IBM Spectrum Protect

Effective Planning and Use of IBM Spectrum Protect Container Storage Pools and Data Deduplication

Document version 3.1

Authors:
Jason Basler
Austen Cook
Dan Wolfe

Document Location

This document is a snapshot of an online document. Paper copies are valid only on the day they are printed. The document is stored at the following location:
http://ibm.biz/ContainerPoolBestPractices

Revision History

Revision Number	Revision Date	Summary of Changes
1.0	08/17/12	Initial publication
1.1	08/31/12	Clarification on data deduplication requires backup option and other minor edits
1.2	06/10/13	General updates on best practices
1.3	06/27/13	Add information for data deduplication of Exchange data
2.0	12/09/13	Major revision to reflect scalability and best practice improvements that are provided by TSM 6.3.4.200 and 7.1.0
2.1	02/09/15	Revised to include references to TSM Blueprints and Solutions
3.1	03/31/17	Major revision for IBM Spectrum Protect container pool data deduplication

Disclaimer

Acknowledgements

The authors would like to express their gratitude to the following people for contributions in the form of adding content, editing, and providing insight into IBM Spectrum Protect technology.

Matt Anglin, IBM Spectrum Protect Server Development
Dave Cannon, IBM Spectrum Protect Architect
Colin Dawson, IBM Spectrum Protect Architect
Robert Elder, IBM Spectrum Protect Performance Evaluation
Tom Hughes, Executive; WW Storage Software
Kathy Mitton, IBM Spectrum Protect Server Development
Harley Puckett, IBM Spectrum Protect Software Development - Executive Consultant
Michael Sisco, IBM Spectrum Protect Server Development
Richard Spurlock, CEO and Founder, Cobalt Iron

Table of Contents

1 Introduction

Container storage pools are a new type of storage pool that combine multiple inline data reduction technologies. Data reduction can occur at either the source (client), inline at the server, or in combination during client ingest. Data deduplication and compression technology work together to provide an effective overall storage savings referred to as data reduction throughout this paper. Because of this inline data reduction, container storage pool I/O is dramatically reduced, enabling the use of lower-cost disk technology. Inline data deduplication and compression eliminate the need for post processing tasks, improve daily client ingest capacity, and simplify the management of the IBM Spectrum Protect server. This document describes the benefits of container storage pools and data deduplication, and provides guidance on how to make effective use of them as part of a well-designed data protection solution. Many of the recommendations throughout this document assume that you are running the latest version of the IBM Spectrum Protect server and client.

The following are key points for IBM Spectrum Protect with a container storage pool:

- A container storage pool with inline data deduplication is an effective tool for reducing overall cost of a backup solution
- Two types of container storage pools can be configured: directory-container and cloud-container storage pools. Directory-container pools are configured with block or network-attached storage. Cloud-container storage pools use either on-premises or off-premises cloud Object Storage.
- With a container storage pool, more resources (DB capacity, CPU, and memory) must be configured for the IBM Spectrum Protect server. However, when properly configured, the benefits of a reduction in storage pool capacity, storage pool I/O, and replication bandwidth requirements are likely to result in a significant overall cost reduction benefit.
- Cost savings is the result of the total data reduction. Container storage pool data deduplication is just one of several methods that IBM Spectrum Protect uses to provide data reduction. Overall data reduction is maximized when deduplication is combined with compression and progressive incremental backup data technologies.
- In addition to data deduplication, a container storage pool uses LZ4 compression for additional data reduction. LZ4 provides a high-performance, low-overhead compression algorithm that is optimal for backup and restore workloads.
- Container storage pool data deduplication and compression can operate on backup, archive, and HSM data, and data that is stored via the IBM Spectrum Protect client API.

IBM Spectrum Protect provides significant flexibility, where a container storage pool can be one of several different storage pool technologies in a solution. Data can be directed to the most cost effective storage pool type based on specific client requirements.

1.1 Scope of this document

This document is intended to provide the technical background, general guidelines, and best practices for using IBM Spectrum Protect container storage pool and data deduplication technologies. Detailed information for configuring servers with container storage pools is provided in the IBM Spectrum Protect Blueprints. It is recommended to use the blueprint documentation and scripts to configure an IBM Spectrum Protect server: http://ibm.biz/IBMSpectrumProtectBlueprints

In addition to the blueprints, guidelines for selecting an IBM Spectrum Protect architecture are documented in the IBM Spectrum Protect Solutions: https://ibm.biz/IBMSpectrumProtectSolutions

This document does not provide comprehensive instruction and guidance for the administration of IBM Spectrum Protect, and must be used in addition to the IBM Spectrum Protect product documentation.

1.2 Overview

1.2.1 Description of container storage pool technology

A container storage pool is designed specifically for data deduplication. Data reduction of backup data is accomplished at the client, inline at the server, or a combination of both. The data goes through all the reduction processing before it is written out to containers on the storage pool disk.

Two types of container storage pools can be configured, directory and cloud based. Directory-container storage pools use block or network-attached storage that is assigned to the storage pool by using the DEFINE STGPOOLDIRECTORY command. Cloud-container storage pools are defined to either on-premises or off-premises Object Storage.

Container storage pools combine the best features from DISK and sequential FILE storage pools. Challenges of sequential file storage pools are avoided with container storage pools. These challenges include management of scratch volumes, identify duplicates processing to identify duplicate data, and reclamation to eliminate empty space. In addition, container storage pools provide a robust audit and repair mechanism that is superior to what is available with a FILE storage pool. Container storage pools tend to follow a large sequential I/O pattern like a FILE storage pool, but can write to empty regions of existing containers randomly like a DISK storage pool. The ability to reuse empty regions avoids the need for expensive reclamation processing.

1.2.1.1 IBM Spectrum Protect container storage pool data deduplication use compared with other data deduplication approaches

Data deduplication technology of any sort requires CPU and memory resources to detect and replace duplicate chunks of data. Software-defined technologies such as IBM Spectrum Protect container data deduplication create similar outcomes to hardware based or appliance technologies.

IBM Spectrum Protect is a software-defined data deduplication solution, so the need to procure specialized and comparatively expensive dedicated hardware is negated. With IBM Spectrum Protect, standard hardware components such as server and storage can be used. Because IBM Spectrum Protect has significant data efficiencies compared to other software-based data deduplication technologies, less duplicate data needs to be removed. Therefore, all other things being equal, IBM Spectrum Protect

requires less standard hardware resources to function compared to other software-based data deduplication technologies.

Care must be taken in planning and implementing data deduplication technology. Under most use cases IBM Spectrum Protect provides a viable proven technical platform. In some use cases, a VTL can provide an appropriate architectural solution. Refer to the following conditions:

Use a VTL when requirements include:

- Managing greater than 4 PB of protected data with a single IBM Spectrum Protect Instance
- Supporting greater than 100 TB of client ingest per day
- Deduplicating NDMP data
- Deduplicating across multiple IBM Spectrum Protect Servers
- LAN-free backup for SAN environments

Use IBM Spectrum Protect container data deduplication when:

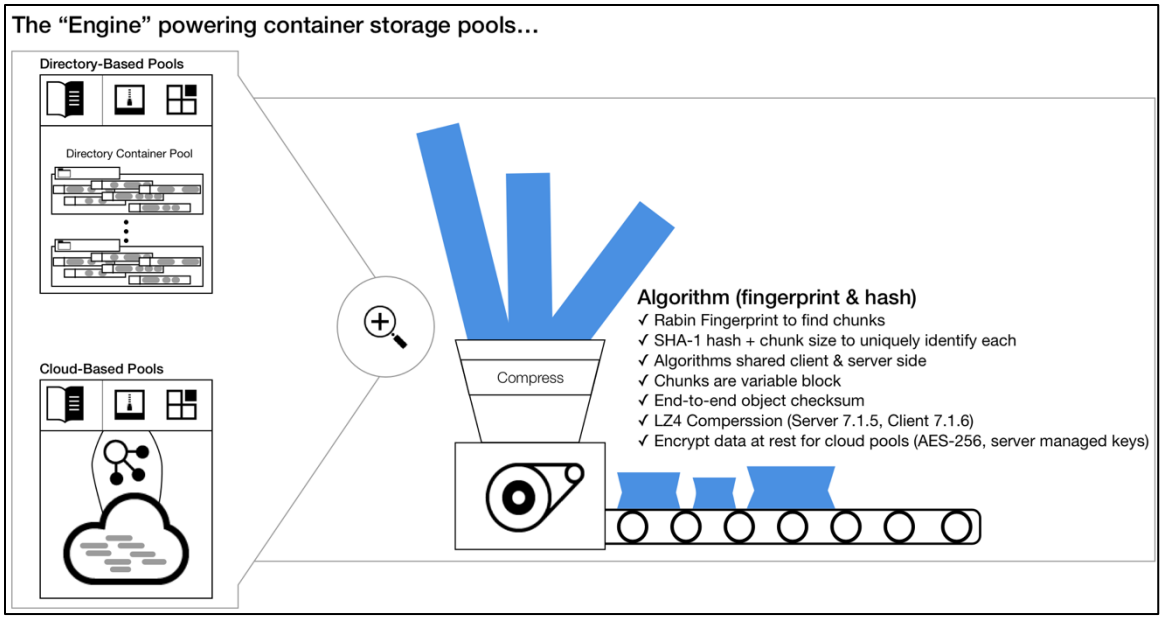
- Managing up to 4 PB of protected data
- Network utilization needs to be minimized by using client-side data deduplication
- Sufficient IBM Spectrum Protect server resources are available for data deduplication and compression
- Desire for a hot-standby replication target and automated client failover

1.2.1.2 How does IBM Spectrum Protect perform data deduplication with container pools?

IBM Spectrum Protect container storage data deduplication uses an algorithm to analyze variable sized, contiguous segments of data, called “chunks” or “extents”, for patterns that are duplicated within the same container storage pool. Identical chunks of data reference the existing chunk with a smaller metadata pointer. As a result, for a pattern that appears multiple times within a collection of data, significant data reduction can be achieved. Unlike compression, data deduplication can take advantage of a pattern that occurs multiple times within a collection of data.

With compression, a single instance of a pattern is represented by a smaller amount of data that is used to algorithmically re-create the original data pattern. Compression cannot take advantage of common data patterns that recur throughout the collection of data, and this aspect significantly reduces the potential reduction capability. However, combining compression with data deduplication further reduces the amount of data storage beyond just one technique or the other. With a container storage pool, LZ4 compression is applied inline to data that is not eliminated by data deduplication for additional data reduction.

Data is organized in container storage pools in either deduplicated containers (.dcf files) or non-deduplicated containers (.ncf files). Data that cannot be deduplicated, such as encrypted data by the client system, is written to non-deduplicated containers. In addition, files that which are smaller than 2 KB are not processed by data deduplication or compression and are written to non-deduplicated containers.



1.2.2 Data reduction and data deduplication

It is important to consider other data reduction techniques that are available, when you are using data deduplication. Unlike other backup products, IBM Spectrum Protect provides a substantial advantage in data reduction through its native capability to back up data only once rather than create duplicate data by repeatedly backing up unchanged files and other data. Combined with incremental-forever backup technology, data deduplication, and compression, the overall data reduction effectiveness must be considered rather than just the reduction from data deduplication alone. Inherent efficiency that is combined with data deduplication, compression, exclusion of specified objects, and appropriate retention policies enables IBM Spectrum Protect container pools to provide highly effective data reduction. If reduction of storage and infrastructure costs is the goal, the focus must be on overall data reduction effectiveness, with data deduplication effectiveness as one component. The following table provides a summary of the data reduction technologies that IBM Spectrum Protect offers:

	Incremental forever	Data deduplication	LZ4 Compression
How is data reduction achieved?	Client only sends changed files, or changed blocks[1]	Eliminates redundant data chunks	Unique data chunks after data deduplication are compressed
Conserves network bandwidth?	Yes	When client-side data deduplication is used	When client-side compression is used
Data supported	Backup	Backup, archive, HSM, API	Backup, archive, HSM, API
Scope of data reduction	Files that do not change between backups on the same node	Redundant data from any data in the same storage pool	Redundant data within a unique chunk
Avoids storing identical files that are renamed, copied, or relocated on client node?	No	Yes	No
Removes redundant data for files from different client nodes?	No	Yes	No

1 IBM Spectrum Protect for Virtual Environments uses change block tracking for incremental forever backups

1.2.3 Server-side and client-side data deduplication

IBM Spectrum Protect provides two options for performing data deduplication: client-side and server-side data deduplication. Both client-side and server-side data deduplication use the same algorithm to identify redundant data, where the data deduplication processing occurs is different. A combination of both techniques can be used with the same storage pool where the most appropriate location for data deduplication is selected for each client.

