and ensure the clean orphan data option is engaged.

As a final step, the CVDiskEraser[6] tool can sanitize the FS media store.

1 Certain sensitive CMI events may require a final disk sanitization operation against the FS media store to digitally sanitize the root file system. After all the previous steps have concluded, this last task can be applied to the FS media store to force a sanitization across the free blocks (un-used space) to eliminate the ability to use forensic tools to recover any deleted block from the source FS media store.

2 Configure and execute the Physical Pruning Status Check workflow[7] to confirm and verify that data from the specified clients has been physically pruned during the data aging process. The workflow will produce a Physical Pruning Status report detailing the client mount paths ready for sanitization.

3 Run the CVDiskEraser tool on these mount paths. The device supports the five standard data erasure methods.

- USD - (Default) the United States Defense Department – This method will clear the free block pool by writing any bit pattern to the entire disk in one pass. The free space will be erased by writing a different bit pattern to the disk in each of three passes. This method takes approximately 4 hours or less to sanitize 1 TB of free space using a standard 64KB buffer size.
- BSI - German BSI Verschlusssachen-IT-Richtlinie – This method will erase the drive with seven passes. For the first six passes, each erasure reverses the bit pattern of the previous erasure. The final pass overwrites the entire disk with the bit pattern "01010101". This standard is commonly considered the most secure method of erasing data. This method takes approximately 6 hours or less to sanitize 1 TB of free space using a standard 64KB buffer size.
- SCH - Bruce Schneier's Algorithm - This method will apply the first pass overwrite with the bit pattern 11, the second pass with 00, and the next five with a randomly generated bit pattern. This method takes approximately 12 hours or less to sanitize 1 TB of free space using a standard 64KB buffer size.
- GUT - Peter Gutmann's Algorithm - This standard erases the data by writing a series of 35 patterns over the region to be erased. The selection of patterns assumes that the user does not know the encoding mechanism used by the drive, so it includes patterns explicitly designed for three types of drives. A user who knows which type of encoding the drive uses can choose only those patterns intended for that drive. An industry with a different encoding mechanism would need different patterns. This method takes approximately 16 hours or less to sanitize 1 TB of free space using a standard 64 KB buffer size.
- DSX - Royal Canadian Mounted Police DSX Method - This standard writes the bit pattern 00 on the first pass, 11 on the second, and a text pattern that consists of the software version number and the date and time the erasure occurred. This method takes approximately 3 hours or less to sanitize 1 TB of free space using a standard 64KB buffer size.

[5] https://documentation.commvault.com/11.25/expert/128883_space_reclamation_online_help.html

[6] https://documentation.commvault.com/11.25/expert/11475_disk_eraser_tool.html

[7] https://documentation.commvault.com/11.25/expert/11477_workflow_physical_pruning_status_check.html