1.1.1.1 Server-side data deduplication

With server-side data deduplication, all the processing of redundant data occurs inline on the IBM Spectrum Protect server. Server-side data deduplication is also called “target-side” data deduplication. The key characteristics of server-side data deduplication are:

- Duplicated data processing is incurred inline at server, but before it is written to the storage pool.
- The duplicate identification and compression processing consumes CPU resources at the server.

1.2.3.1 Client-side data deduplication

Client-side data deduplication processes the redundant data during the backup process on the system where the source data is located. The results of client data deduplication are the same as with server-side data deduplication. Data that is duplicated at the client requires only a small signature to be sent to the server. Client-side data deduplication immediately removes redundant data before it is sent to the server, and it can be especially effective when it is important to conserve bandwidth between the client and server. In some cases, client-side data deduplication has the potential to be more scalable than

server-side data deduplication. The increase in scalability is attributed to the reduced network demands and distribution of data deduplication processing. The following conditions must exist to effectively scale client-side data deduplication:

- Sufficient client CPU resource to perform the duplicate identification processing during backup.
- The combination of the IBM Spectrum Protect database and activity log running on SSD disk, and a high-bandwidth low-latency network between the clients and server.

1.2.3.2 Client data deduplication cache

Although it is necessary for the backup client to “check in” with the server to determine whether a chunk is unique or a duplicate, the amount of data transfer is small. The client must query the server for each chunk of data that is processed. The network activity that is associated with this query process can be reduced substantially by configuring a cache on the client. The cache allows previously discovered chunks on the client (during the backup session) to be identified without a query to the server. If multiple, concurrent client sessions are configured, a separate cache needs to be configured for each session. For applications that use the IBM Spectrum Protect API, the data deduplication cache must not be used due to the potential for backup failures that are caused by an out of sync cache with the Spectrum server.

Faster performance might be possible when the data deduplication cache is disabled. When the network between the clients and server has high bandwidth and low latency, and the server database is on fast storage, the data deduplication queries directly to the server can outperform queries to the local cache.

1.2.3.3 Use of compression with data deduplication

Further reduction of backup data can be achieved with LZ4 compression in addition to data deduplication. Like data deduplication, compression can occur either client-side or server-side. However, to use client-side LZ4 compression, client-side data deduplication must also be enabled. Backups with high data deduplication rates incur lower compression processing, as only unique data chunks are compressed. For optimal data reduction, the data deduplication process must occur before compression. Use of compression before data deduplication will most likely result in lower overall data reduction benefits. For example, the combination of client compression and server data deduplication must be avoided. The following data deduplication and compression combinations provide the optimal ordering for data reduction:

- Server data deduplication and Server LZ4 Compression
- Client data deduplication, Client LZ4 Compression
- Client data deduplication and Server LZ4 Compression

Although actual results are highly dependent upon the actual source data, the use of LZ4 compression can reduce the amount of backup data by an extra 25 - 80% above the reduction from data deduplication. Consider the following points with container storage pool compression:

- Compression requires an extra processing step, and therefore requires more CPU resource. Typically, the additional processing for compression is not an issue unless the CPU resource is already heavily used. Backup performance is also affected by this additional step, although this impact is often mitigated by the reduced amount of data transfer and storage pool writes.

- Due to the additional processing step to decompress the data, compression might have a small effect on restore performance.

1.2.4 Prerequisites for configuring IBM Spectrum Protect container storage pool data deduplication

Section 1.2.4 provides general information on prerequisites for IBM Spectrum Protect container storage pool data deduplication. For a complete list of prerequisites, refer to the IBM Spectrum Protect administrator documentation.

1.2.4.1 Prerequisites common to client and server-side data deduplication

- The destination storage pool must be of type container.
- The server database and storage pool file systems must be configured according to best practices for high performance. Refer to the IBM Spectrum Protect Blueprints documentation: http://ibm.biz/IBMSpectrumProtectBlueprints

1.2.4.2 Prerequisites specific to client-side data deduplication

To configure client-side data deduplication, the following requirements must be met:

- The client must be at version 6.2.0 or later for client-side data deduplication, and version 7.1.6 or later for client-side data deduplication and LZ4 compression. The latest maintenance version is always recommended.
- The client-side data deduplication option is enabled at the client (DEDUPLICATION YES).
- The server must enable the node for client-side data deduplication with the DEDUP=CLIENTORSERVER parameter by using either the REGISTER NODE command, UPDATE NODE command, or through the Operations Center GUI.
- Files must be bound to a management class with the destination parameter that points to a container storage pool.
- By default, all client files that are at least 2 KB and smaller than the value that is specified by the server *clientdeduptxlimit* option are processed with data deduplication. The *exclude.dedup* client option provides a feature to selectively exclude certain files from client-side data deduplication processing.

The following IBM Spectrum Protect features are incompatible with client-side data deduplication:

- Client encryption
- UNIX HSM client
- Subfile backup

1.2.5 Comparing IBM Spectrum Protect data deduplication and appliance data deduplication

Container storage pool data deduplication provides the most cost-effective solution for reducing backup storage costs. IBM Spectrum Protect data deduplication requires no additional software license, and it does not require special purpose hardware appliances. Data deduplication of backup data can also be accomplished by using a data deduplication storage device in the IBM Spectrum Protect storage pool

hierarchy. Data deduplication appliances such as IBM's ProtecTIER provide data deduplication capability at the storage device level. NAS devices are also available that provide NFS or CIFS-mounted storage that removes redundant data through data deduplication.

Both IBM Spectrum Protect container storage pool data deduplication and data deduplication appliances can be used in the same environment for separate storage hierarchies or in separate server instances. For example, IBM Spectrum Protect client-side data deduplication is an ideal choice for backing up remote environments, either to a local server or to a central datacenter. IBM Spectrum Protect replication (storage pool protection and node replication) can then take advantage of the deduplicated storage pools to reduce data transfer requirements between IBM Spectrum Protect servers, for disaster recovery purposes.

For a data deduplication NAS device, a directory-container pool could be created on the NAS. However, this combination of software and appliance data deduplication is not recommended, since the data is already deduplicated and compressed by IBM Spectrum Protect.

1.2.5.1 Factors to consider when comparing IBM Spectrum Protect and appliance data deduplication

Major factors to consider on which data deduplication technology to use:

- Scale
- Scope
- Cost

1.2.5.1.1 Scale

The IBM Spectrum Protect data deduplication technology is a scalable solution that uses software technology that makes heavy use of database transactions. For a specific IBM Spectrum Protect server hardware configuration (for example, database disk speed, processor and memory capability, and storage pool device speeds), there is a practical limit to the amount of data that can be backed up using data deduplication.

The two primary points of scalability to consider are the daily amount of new data that is ingested, as well as the total amount of data that will be protected over time. The practical limits that are described are not hard limits in the product, and vary based on the capabilities of the hardware that is used. The limit on the amount of protected data is presented as a guideline with the purpose of keeping the size of the IBM Spectrum Protect database below the recommended limit of 6 TB. A 6 TB database corresponds roughly to 1 to 4 PB of protected data (original data plus all retained versions). Daily ingest limits are prescribed with the goal of allowing enough time each day for the maintenance tasks to run efficiently.

Data deduplication appliances are configured with dedicated resources for data deduplication processing, and do not have a direct impact on IBM Spectrum Protect server performance and scalability. If it is desired to scale up a single-server instance as much as possible, beyond approximately 4 PB of protected data, then appliance data deduplication can be considered. However, often a more cost-effective approach is to scale out with more IBM Spectrum Protect server instances. Using more server instances provides the ability to manage many multiples of 4 PB protected data.

In addition to the scale of data that is stored, the scale of the daily amount of data that is backed up has a practical limit with IBM Spectrum Protect. The daily ingest is established by the capabilities of system resources as well as the inclusion of secondary processes such as replication. Occasionally exceeding the limit for daily ingest is okay. Regularly exceeding the practical limit on daily ingest for your specific hardware can cause backup durations to run longer than wanted, or not leave enough time for replication to complete.

Since data deduplication appliances are single-purpose devices, they have the potential for greater throughput due to the use of dedicated resources. A cost/benefit analysis must be considered to determine the appropriate choice or mix of data deduplication technologies. The following table provides some general guidelines for daily ingest ranges for each IBM Spectrum Protect server relative to hardware configuration choices.

Ingest range	Server requirements	Storage requirements
Up to 10 TB per day	✓ 12 CPU cores (Intel) ✓ 6 CPU cores (Power 8) ✓ 64 GB RAM	✓ Database and active log on SSD/flash or SAS/FC 15K rpm ✓ Storage pool on NL-SAS/SATA
10 - 20 TB per day	✓ 16 CPU cores (Intel) ✓ 10 CPU cores (Power 8) ✓ 128 GB RAM	✓ Database and active log on SSD/flash ✓ Storage pool on NL-SAS/SATA
20 - 100 TB per day	✓ 44 CPU cores (Intel) ✓ 20 CPU cores (Power8) ✓ 256 GB RAM	✓ Database and active log on SSD/flash storage ✓ Storage pool on NL-SAS/SATA

1.2.5.1.2 Scope

The scope of IBM Spectrum Protect data deduplication is limited to a single-server instance and more precisely within a single container storage pool. A single shared data deduplication appliance can provide data deduplication across multiple servers and storage pools.

When node replication or storage pool protection is used in a many-to-one architecture, the container storage pool on the replication target can deduplicate across the data from multiple source servers.

1.2.5.1.3 Cost

IBM Spectrum Protect data deduplication functionality is embedded in the product without an extra software license cost. In fact, software license costs will reduce when capacity-based licensing is in use because the capacity is calculated after data reduction. It is important to consider that hardware resources must be sized and configured for data deduplication. However, these additional costs can easily be offset by the savings in disk storage.

Data deduplication appliances are priced for the performance and capability that they provide, and generally are considered more expensive per GB than the hardware requirements for IBM Spectrum Protect software defined data deduplication. A detailed cost comparison must be done to determine the most cost-effective solution.

1.3 Conditions for effective use of container storage pool data deduplication

Although IBM Spectrum Protect data deduplication provides a cost-effective and convenient method for reducing the amount of disk storage for backups. Under some conditions, container storage pools might not be effective in data reduction, and in fact can reduce the efficiency of a backup operation.

Conditions that lead to effective use of IBM Spectrum Protect container storage pool data deduplication include:

- Need for reduction of the disk space for backup storage.
- Need for remote backups over limited bandwidth connections.
- Use of IBM Spectrum Protect storage pool protection and node replication for disaster recovery across geographically dispersed locations.
- Total amount of backup data and data that is backed up per day are within the recommended limits of less than 4 PB total and 100 TB per day for each server instance.
- Backup data must be a good candidate for data reduction through data deduplication.
- High-performance SSD/Flash disk must be used for the IBM Spectrum Protect database to provide acceptable data deduplication performance.

1.3.1 Traditional IBM Spectrum Protect architectures that are compared with data deduplication architectures

A traditional IBM Spectrum Protect architecture ingests data into disk storage pools, and moves this data to tape on a frequent basis to maintain adequate free space on disk for continued ingestion. An architecture that includes data deduplication changes this model to store the primary copy of data in a container storage pool for its entire lifecycle. Data deduplication provides enough storage savings to make keeping the primary copy on disk affordable.

Tape storage pools still have a place in this architecture for maintaining a secondary storage copy for disaster recovery purposes, or for data with long retention periods.

1.4 When is it not appropriate to use IBM Spectrum Protect data deduplication?

IBM Spectrum Protect container storage pools are not appropriate for the following solutions:

- When the desired architecture includes moving data from disk to tape
- Encrypted client data
- NDMP backups
- With data that is known to not benefit from data deduplication or compression

1.4.1 Restore performance considerations

Restore performance from deduplicated storage pools is slower than from a comparable disk storage pool that does not use data deduplication. However, restore performance from a deduplicated storage pool can compare favorably to restore from tape devices for certain workloads.

If fastest restore performance from disk is a high priority, then restore performance benchmarking must be done to determine whether the effects of data deduplication can be accommodated. The following table compares the restore performance of small and large object workloads across several storage scenarios.

Storage pool type	Small object workload	Large object workload
Tape	Typically slower due to tape mounts and seeks	Typically faster due to streaming capabilities of modern tape drives
Non-deduplicated disk	Typically faster due to absence of tape mounts and quick seek times	Comparable to or slightly slower than tape
Deduplicated disk	Faster than tape, slower than non-deduplicated disk	Slightly slower than non-deduplicated disk since data must be rehydrated.

1.5 Container storage pool protection and replication

IBM Spectrum Protect container storage pool data can be protected at a storage pool level with the PROTECT STGPOOL command, and at the inventory with the REPLICATE NODE command.

Storage pool protection (PROTECT STGPOOL) is a storage level protection mechanism for directory-container pools that is similar to the copy storage pool feature, and allows for the repair of damaged chunks. Data chunks are sent to the replication target server, or to a tape device in their deduplicated and compressed form.

Node replication is an inventory or node level protection feature that replicates data chunks and inventory level data to a target server. Node replication allows for an active/active replication architecture where a client restore can fail over to the target replication server.

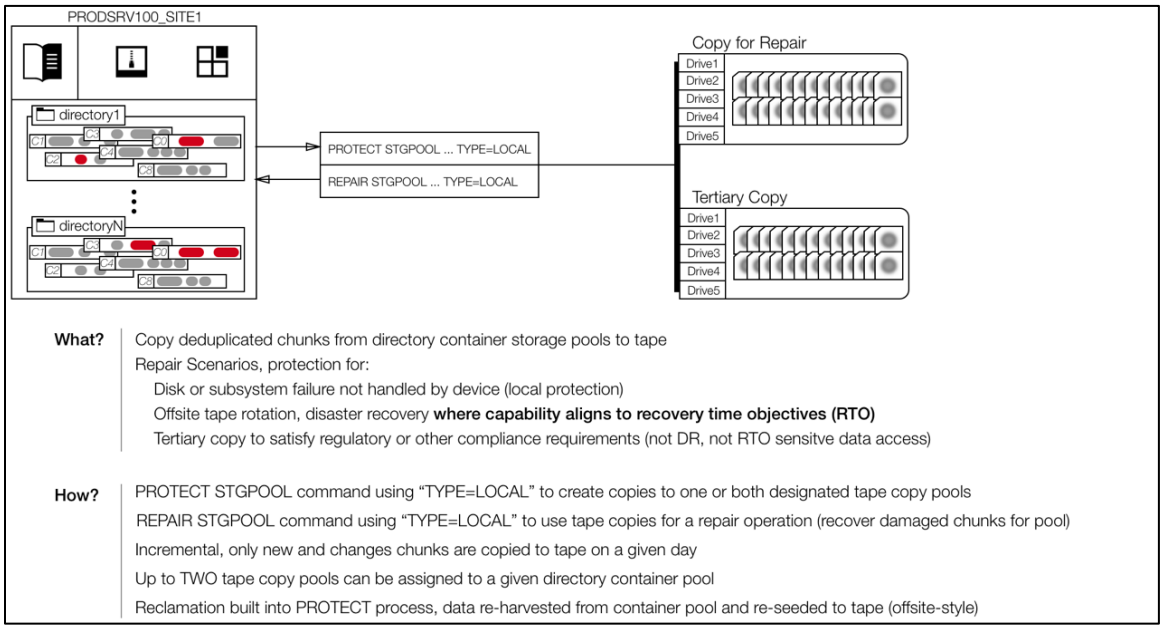
Both protect storage pool and node replication features should be combined for both a chunk and inventory level recovery. When combined, both are achieved with only one transfer of the data. Also, protect storage pool is more efficient in transferring chunks than node replication. When protect storage pool is issued before node replication and completes, the overall replication process completes faster.

1.5.1 Container storage pool protection with a replication target server

Storage pool protection (PROTECT STGPOOL) is a storage pool level protection feature that replicates deduplicated chunks to a replication target server. Protect storage pool replication requires a directory-container storage pool at both the source and target. Protect storage pool replication can be set up to replicate either one-way (Server A to Server B), or bi-directionally (Server A to Server B and Server B to Server A). Protect storage pool replication is incremental, and transfers unique chunks that do not exist on the replication target container pool. By using a single directory-container storage pool for local client backups and as the replication target, the larger data deduplication scope can potentially improve data deduplication reduction. Damaged or missing chunks can be repaired with the repair storage pool command (REPAIR STGPOOL). Client/Nodes cannot restore directly from a replication target server that is only using protect storage pool replication. Node replication must be used along with protect storage pool to ensure data protection for full failover capabilities.

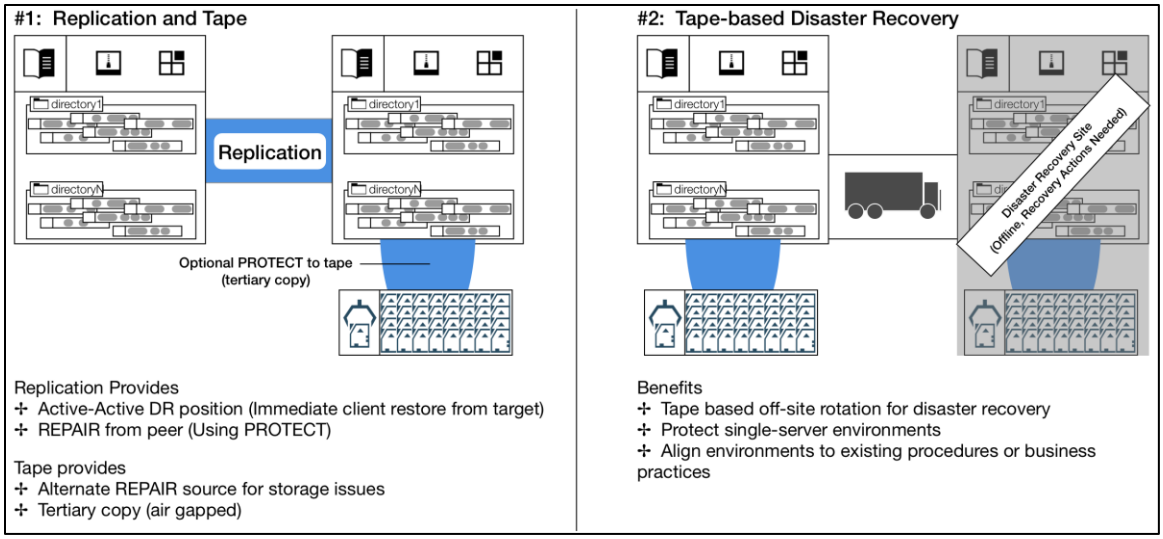
1.5.2 Container Storage Pool Protection with Tape

IBM Spectrum Protect directory-container pools can also be protected by copying the data or chunks to tape. Like replication-based protect storage pool, the tape copy can be used to repair damaged or missing chunks. Storage pool protection to tape is incremental, keeps data in its deduplicated and compressed form, and the protect storage pool operation automatically manages the maintenance of reclamation. The tape copy can be used for compliance, as an air gap copy, and for environments where a replication server is not available. Protect to tape can also be used as a disaster recovery copy with tape-based off-site rotation. Similar to the replication method of protect storage pool, clients/nodes cannot restore data directly from the tape copy. The directory-container pool must be repaired completely before clients can restore. Depending on the size of the container pool and the extent of the damaged and missing chunks, the repair storage pool process can take several days to process.



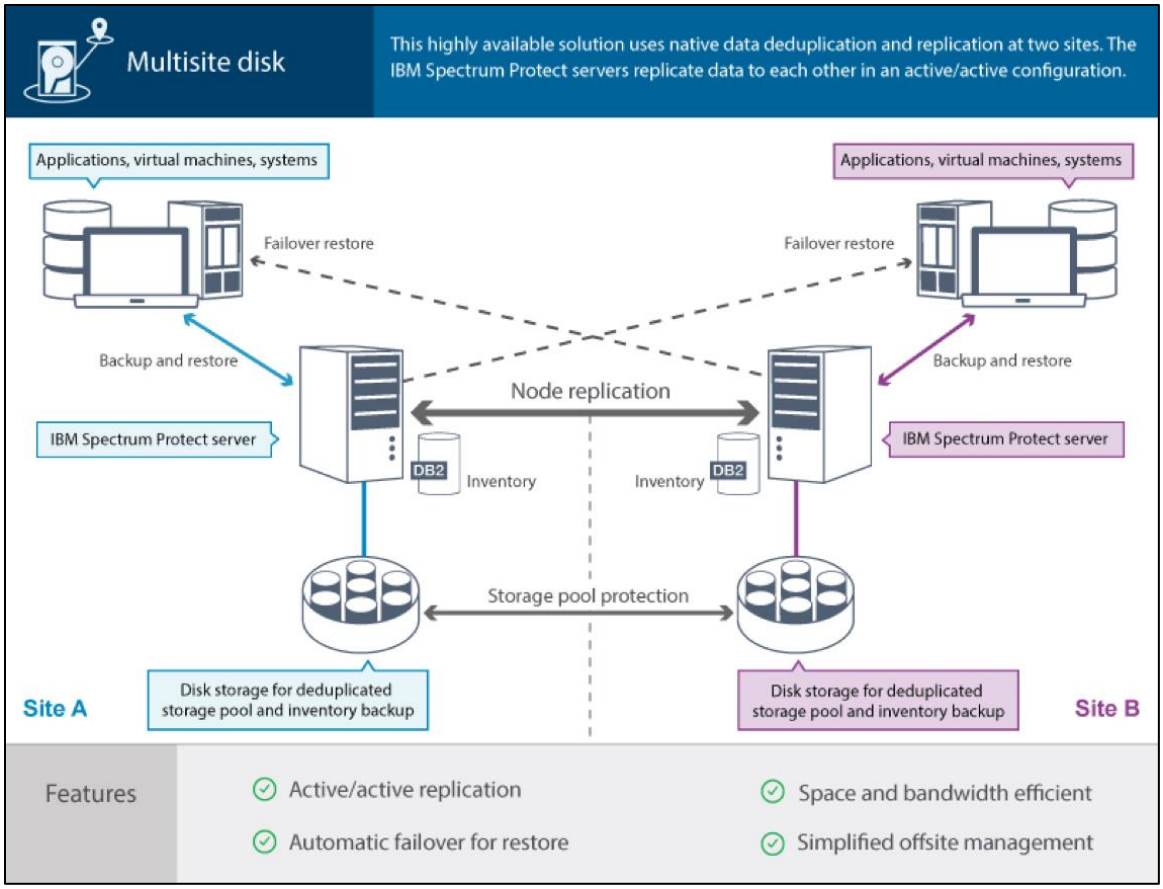
1.5.3 Combining Replication and Tape Container Pool Protection

Both replication and tape storage pool protection methods can be combined for more recovery options, as well as providing a tertiary copy of the storage pool data. The repair storage pool process can use either the replica or tape copy, based on availability.



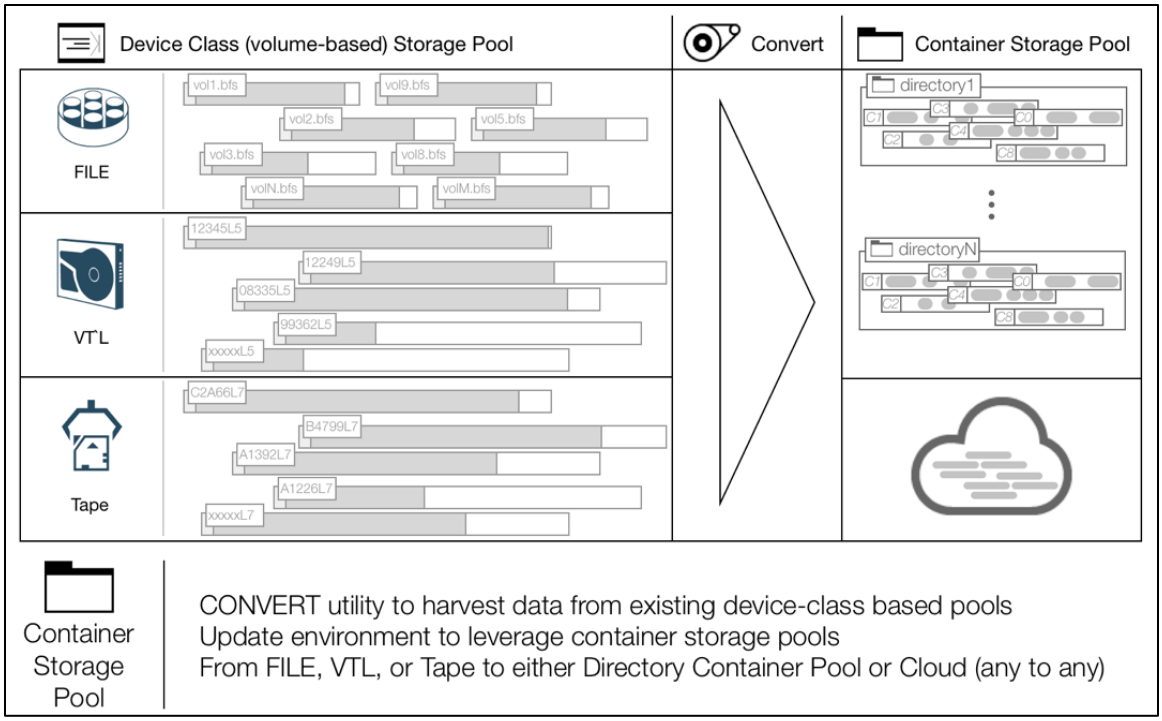
1.5.4 Node Replication

For a complete replication strategy, node replication must be used along with protect storage pool. Node replication is a node inventory or file level protection, allowing clients to restore from the replication target server when there is a failover. Completing the protect storage pool process before node replication improves the performance of the node replication process. Although both replication techniques are used, unique data needs to be transferred one time. The dissimilar policies capability of node replication must only be combined with storage pool protection when the replication target server is configured to retain data longer than the source server.



1.6 Convert Storage Pool

With IBM Spectrum Protect 8.1.0, existing file device, virtual tape library, and physical tape pools can be converted or migrated to a container pool by using the convert storage pool command (CONVERT STGPOOL). The convert storage pool is a one-time process that can be run incrementally over several days or weeks. During the convert process, all source data is moved from the source pool to the container storage pool. When the process is completed, the source pool becomes unavailable and is eligible for deletion after the conversion process. Data from copy and active-data storage pools that are associated with the source pool are deleted during the convert process.



Storage pool conversion typically follows this sequence:

1. Create the container storage pool that is the target of the conversion.
2. Update policy so that new client backups target the new container storage pool.
3. Schedule the convert storage pool to run during the server's maintenance window for as many hours per day as possible.

For the conversion of a FILE device pool to a container storage pool, the simplest conversion is to use new storage for the target container storage pool. Alternatively, the conversion can be performed in-place, where the target container storage pool is using the same disk storage as the existing FILE device pool. For an in-place conversion, consider this information:

- Available space to the target storage pool must be 50% or more of the used space within the source storage pool.
- The conversion process must periodically be stopped by using the DURATION parameter. Pausing the conversion process allows for deletion processing in the source storage pool to catch up and free more space for the target storage pool.
- During the conversion, damaged objects in the source storage pool can prevent the release of storage to the target storage pool. Issue the QUERY CLEANUP command to monitor for damage that might require cleanup. Refer to the link for more details on how to clean up a storage pool: http://ibm.biz/StgpoolCleanUp

Considerations for a VTL or TAPE device class conversion to a container storage pool:

- Do not use all available tape drives for the storage pool conversion. Reserve a few drives for database backup, reclamation, and client restore operations.

- The conversion process can take weeks, and depends on the total amount of data to be converted.

1.7 Recovering a container storage pool

If the IBM Spectrum Protect server database is recovered, or a file system that is associated with a directory container pool is damaged, corrupted, or lost, use the audit container command (AUDIT CONTAINER). The audit container command is used to identify inconsistencies between the database and what is stored in the directory container storage pool. After the audit container process completes, the query damage command (QUERY DAMAGE) displays any damaged chunks in the pool.

A container storage pool that is protected by local tape or a remote secondary server can recover the damaged data by issuing the repair storage pool command (REPAIR STGPOOL). After the repair storage pool command completes, issue a QUERY DAMAGE command to verify that no damaged chunks remain. Damaged chunks can also be repaired with client backups. If a client backup references a damaged chunk, it automatically replaces the damaged chunk.

2 Resource requirements for IBM Spectrum Protect container storage pool data deduplication

IBM Spectrum Protect data deduplication provides significant benefits because of its data reduction technology, particularly when combined with other data reduction techniques. However, the use of data deduplication adds extra requirements for server hardware, and database/log storage, which are essential for a successful implementation. To configure container storage pools, you must ensure that proper resources are allocated to support the use of the technology. The resource considerations include:

- Hardware requirements necessary to meet the additional processing that is performed during data deduplication.
- Increase in storage requirements for the database records used to store the data deduplication catalog.
- Increase in storage requirements for the server database logs.

The IBM Spectrum Protect internal database plays a central role in enabling the data deduplication technology. Data deduplication requires more database capacity to be available. Also, the database has a significant increase in the frequency of references to records during backup, restore, replication, and expiration operations. These demands on the database require that the database and active log disk storage can sustain higher rates of I/O operations than without the use of data deduplication.

As a result, planning for resources that are used by the IBM Spectrum Protect database is critical for a successful data deduplication deployment. This section guides you through the estimation of resource requirements to support container storage pool data deduplication.

2.1 Database and log size requirements

2.1.1 IBM Spectrum Protect database capacity estimation

Use of data deduplication significantly increases the capacity requirements of the IBM Spectrum Protect database. This section provides some guidelines for estimating the capacity requirements of the database. It is important to plan for the database capacity so enough high-performing disk can be reserved for the database (refer to the next section for performance requirements).

The estimation guidelines are approximate, since actual requirements depend on many factors. Some factors that cannot be predicted ahead of time (for example, a change in the data backup rate, the exact amount of backup data, and other factors).

2.1.1.1 Planning database space requirements

The use of data deduplication requires more storage space in the IBM Spectrum Protect server database than without the use of data deduplication. Each “chunk” of data that is stored in a container storage pool is referenced by an entry in the database. The use of storage pool compression does not have a significant impact on database size.

Without data deduplication, each backed-up object (typically a file) is referenced by a database entry, and the database grows proportionally to the number of objects that are stored. With data deduplication, the database grows proportionally to the total amount of data that is backed up. Other factors that influence database size include the average data deduplication chunk size that varies from one data source to the next. The following table provides an example to illustrate these points and provides a guide that can be used for planning database sizes for container storage pools:

Workload type		Number of objects stored	Amount of data being managed	Database Storage requirements
Unstructured data (smaller average object sizes)	**Without data deduplication**	500 million	200 TB	475 GB *
	With data deduplication	500 million	200 TB	2000 GB **
Structured data (larger average object sizes)	**Without data deduplication**	5 million	200 TB	5 GB *
	With data deduplication	5 million	200 TB	500 GB ***

* Using non-deduplicated estimate of 1 KB of database space per object stored
** Using unstructured estimate of 1000 GB of database space per 100 TB of data that is managed with average object sizes 512 KB or smaller
*** Using structured estimate of 250 GB of database space per 100 TB of data managed

2.1.1.2 Planning active log space requirements

The database active log stores information about database transactions that are in progress. With data deduplication, transactions can run longer, requiring more space to store the active transactions.

Tip: Use a minimum of 128 GB for the active log and 256 GB for very active servers.

2.1.1.3 Planning archive log space requirements

The archive log stores older log files for completed transactions until they are cleaned up as part of the IBM Spectrum Protect server database backup processing. The archive log file system must be given sufficient capacity to avoid running out of space, which can cause the server to be halted. Space is freed in the archive log every time a full backup is performed of the server's database.

For detailed information on how to calculate the space requirements for the server archive log: http://ibm.biz/SizingTheArchiveLog

Tip: A file system with 4 TB of free space is more than adequate for a large-scale server that ingests up to 100 terabytes a day of new data into deduplicated storage pools and performs a full database backup once a day.

2.2 Estimating capacity for container storage pools

IBM Spectrum Protect container storage pool data reduction ratios typically range from 2:1 (50% reduction) to 15:1 (93% reduction). The ratio of 15:1 corresponds to an overall data reduction when factoring in the data reduction benefits of progressive incremental backups. Data reduction depends on both the type of data and type of backups performed. Lower data reduction ratios are associated with backups of unique data (for example, such as progressive incremental data). Higher data reduction ratios are associated with backups that are repeated, such as repeated full backups of databases or

virtual machine images. Mixtures of unique and repeated data results in ratios within that range. If you aren't sure of what type of data you have and how well it reduces, for planning purposes, use a 4:1 data reduction for progressive incremental workloads, or an 8:1 data reduction for daily full workloads.

2.2.1 Estimating storage pool capacity requirements

2.2.1.1 Plan for reserved free space in your container storage pool

Due to the latency for deletion of data chunks, “transient” space is needed in a container storage pool. Data chunks are eligible for deletion based on reference counts and the value of the storage pool REUSEDELAY parameter. When a data extent is eligible for deletion, it is removed from the container storage pool. The reuse delay is important when the IBM Spectrum Protect database needs to be recovered. The reuse delay default is 1 day, and can be increased if longer reuse periods and recovery points. Consider reserving up to an extra 20% of storage pool capacity for this transient storage and as a contingency for changes in requirements.

2.2.1.2 Estimating physical storage requirements for a container storage pool

You can roughly estimate physical storage requirements for a container storage pool by using the following technique:

- Estimate the base size of the source data
- Estimate the daily backup size, by using an estimated change and growth rate
- Determine retention requirements
- Estimate the total amount of source data by factoring in the base size, daily backup size, and retention requirements.
- Apply the data reduction ratio factor
- Add additional reserved space
- Add annual growth factor

The following example illustrates the estimation method:

Parameter	Value	Notes
Base size of the source data	100TB	Data from all clients that are backed up to the deduplicated storage pool.
Estimated daily change rate	5%	Includes new and changed data
Retention requirement	30 days	
Estimated data reduction ratio[1]	4:1	4:1 combined data reduction with data deduplication and compression
Add additional reserved space	20%	
Factor annual growth of client data	3 years at 20%	Future value = present value (1 + rate) ^ (number of years)

1. You can use an 8:1 data reduction for workloads consisting primarily of daily full backups.

Computed Values:

Parameter	Computation	Result
Base source data	100 TB	100 TB
Estimated daily backup amount	100 TB * 0.05 change rate	5 TB
Total changed data retained	(30 – 1) * 5 TB daily backup	145 TB
Total managed data (use for planning DB size)	100 TB base data + 145 TB retained	245 TB
Retained data after data reduction (4:1 ratio)	245 TB / 4	61.25 TB
Add reserved space (20%)	61.25 TB + (61.25 * 0.2)	73.5 TB
Factor annual growth (3 years at 20%)	73.5 TB (1 + .20) ^ 3	127 TB

2.3 Hardware recommendations and requirements

The use of container storage pools requires additional processing, which increases the IBM Spectrum Protect server hardware requirements beyond what is required without the use of data deduplication. The most critical hardware requirements for data deduplication is the I/O capability of the disk system that is used for the IBM Spectrum Protect database and CPU capacity. **The most detailed and comprehensive guidance for hardware requirements for an IBM Spectrum Protect server with container storage pools is provided in the IBM Spectrum Protect Blueprints documentation: http://ibm.biz/IBMSpectrumProtectBlueprints.** The blueprints are designed with sufficient CPU, memory, and database disk to handle both deduplication and compression.

2.3.1 Database I/O requirements

For optimal performance, solid-state disk (SSD) or Flash technology storage is recommended for the IBM Spectrum Protect database. Due to the random-access I/O patterns of the database, minimizing the latency of operations that access the database volumes is critical for optimizing performance. The large tables that are used for storing data deduplication information in the database increases the demand for disk storage that can handle many input/output operations per second (IOPS).

Details about how to configure high performing disk storage are beyond the scope of this document. The following key points must be considered for configuring IBM Spectrum Protect database disk storage:

- SSD or Flash disk devices must be used for the IBM Spectrum Protect database storage and the active log. The cumulative IOPS for database volumes often exceed 20,000 and even 50,000 for a large server. Slower disk technology is acceptable for the archive log.
- Disk devices or storage systems capable of a minimum of 10,000 IOPS are suggested for the database disk storage. An additional 1,000 IOPS per TB of daily ingested data (pre-deduplication) with a ceiling of 50,000 must be considered.
- IBM Spectrum Protect database and logs must be configured on separate disk volumes (LUNS), and must not share disk volumes with the IBM Spectrum Protect storage pools or any other application or file system.

2.3.1.1 Using flash storage for the IBM Spectrum Protect database

A significant benefit to data deduplication and replication scalability is achieved by using flash storage for the IBM Spectrum Protect database. Many choices available when moving to flash technology. Container storage pool testing has been performed with the following classes of flash-based storage for the server database:

- Flash acceleration that uses in-server PCIe adapters. For example, the High IOPS MLC (Multi Level Cell) and Enterprise Value Flash adapters.
- Solid-state drive modules (SSDs) as part of a disk array. For example, the SSD options available with the IBM Storwize family of disk arrays.
- Flash memory appliances, which provide a solution where flash storage can be shared across more than one IBM Spectrum Protect server. For example, the IBM FlashSystem family of products.

The following are some general guidelines to consider when implementing the IBM Spectrum Protect database with solid-state storage technologies:

- Solid-state storage provides the most significant benefit for the database and active log.
- The archive log has no substantial benefit on solid-state storage.
- SSD arrays that use RAID10 do not provide enough of a performance advantage over RAID5 to justify the cost of the additional capacity loss.
- Faster database access from using solid-state technology allows pushing the parallelism to the limit with tuning parameters for tasks such as backup sessions, expire inventory resources, and protect/replication max sessions.

2.3.2 CPU

The use of container storage pools requires additional CPU resources on the IBM Spectrum Protect server, particularly for performing the task of data reduction that uses data deduplication and compression. You must consider a minimum of at least twelve Intel Xeon (Haswell, Broadwell, or newer) or six Power8 physical processor cores in any IBM Spectrum Protect server that is configured with container storage pools. The following table provides CPU recommendations for different ranges of daily ingest.

Daily ingest	Recommended Intel CPU cores	Recommended Power 8 CPU cores
Up to 10 TB	12	6
6 TB to 20 TB	16	10
20 TB to 100 TB	44	20

2.3.3 Memory

Memory or RAM is used to optimize the frequent lookup of data deduplication chunk information that is stored in the IBM Spectrum Protect database. Memory sizes are planned based on the planned size of the database.

A minimum of 64 GB of system memory must be considered for IBM Spectrum Protect servers that use container storage pools. As the database capacity grows, the memory requirement can be as high as 256 GB. The following table provides system memory guidance for different database sizes.

Database Size	Recommended system memory
1 TB	64 GB
2 TB	128 GB
6 TB	256 GB

2.3.4 Considerations for directory-container storage pool disk

The speed of the disk technology that is used for a container storage pool also has significant implications to the overall performance of a data deduplication solution. In general, using low cost, higher capacity disk such as Nearline-SAS (NL-SAS) in RAID6 arrays is desirable for the storage pool to keep the overall cost down. To prevent the use of slower disk technology from impacting performance, it is important to distribute the storage pool I/O across many disks. The quantity and capacity of NL-SAS drives to use needs to consider the throughput of storage pool writes that are required during backup. For planning purposes, consider:

- The storage pool I/O pattern during backup is 100% large sequential write with an average I/O size of approximately 256 KB.
- Plan for approximately 90 MB/s of throughput to be available from every 12 drive 10+p+q array or 500 MB/s of throughput to be available from every distributed RAID6 array with 45 or more drives.
- The actual storage pool throughput that is required is determined by taking the daily ingest requirement, which is divided by the number of hours in the backup window, and reducing by the expected data reduction.

The following shows an example of planning the storage pool arrays for a server that handles a maximum daily ingest of 15 TB/day in an 8-hour backup window, with a physical storage pool capacity of 180 TB.

Parameter	Computation	Result
Determine throughput requirement	15 TB / 8 hours	1.9 TB/hr
Convert to MB/sec	1.9 *1024 * 1024 / 60 / 60	553 MB/sec
Factor in data reduction (4:1)[1]	553 / 4	138.2 MB/sec
Determine rounded up number of drives from throughput per 12 drive array	138.2 / 90 = 1.5 rounded up to 2	2 arrays of 24 drives
Determine rounded up drive size that is needed to meet the physical capacity	180 TB / 24 drives = 7.5 TB/drive	8 TB drives

1. You can use an 8:1 data reduction for workloads consisting primarily of daily full backups.
2.

Consider some other general guidelines for configuring storage for a container storage pool:

- Container pool performance does not depend on having a large quantity of storage pool directories assigned. However, the blueprints create in the range 20 - 75 volumes to avoid

unreasonably large file systems, and minimize the scope of damage if one file system becomes corrupted.
- Maintain a one-to-one relationship between the disk system volume (or LUN), the operating system logical volume, and file system.

2.3.5 Hardware requirements for IBM Spectrum Protect client data deduplication

Client-side data deduplication (and compression if used with data deduplication) requires resources on the client system for processing. Before deciding to use client-side data deduplication, you must verify that client systems have adequate resources available during the backup window to perform the data deduplication processing. Using client-side data deduplication is most appropriate for configurations where a dedicated data mover that is performing the backup, and where CPU resources are not being shared with an application. For example, Data Protect for VMware can use a dedicated data mover.

- In addition to application needs on the client, a suggested minimum CPU requirement is the equivalent of one CPU core per backup session with client-side data deduplication.
- No significant additional memory requirement for client systems that use client-side data deduplication or compression.

3 Implementation guidelines

A successful implementation of an IBM Spectrum Protect server with a container storage pool requires careful planning in the following areas:

- Implement an appropriate architecture suitable for using a container storage pool
- Properly size your server hardware and storage
- Configure the server by following the blueprint documentation

3.1 Deciding between client and server data deduplication

After you decide on an architecture that uses IBM Spectrum Protect data deduplication. You need to decide whether you will perform data deduplication on the clients, the server, or using a combination of the two. The container pool implementation allows for a storage pool to manage data deduplication that is performed by both clients and the server. The server is optimized to perform data deduplication on data that has not already been deduplicated by the clients. Furthermore, duplicate data can be identified across objects regardless of whether the data deduplication is performed on the client or server. These benefits allow for hybrid configurations that efficiently apply client-side data deduplication to a subset of clients, and use server-side data deduplication for the remaining clients.

Typically, a combination of both client-side and server-side data deduplication is appropriate. Some further points to consider:

- Compression can also be performed at the client, allowing the optimal combination of data deduplication followed by compression.
- Client-side data deduplication and compression can reduce the amount of data that is sent across the backup network. Client-side data deduplication can allow for a larger total ingest by distributing the CPU processing for data deduplication and compression across a larger number of systems.

Tips:

Perform data deduplication at the client in combination with compression in the following circumstances:

- Your backup network speed is a bottleneck.
- The client system has available CPU resource or is a dedicated data mover.
- The total daily ingest to be processed exceeds the CPU capability of your server.

3.2 Container storage pool configuration recommendations

NOTE: The IBM Spectrum Protect Blueprints and scripts should be considered when configuring a server that uses container storage pools. The latest best practices for configuring the operating system, storage layout, and configuring IBM Spectrum Protect are included in the Blueprints. This section provides information on some of the specific configuration details that are recommended for container storage pools.

3.2.1 Recommendations for container storage pools

The IBM Spectrum Protect server can be configured with more than one container storage pool, but duplicate data is not identified across different storage pools. In most cases, by using a single large container storage pool is recommended. With container storage pools, data deduplication is always enabled. Although the use of compression is optional, it will be enabled by default. Leaving it enabled is recommended.

The following commands provide an example of setting up a container storage pool on the server. Alternatively, the Operations Center storage pool wizard can be used to create a container storage pool. Some parameters are explained in further detail to give the rationale behind the values that are used, and later sections build upon those settings.

3.2.1.1 Storage pools

The storage pool is the repository for deduplicated data that uses space on one or more storage pool directories, or cloud Object Storage. Examples of the commands for defining storage pools.

Define a directory-type container storage pool, that is used with block or network-attached storage:

```
define stgpool deduppool stgtype=directory
```

Define a cloud-type container storage pool with Object Storage that uses the S3 protocol:

```
define stgpool deduppool stgtype=cloud cloudtype=s3 cloudurl=”https://cloud_ipaddress1/ |
https://cloud_ipaddress2/ | …” id=”access_key_id” password=”secret_access_key”
```

3.2.1.2 Storage pool directories

Once a container storage pool is defined, storage pool directories need to be defined to provide storage space. A storage pool directory must correspond to a dedicated file system on the server. Multiple storage pool directories (file systems) can be defined to a container storage pool, which increases the throughput potential and storage space available to the storage pool. Refer to the blueprint documentation on storage pool directory and storage layout recommendations. Storage pool directories are used differently by directory and cloud container pools. Directory container storage pools use storage pool directories to provide permanent storage for containers. Cloud container pools use storage pool directories as a temporary storage, or accelerator cache, for containers until the reach an optimal size for transfer to Object Storage. The accelerator cache provides fast local storage as a temporary landing spot for incoming backups. The accelerator cache also enhances data transfer out to the cloud Object Storage by grouping deduplicated chunks into larger objects.

An example of the define storage pool directory command follows. On UNIX operating systems, assign ownership of the directories to the database instance user before defining the storage pool directories:

```
define stgpooldirectory deduppool directory=/storage/fsmnt1,/storage/fsmnt2,…
```

For additional accelerator cache details and considerations, refer to the following link:
http://ibm.biz/OptimizePerformanceForCloudObjectStorage

3.2.1.3 Container storage pool encryption

Inline encryption is optional for on-premises and off-premises cloud-container storage pools. When enabled, the server encrypts the data at a data chunk level with AES 256-bit encryption before it is written to the storage pool. When data is retrieved, it is decrypted at the server.

To enable encryption, the server uses a master encryption key, which is created when the server password is set, and is stored as part of the server password file, dsmserv.pwd. The master encryption key is required to decrypt encrypted data, and for this reason, it is important that the server password file is protected. The master encryption key can be protected with the DATABASE BACKUP command when configured to include the master encryption key. Use the SET DBRECOVERY command to enable protection of the master encryption key and to password protect the database backup. The same password is required to restore the master encryption key from the database backup. Here is an example of enabling protect of the master encryption key:

```
set dbrecovery device_class_name protectkeys=yes password=password
```

3.2.1.4 Container copy pools

Where physical tape copies are required, a container copy pool can be configured to hold a copy of data from a container storage pool. A container copy pool can be used as a secondary or air-gap copy of the deduplicated chunks. Container copy pools are restricted to physical tape library devices and directory container pools. An example of creating a copy container pool and updating the directory storage pool to map to the container copy pool follows:

```
define stgpool copycontainerpool tapedevc pooltype=copycontainer maxscratch=100
update stgpool deduppool protectstgpool=copycontainerpool
```

3.2.1.5 Policy settings

The final configuration step involves defining policy settings on the IBM Spectrum Protect server that allow data to ingest directly into the newly create container storage pool. Policy requirements vary for each customer. The following example shows policy that retains extra backup versions for 30 days:

```
define domain DEDUPDISK
define policy DEDUPDISK POLICY1
define mgmtclass DEDUPDISK POLICY1 STANDARD
assign defmgmtclass DEDUPDISK POLICY1 STANDARD
define copygroup DEDUPDISK POLICY1 STANDARD type=backup destination=DEDUPPOOL
VEREXISTS=nolimit VERDELETED=10 RETEXTRA=30 RETONLY=80
define copygroup DEDUPDISK POLICY1 STANDARD type=archive destination=DEDUPPOOL
RETVER=365
activate policyset DEDUPDISK POLICY1
```

3.2.2 Recommended options for data deduplication

The server has several tuning options that control data deduplication processing. The following table summarizes these options. The options that control the maximum data deduplication transaction sizes do not apply to container storage pools.

Server options	Allowed values	Recommended value	Explanation
deduptier2filesize	Min: 20 Max: 9999 Default: 100	Default	Changing the default tier settings is not recommended. Small changes can be tolerated, but avoid frequent changes to these settings, as changes will prevent matches between previously ingested backups and future backups.
deduptier3filesize	Min: 90 Max: 9999 Default: 400	Default	See above.

Storage pool parameters	Allowed values	Recommended value	Explanation
reusedelay	Min: 0 Max: 9999 Default: 1	Default or higher	Avoid setting reusedelay=0
maxwriters	Min: 1 Max: nolimit Default: nolimit	Default	
compression	Yes | No	Default	Use of compression is strongly recommended.

Node parameters	Allowed values	Recommended value	Explanation
maxnummp	Min: 1 Max: 999 Default: 1	99	This parameter is used in determining the maximum number of no-query restore sessions.
deduplicate	clientorserver serveronly	clientorserver	

3.2.3 Best practices for ordering backup ingestion and data maintenance tasks

A successful implementation of data deduplication with IBM Spectrum Protect requires separating the tasks of ingesting client data and performing server data maintenance tasks into separate time windows. Furthermore, the server data maintenance tasks have an optimal ordering, and in some cases, need to be performed without overlap to avoid resource contention problems. The following is the recommended ordering of maintenance with replication to a second server:

1. Storage pool protection (PROTECT STGPOOL command)
2. Node replication (REPLICATE NODE command)
   For both storage pool protection and node replication, the optimal ordering is to run storage pool protection followed by node replication to the secondary server. These two tasks are efficient, in that deduplicated chunks only need to be transferred to the replication target one time. The protect stgpool command transfers chunks faster than node replication, so it is

recommended to run the protect stgpool process before the node replication process. With this ordering the protect storage pool command transfers the deduplicated chunks first, followed by the replicate node command, which replicates the inventory metadata.

3. Protect the IBM Spectrum Protect database (BACKUP DATABASE command). The database backup of the IBM Spectrum Protect server is needed to create a recovery point and to prune the server's database archive log. The recovery point must be created after the daily client backups and protect storage pool and node replication processing. Upon successful completion of a database backup, the delete volume history command must be run to delete older database backup copies that are no longer needed.
4. Perform inventory expiration processing (EXPIRE INVENTORY command). The expire inventory command removes data that exceeds the retention that is specified by policy. This process is generally a lower priority than protecting the data and metadata of the IBM Spectrum Protect server and is run last. Also, after the expire inventory process completes, background chunk deletion threads run in the background to update the reference counts for data chunks. This processing incurs database lookups and updates and can slow down replication and database backup maintenance tasks if run at the beginning of the maintenance window.

Recommended ordering of maintenance tasks with local storage pool protection to tape:

1. Storage pool protection to tape (PROTECT STGPOOL TYPE=LOCAL command)
   With local storage pool protection to tape, separate tape reclamation processing is not needed. Reclamation processing is a built-in step of the storage pool protection process.
2. Optional: Protect data to a secondary server with PROTECT STGPOOL and REPLICATE NODE.
3. Protect the IBM Spectrum Protect database (BACKUP DATABASE command)
4. Perform inventory expiration processing (EXPIRE INVENTORY command)
5. IBM Spectrum Protect can schedule these activities to follow these best practices, and if you use the blueprint configuration scripts these schedules are created automatically.

3.2.3.1 Define scripts that run each required maintenance task

The following scripts, once defined, can be called by scheduled administrative commands. A few points to note regarding these scripts:

- The replication scripts are not used for cloud container storage pools.
- Protect stgpool type=local assumes you already defined a directory container pool, and a copy storage pool named containercopypool.
- The database backup script requires a device class that typically also uses file device or tape storage.
- If you have a large server database, you can further optimize the (BACKUP DATABASE) command by using multiple streams.

Replication scenarios:

Directory-container protect storage pool, followed by node replication to a secondary server:

```
def script REPLICATE description="Run stgpool protection and node replication."
upd script REPLICATE "protect stgpool DEDUPPOOL maxsessions=50 wait=yes" line=010
upd script REPLICATE "replicate node * maxsessions=40 wait=yes" line=020
```

Directory-container protect local to tape, and optionally protect storage pool and node replication to a secondary server:

```
def script REPLICATE description="Run stgpool protection and node replication."
upd script REPLICATE "protect stgpool type=local DEDUPPOOL wait=yes" line=010
upd script REPLICATE "protect stgpool DEDUPPOOL maxsessions=50 wait=yes" line=020
upd script REPLICATE "replicate node * maxsessions=40 wait=yes" line=030
```

Database backup:

```
def script DBBACKUP description="Run full server database backup and remove old backups."
upd script DBBACKUP "backup db devc=DBBACK_FILEDEV type=full numstreams=12 wait=yes" line=010
upd script DBBACKUP "if(error) goto done" line=020
upd script DBBACKUP "backup volhist" line=030
upd script DBBACKUP "backup devconf" line=040
upd script DBBACKUP "delete volhist type=dbb todate=today-5 totime=now" line=050
upd script DBBACKUP "done:exit" line=060
```

Expiration:

```
def script EXPIRE description="Run expire inventory to remove backup objects that exceed the retention."
upd script EXPIRE "expire inventory wait=yes resource=40" line=010
```

3.2.3.2 Define schedules to run the data maintenance tasks

The IBM Spectrum Protect server can schedule commands to run, where the scheduled action is to run the various scripts that were defined in the previous sections. The following examples give specific start times that have proven to be successful in environments where backups run from midnight until 07:00 AM on the same day. You need to change the start times to appropriate values for your environment.

REPLICATE, 8 hours after time of backup start:

```
def sched REPLICATE type=admin cmd="run REPLICATE" active=no description="Run replicate node."
startdate=today starttime=08:00 dur=15 durunit=minutes period=1 perunit=day
```

DB BACKUP, as long as needed beginning @ backup start + 14 hours:

```
def sched DBBACKUP type=admin cmd="run DBBACKUP" active=yes description="Run database backup and
remove old backups." startdate=today starttime=14:00 dur=15 duru=minute period=1 perunit=day
```

EXPIRE, as long as needed beginning @ backup start + 17 hours:

```
def sched EXPIRE type=admin cmd="run expire" active=yes desc="Run expiration to remove backup objects
that exceed retention." startdate=today starttime=17:00 dur=15 durunit=minutes period=1 perunit=day
```

4 Optimizing clients for container pools

Container storage pools can provide both significant data reduction and excellent performance. This section covers techniques for tuning IBM Spectrum Protect clients to achieve both effective data reduction and good performance. There are some general recommendations to consider, as well as considerations for specific client types that have unique requirements.

The data reduction in container storage pools is performed inline, and requires substantial computation, which can have an impact on the throughput of backup ingest. However, the data reduction also results in reduced I/O to both the network and storage systems, which offsets some or all the cost of this computation. In addition, container storage pools can handle many simultaneous sessions. The computational cost can also be offset by increasing the number of parallel sessions that are used for clients that are protecting large amounts of data. If you are moving from a different type of storage pool that did not perform data deduplication, you might need to double the number of sessions for certain types of clients to keep consistent performance.

4.1 Deciding between performing data reduction at client or server

After you decide on an architecture using data deduplication for your IBM Spectrum Protect server, you need to decide whether you will perform data reduction processing on the clients, the server, or by using a combination of the two. The server is optimized to perform data reduction processing on data that was not deduplicated or compressed by the clients.

4.1.1 Client data deduplication processing

Deduplicated data can be identified across objects regardless of whether the data deduplication is performed on the client or server. These benefits allow for hybrid configurations that efficiently apply client-side data deduplication to a subset of clients, and use server-side data deduplication for the remaining clients.

Typically, a combination of both client-side and server-side data deduplication is the most appropriate. Some further points to consider:

- Client-side data deduplication can outperform server-side data deduplication with a high-performing client resources and a low-latency network connection between the client and server
- Data deduplication on the client can be combined with compression to provide the largest possible network bandwidth savings.
- Client-side data deduplication processing can increase backup durations. Expect increased backup durations if network bandwidth is not restrictive. The increased backup durations can be mitigated by increasing the number of client sessions for the backup. Refer to section 4.3 for recommendations.
- The daily ingest limits of a server can be increased by selective use of client-side data deduplication to distribute the processing of data across more systems.

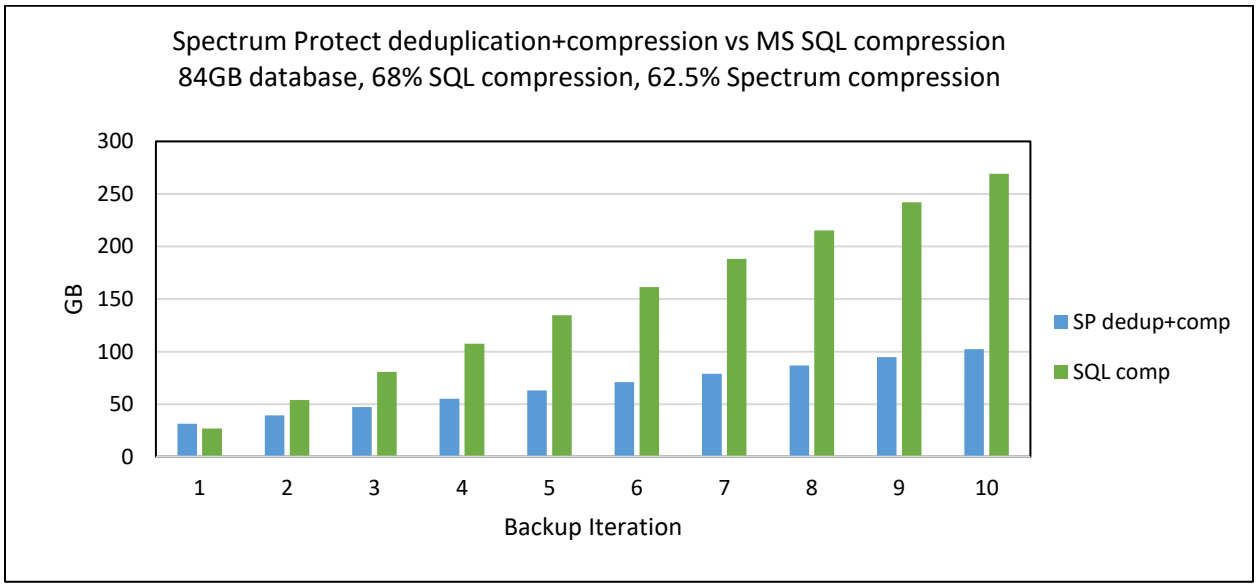
4.1.2 Client compression processing

Data deduplication processing must always occur before compression for optimal data reduction (refer to section 1.2.3.3 for additional details). For optimal data reduction, client-side compression should only be enabled when client-side data deduplication is enabled. Also, be aware that many applications can perform compression for the backup stream, which is another case of suboptimal ordering of data deduplication and compression. Application compression must be disabled to allow IBM Spectrum Protect to perform both the data deduplication and compression. However, application compression features, such as database table or index compression where the compressed data remains relatively static, are acceptable to use with data deduplication processing. Refer to section 4.3 for application compression options that negatively affect data deduplication efficiency.

Client-side compression processing must be combined with client-side data deduplication under the following conditions.

- The backup network speed is a bottleneck
- The client system can afford the additional CPU processing for compression

The ordering of compression and data deduplication is a significant consideration for the overall result of data reduction. The following test results demonstrate how the combination of IBM Spectrum Protect data deduplication and compression provides a significantly better data reduction result than relying solely on database features compress a backup stream. In this example where cumulative data that is stored after 10 days of Microsoft SQL database backups, the data deduplication and compression combination more than doubles the data reduction.



4.2 Client encryption

Encryption before data deduplication and compression processing can negatively impact data reduction. Data deduplication and compression processing is disabled when the IBM Spectrum Protect client-side encryption feature is enabled. With application encrypted data, such as an encrypted database table spaces, it might not be possible to have the data decrypted before it is processed by IBM Spectrum Protect. Testing shows that favorable data reduction can still be obtained for application encrypted data. The data reduction depends on how it changes between backup versions. Identical data which exists at different sources, and is encrypted with different keys, will of course not benefit from data reduction.

An alternative to client-side encryption that does not negatively impact data reduction processing can be achieved with:

- Secure the client to server communication pathway by enabling the IBM Spectrum Protect SSL/TLS feature.
- Use encryption at rest to secure the data in the container storage pool. Encryption at rest can be done within the IBM Spectrum Protect Server with cloud container pools, or at the disk subsystem level.

4.3 Client option recommendations

The recommendations are starting values for client systems that back up to a container storage pool. Many factors affect backup performance, including client to server network connectivity, CPU, memory, and disk I/O capabilities on the client and server. The number of client sessions, along with the data deduplication and compression processing location might need to be tuned from these starting values to achieve optimal performance.

Hints:

- If the client system has high CPU or high disk read latencies, lowering the number of client sessions can improve backup performance.
- If the backup network is saturated, additional client sessions might offer little to no performance value. If the client system has reserve CPU resources, consider enabling client-side data deduplication or a combination of client-side data deduplication and compression.

4.3.1 Backup-Archive client / client API

Backup-Archive client with limited, high latency network (WAN backups):

TCPWINDOWSIZE	512
RESOURCEUTILIZATION	4
COMPRESSION	Yes
DEDUPLICATION	Yes
ENABLEDEDUPCACHE	Yes

Tip: Do not use the client data deduplication caching for applications that use the IBM Spectrum Protect API. Refer to section 1.2.3.2 for additional details.

Backup/Archive client or Client API with limited network (Gigabit LAN backups):

TCPWINDOWSIZE	512
RESOURCEUTILIZATION	10
COMPRESSION	Yes
DEDUPLICATION	Yes
ENABLEDEDUPCACHE	No

Backup/Archive client or Client API with high-speed network (10 Gigabit + LAN backups)

TCPWINDOWSIZE	512
RESOURCEUTILIZATION	10
COMPRESSION	No
DEDUPLICATION	No
ENABLEDEDUPCACHE	No

Tip: For optimal data reduction, **avoid** the following client option combination:

COMPRESSION	Yes
DEDUPLICATION	No

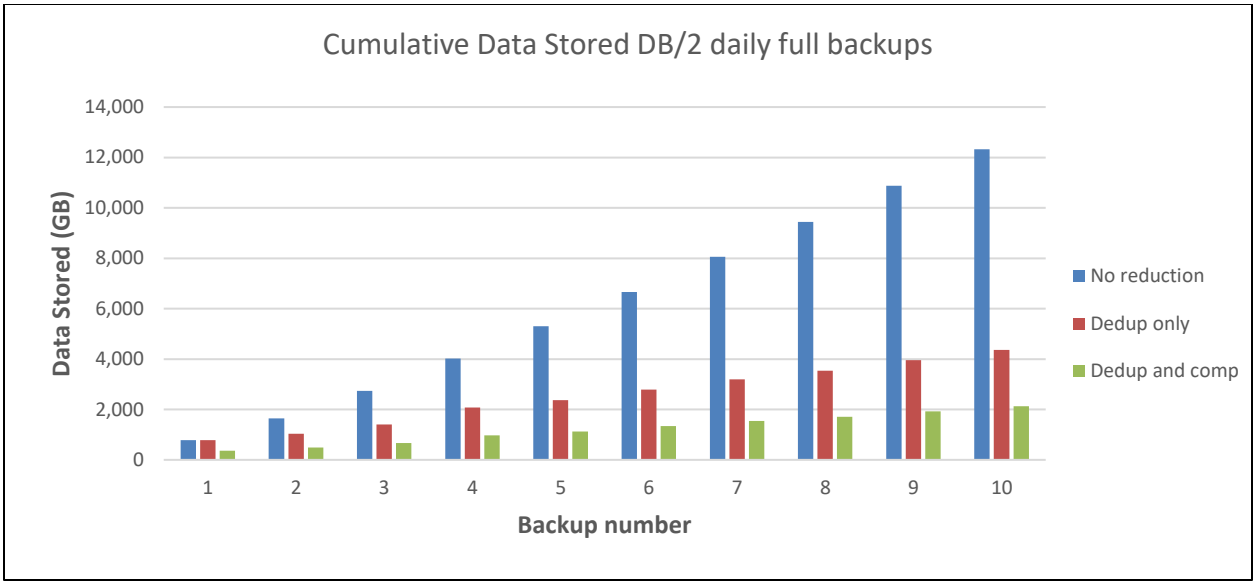
4.3.2 DB2

The DB/2 database provides a backup feature capable of using the IBM Spectrum Protect API to stream backup data directly to a server without temporarily writing the backup data to disk. Daily full backups typically experience significant combined data reduction from data deduplication and compression. Use the SESSIONS option of the DB/2 BACKUP DATABASE command to control the number of backup sessions. The DB2 BACKUP DATABASE options for controlling the number of buffers, buffer size, and parallelism typically self-tune themselves optimally. As a recommended starting point, use 10 sessions. For databases stored on fast disk such as flash, up to 20 sessions might be required to achieve maximum throughput. The DB/2 BACKUP DATABASE command also provides a DEDUP DEVICE option, which is required for some backup systems to optimize data reduction results. IBM Spectrum Protect container storage pool testing demonstrates that the DEDUP DEVICE option is not required for excellent data reduction, and by using it can limit backup throughput by restricting how uniformly backup data is sent across many sessions.

The following BACKUP DATABASE command is optimal for a 1.3 TB database that is stored on an SSD storage array, and is backed up to a container storage pool on a large blueprint server. The aggregate backup throughput exceeds 1300 MB/sec.

```
db2 backup db DATABASE1 online use TSM open 20 sessions
```

Testing shows that the previous db2 backup command cumulatively resulted in more than a 5 to 1 data reduction with daily full backups over 10 days. Also, compression contributed significantly to the overall data reduction savings. Close to 10 TB of storage space was saved.



Totals after 10 backups	Stored (GB)	Saved (GB)	%Reduced
No reduction	12,326.1	n/a	n/a
Data deduplication only	4,357.3	7,968.9	64.7%
Dedup and compression	2,131.1	10,195.1	82.7%

DB2 backup database options to avoid or disable for data reduction and performance purposes:

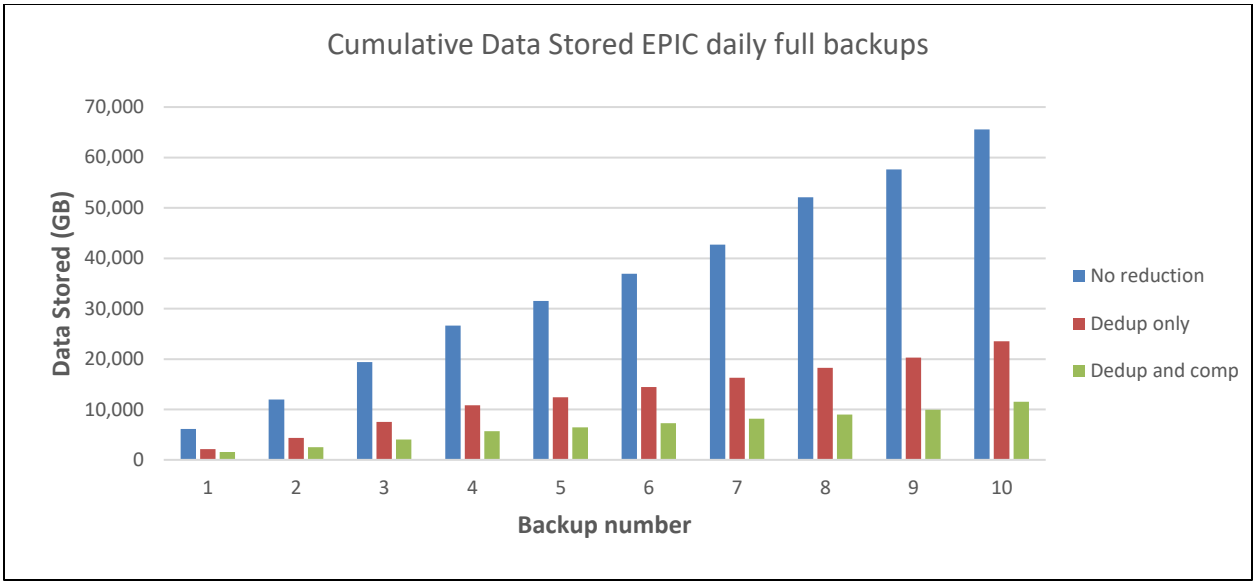
Compress
Encrypt
Dedup device

4.3.3 EPIC DB

EPIC client backups that use FlashCopy, and a dedicated proxy client to perform daily full image backups with the backup-archive client, can experience significant data reduction. A resourceutilization greater than 10 can be used if there are more than 10 file systems to back up in parallel and the datamover can handle the additional workload.

RESOURCEUTILIZATION	10
IMAGEGAPSIZE	128K
SNAPSHOTCACHESIZE	1

The following results from a customer show how data deduplication and compression cumulatively resulted in more than a 5 to 1 data reduction with daily full backups over 10 days. Close to 54 TB of storage space was saved.



Totals after 10 backups	Stored (GB)	Saved (GB)	%reduced
No reduction	65,587.7	n/a	n/a
Data deduplication only	23,551.1	42,036.6	64.1%
Dedup and compression	11,518.2	54,069.5	82.4%

4.3.4 Microsoft SQL

Two backup methods are supported with Microsoft SQL, VSS and Legacy.

With VSS backups, use the RESOURCEUTILIZATION option in the client API option file to tune performance. Recommended starting values for performance:

```
RESOURCEUTILIZATION 10
```

With legacy backups, use the STRIPES option to tune performance. The sqlbuffersize and buffersize options also affect performance and data reduction rates. A larger buffer size typically offers better data deduplication reduction, though it can have a larger impact to the disk IO during the backup. Recommended starting values for performance.

```
Stripes=4
Sqlbuffersize=1024
Buffersize=1024
```

Options to avoid or disable for data reduction and performance purposes:

```
SQLCOMPression
WITH COMPRESSION
ENCRYPTION
```

4.3.5 Oracle RMAN

With Oracle RMAN, use the CHANNELS and SECTION SIZE options to tune backup performance and data reduction. To improve data reduction, the section size must be less than 100 GB to avoid dedup tier 2 processing. Also, the section size improves backup parallelization of large table spaces. Use Oracle RMAN section size to send a single data file across multiple channels. Section size = total data size/channels (if greater than 100 GB, use less than 100 GB). Recommended starting values for performance:

```
Channels 10
SECTION SIZE 10 GB
```

RMAN options to avoid or disable for optimal data reduction and performance purposes:

```
Compression
Encryption for database
Encryption for tablespace
```

4.3.6 SAP HANA

With SAP HANA, the default options result in poor data reduction results due to the small default setting for the BUFFSIZE option. Testing demonstrates excellent data savings are possible with a buffer size up to 8 MB. However, this change has the side-effect of causing significantly more data to be stored during log backups. To overcome this issue, use a separate options file for data and log file backups. The only option that needs to vary between these two files is the BUFFSIZE setting. By default, SAP HANA backups will not use multiple sessions, which can limit backup throughput. SAP HANA also limits the use of multiple sessions to database that is larger than 128 GB. The following are recommended starting values provide good performance and reduction for repeated full backups, which approaches 80% savings. In the HANA administration studio, set the following options:

```
data_backup_parameter_file  /usr/sap/ALM/SYS/global/hdb/opt/hdbconfig/initALM.utl
log_backup_parameter_file   /usr/sap/ALM/SYS/global/hdb/opt/hdbconfig/initALM_log.utl
parallel_data_backup_backint_channels  4
```

In the TDP for HANA option file for data backups, include the following options:

```
MAX_SESSIONS            4
MAX_BACK_SESSONS        4
MAX_ARCH_SESSIONS       2
RL_COMPRESSIONS         NO
MULTIPLEXING            1
BUFFSIZE                8388608
```

In the TDP for HANA option file for log backup, include the same options except for:

```
BUFFSIZE                262144
```

4.3.7 VMware

The IBM Spectrum Protect Data Protection for VMware product is suited for container storage pools because it optimizes restores by keeping data permanently on a disk-based storage pool. Significant data reduction is possible due to the nature of data in virtual environments. Frequent use of cloning and templates results in redundant data that tends to deduplicate well. Data Protection for VMware uses VMware's Change Block Tracking (CBT) capability to only back up changed disk sectors, further reducing the amount of backup data.

Because a dedicated data mover is used in the solution, client-side data deduplication and compression are usually a good choice. Only enable client-side compression if client-side data deduplication is also enabled.

Some tips:

- A container storage pool is suitable for both the data and control files from DP for VMware backups.
- Use the VMMAXPARALLEL option to drive parallel sessions. Also, beginning with DP for VMware 8.1, two new options VMMAXBACKUPSESSIONS and VMMAXRESTORESESSIONS allow for multiple parallel sessions per virtual disk to optimize the backup and restore performance for large virtual machines.
- Use the VMLIMITPERHOST and VMLIMITPERDATASTORE options to more evenly distribute backups across multiple hosts and data stores.

5 Monitoring container storage pools

Several helpful capabilities are available for monitoring the health of a container storage pool. The sections that follow provide an overview of commands available for monitoring and an explanation of how to interpret the results. In addition, the Operations Center provides powerful capabilities for monitoring and alerting for container storage pools. The following indicators of container pool health must be monitored:

- Available storage pool space
- Whether storage pool protection and node replication are keeping pace with daily ingest. Replication can be monitored with the Operations Center web interface.
- The effectiveness of data reduction.

Since the scope of data deduplication includes multiple backups across multiple hosts, it takes time to accumulate sufficient data in the container storage pool to be effective at eliminating duplicates. Therefore, it is important to sample results at regular intervals to obtain a valid report of the results.

5.1 Simple Server Queries

5.1.1 Query Stgpool

The QUERY STGPOOL command provides a basic and quick method for evaluating storage pool free space and data reduction results.

Example command:

```
QUERY STGPOOL deduppool format=detailed
```

Example output:

```
Storage Pool Name:                      DEDUPPOOL
Storage Pool Type:                      Primary
Device Class Name:
Storage Type:                           DIRECTORY
Cloud Type:
Cloud URL:
Cloud Identity:
Cloud Location:
Estimated Capacity:                     45,806 G
Space Trigger Util:
Pct Util:                               40.8
< ... >
Additional space for protected data:    37,611 M
Total Unused Pending Space:             32,660 M
Deduplication Savings:                  145,093 G (79.33%)
Compression Savings:                    19,175 G (50.71%)
Total Space Saved:                      164,268 G (89.81%)
```

Data reduction statistics for a container storage pool are broken down into three fields. The three fields allow you to see the contribution to data reduction from data deduplication and compression individually, as well as, the combined total reduction from the two. The compression savings percentage is based on the amount of reduction on the data that remained after the data deduplication savings. In the example above, after the data deduplication reduction, the remaining data is compressed by 50.71%. The total space saved field indicates the combined amount of space that is saved in a pool from both data deduplication and compression. In the example above, the total savings percentage is related to the data deduplication and compression savings as follows:

89.81% = 79.33% + (100% - 79.33%) * 50.71%

5.1.2 Dedup Stats

With directory and cloud container storage pools, detailed node and file space data reduction statistics can be generated for later reporting. A significant amount of database processing is required during generate dedupstats processing, so these commands are intended to be issued on a periodic basis and are not practical for frequent reporting. The following commands are available:

- GENERATE DEDUPSTATS: Use this command to generate data reduction statistics
- QUERY DEDUPSTATS: Use this command to display data reduction statistics
- DEDUP DEDUPSTATS: Use this command to delete or prune data reduction statistics

The following example shows how the GENERATE DEDUPSTATS command can be run against a group of nodes that use multiple processes. After the generate dedupstats process completes, the QUERY DEDUPSTATS command to view the results for a single node.

```
GENERATE DEDUPSTATS deduppool node_grp1 maxprocess=10
QUERY DEDUPSTATS stgpool node_name format=detailed
```

Example output:

```
Date/Time:                          01/17/2017 16:07:07
Storage Pool Name:                  STGPOOL
Node Name:                          NODE_NAME
Filespace Name:                     \\NODE\c$
FSID:                               1
Type:                               Bkup
Total Data Protected (MB):          79,924
Total Space Used (MB):              41,302
Total Space Saved (MB):             38,622
Total Saving Percentage:            48.32
Deduplication Savings:              23,571,634,898
Deduplication Percentage:           28.13
Non-Deduplicated Extent Count:      39,218
Non-Deduplicated Extent Space Used: 35,720,848
Unique Extent Count:                157,226
Unique Extent Space Used:           54,990,126,334
Shared Extent Count:                182,625
```

```
Shared Extent Data Protected:        28,780,085,996
Shared Extent Space Used:            5,186,090,014
Compression Savings:                 16,926,094,178
Compression Percentage:              28.10
Compressed Extent Count:             339,660
Uncompressed Extent count:           39,409
```

5.1.3 Query Container

The QUERY CONTAINER command is used to display information about a container.

An example is shown here:

```
Q CONTAINER * f=d
```

Example output:

```
Container:                   /tsminst1/TSMfile00/09/0000000000000923.dcf
Storage Pool Name:           DEDUPPOOL
Container Type:              Dedup
State:                       Available
Free Space(MB):              1
Maximum Size(MB):            10,240
Approx. Date Last Written:   02/04/2017 18:55:25
Approx. Date Last Audit:
Cloud Type:
Cloud URL:
Space Utilized (MB):
Object Count:
```

5.1.4 Query Damaged

The QUERY DAMAGED command displays information about damaged extents in a container storage pool. The query damaged command has three different output summaries by inventory, node, and container. The default, inventory, displays the sum of damaged dedup and non-dedup extents. The node option lists the client or node data that is affected by the damage. Type container lists the containers that are affected by the damaged data.

An example is shown here:

```
QUERY DAMAGED deduppool type=inventory
```

Example output:

```
Storage Pool  Non-Dedup Data   Dedup Data     Cloud Orphaned
Name          Extent Count     Extent Count   Extent Count
------------  --------------   --------------  --------------------
DEDUPPOOL     58               145
```

5.1.5 Query ExtentUpdates

The QUERY EXTENTUPDATES command displays useful information for monitoring the progress of deleting unused extents. The number of extents pending update output indicates the number of extents that are pending an update to their reference count. The number of extents not referenced indicates extents that are no longer referenced and will be eligible for deletion after the reuse delay duration is met. The number of extents eligible for deletion is the queue of extents that are no longer referenced and exceeds the reuse delay period and are awaiting removal. An example is shown here:

```
QUERY EXTENTUPDATES deduppool
```

Example output:

```
Number of Extents Pending Update:          2,874,058
Number of Extents Not Referenced:          2,600,418
Number of Extents Eligible for Deletion:   850,565
Extent Reuse Delay (Days):                 1
```

5.2 IBM Spectrum Protect client reports

For client-side data deduplication, the client summary report shows the data reduction that is associated with data deduplication as well as compression.

An example is shown here:

```
Total number of objects inspected:        38,208
Total number of objects backed up:        38,208
Total number of objects updated:          0
Total number of objects rebound:          0
Total number of objects deleted:          0
Total number of objects expired:          0
Total number of objects failed:           0
Total number of objects encrypted:        0
Total objects deduplicated:               23,746
Total number of objects grew:             0
Total number of retries:                  0
Total number of bytes inspected:          77.05 GB
Total number of bytes processed:          56.52 GB
Total bytes before deduplication:         77.04 GB
Total bytes after deduplication:          56.51 GB
Total number of bytes transferred:        40.56 GB
Data transfer time:                       171.87 sec
Network data transfer rate:               247,425.94 KB/sec
Aggregate data transfer rate:             53,226.12 KB/sec
Objects compressed by:                    29%
Deduplication reduction:                  26.65%
Total data reduction ratio:               47.37%
Elapsed processing time:                  00:13:18
```

Client reports can also be collected from the server extended summary table by using a SELECT command. The following fields provide backup or archive statistics:

- BYTES_PROTECTED: <Bytes that have been protected prior to data reduction>
- BYTES_WRITTEN: <Amount of data remaining after data reduction>
- DEDUP_SAVINGS: <Savings from deduplication processing>
- COMP_SAVINGS: <Savings from compression>

An example is shown here:

```
SELECT * from summary_extended where entity='NODE_NAME' and activity IN ('BACKUP','ARCHIVE')
```

Example output:

```
START_TIME:                     2017-01-17 11:04:14.000000
END_TIME:                       2017-01-17 11:24:02.000000
ACTIVITY:                       BACKUP
ACTIVITY_DETAILS: SESSION_LIST: 22479,22478,22484,22483,22481,22480,22477
ACTIVITY_TYPE:                  SESSION_END
NUMBER:                         22472
ENTITY:                         NODE_NAME
AS_ENTITY:
SUB_ENTITY:
COMMMETH:                       Tcp/Ip
ADDRESS: NODE_IP:               51254
SCHEDULE_NAME:
EXAMINED:                       38208
AFFECTED:                       38208
FAILED:                         0
BYTES:                          43544212649
BYTES_PROTECTED:                82752759497
BYTES_WRITTEN:                  43275409789
DEDUP_SAVINGS:                  22042701708
COMP_SAVINGS:                   17148564126
IDLE:                           1775
MEDIAW:                         0
PROCESSES:                      9
COMPLETION_CODE:                0
SUCCESSFUL:                     YES
VOLUME_NAME:
DRIVE_NAME:
LIBRARY_NAME:
LAST_USE:
COMM_WAIT:                      1108
NUM_OFFSITE_VOLS:
```

5.3 Summary Table Queries

The following SQL queries can be collected from the server extended summary table by using the SELECT command. These queries are low cost, and can be added to the Operations Center custom reports. The queries below provide daily ingest and data reduction statistics, as well as identify clients with low data deduplication or compression rates. Further details on clients with low data reduction rates can be examined with the QUERY DEDUPSTATS command.

Daily Client Workload: Provides a daily aggregate client workload and data reduction statistics for the server. An example is shown here:

```
SELECT DATE(s.START_TIME) AS Date, (CAST(FLOAT(SUM(s.bytes_protected))/1024/1024 AS DECIMAL(12,2))) AS
PROTECTED_MB, (CAST(FLOAT(SUM(s.bytes_written))/1024/1024 AS DECIMAL(12,2))) AS WRITTEN_MB,
(CAST(FLOAT(SUM(s.dedup_savings))/1024/1024 AS DECIMAL(12,2))) AS DEDUPSAVINGS_MB,
(CAST(FLOAT(SUM(s.comp_savings))/1024/1024 AS DECIMAL(12,2))) AS COMPSAVINGS_MB,
(CAST(FLOAT(SUM(s.dedup_savings))/FLOAT(SUM(s.bytes_protected))*100 AS DECIMAL(5,2))) AS DEDUP_PCT,
(CAST(FLOAT(SUM(s.bytes_protected) - SUM(s.bytes_written))/FLOAT(SUM(s.bytes_protected))*100 AS
DECIMAL(5,2))) AS SAVINGS_PCT from summary s WHERE activity='BACKUP' or activity='ARCHIVE' GROUP BY
DATE(S.START_TIME)
```

Example output:

DATE	PROTECTED_MB	WRITTEN_MB	DEDUPSAVINGS_MB	COMPSAVINGS_MB	DEDUP_PCT	SAVINGS_PCT
2017-02-04	81544974.73	18618531.50	45393524.18	17532884.53	55.66	77.16
2017-02-05	145352668.34	35421956.45	76472838.00	33447607.20	52.61	75.63
2017-02-06	100983732.79	27366562.23	47030394.41	26562826.58	46.57	72.90
2017-02-07	112150778.74	27995259.37	57687301.06	26468172.06	51.43	75.03
2017-02-08	66595860.38	16897757.29	33717502.22	15976329.51	50.63	74.62

< … >

25 Worst Deduplicating Nodes sorted by data deduplication percent: Provides a list of client nodes that are deduplicating poorly. An example is shown here:

```
SELECT SUBSTR(s.ENTITY,1,10) AS NODE, (CAST(FLOAT(SUM(s.bytes_protected))/1024/1024/1024 AS
DECIMAL(12,2))) AS PROTECTED_GB, (CAST(FLOAT(SUM(s.dedup_savings))/1024/1024/1024 AS DECIMAL(12,2)))
AS DEDUPSAVINGS_GB, (CAST(FLOAT(SUM(s.comp_savings))/1024/1024/1024 AS DECIMAL(12,2))) AS
COMPSAVINGS_GB, (CAST(FLOAT(SUM(s.dedup_savings)) / FLOAT(SUM(s.bytes_protected))*100 AS
DECIMAL(5,2))) AS DEDUP_PCT, (CAST(FLOAT(SUM(s.comp_savings)) / FLOAT(SUM(s.bytes_protected)-
SUM(s.dedup_savings))*100 AS DECIMAL(5,2))) AS COMP_PCT from summary_extended s WHERE
activity='BACKUP' or activity='ARCHIVE' GROUP BY S.ENTITY ORDER BY DEDUP_PCT ASC FETCH FIRST 25 ROWS
ONLY
```

Example output:

NODE	PROTECTED_GB	DEDUPSAVINGS_GB	COMPSAVINGS_GB	DEDUP_PCT	COMP_PCT
CF1	35.09	0.01	0	0.02	0
CETVM68	77.06	20.52	15.97	26.63	28.24
C2T1	634.91	286.52	0	45.12	0
C2T2	621.76	286.82	0	46.13	0
C2T148	770.3	357.68	0	46.43	0

<…>

25 Worst Compressed Nodes by compression percent: Provides a list of client nodes that are compressing poorly. An example is shown here:

```
SELECT SUBSTR(s.ENTITY,1,10) AS NODE, (CAST(FLOAT(SUM(s.bytes_protected))/1024/1024/1024 AS
DECIMAL(12,2))) AS PROTECTED_GB, (CAST(FLOAT(SUM(s.dedup_savings))/1024/1024/1024 AS DECIMAL(12,2)))
AS DEDUPSAVINGS_GB, (CAST(FLOAT(SUM(s.comp_savings))/1024/1024/1024 AS DECIMAL(12,2))) AS
COMPSAVINGS_GB, (CAST(FLOAT(SUM(s.dedup_savings)) / FLOAT(SUM(s.bytes_protected))*100 AS
DECIMAL(5,2))) AS DEDUP_PCT, (CAST(FLOAT(SUM(s.comp_savings)) / FLOAT(SUM(s.bytes_protected)-
SUM(s.dedup_savings))*100 AS DECIMAL(5,2))) AS COMP_PCT from summary_extended s WHERE
activity='BACKUP' or activity='ARCHIVE' GROUP BY S.ENTITY ORDER BY COMP_PCT ASC FETCH FIRST 25 ROWS
ONLY
```

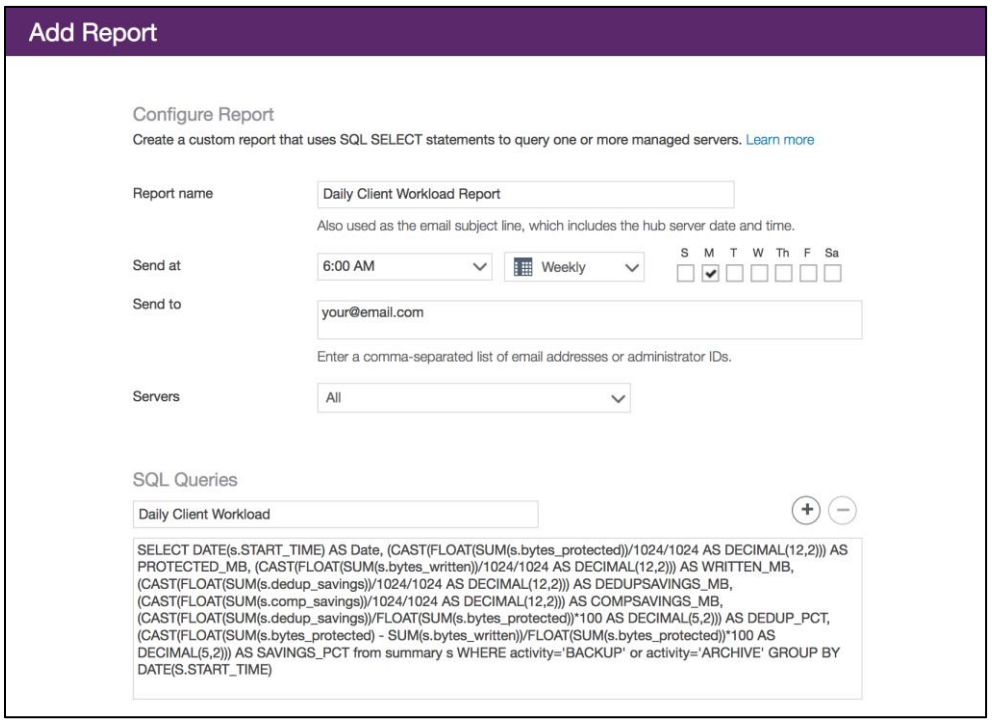
Example output:

NODE	PROTECTED_GB	DEDUPSAVINGS_GB	COMPSAVINGS_GB	DEDUP_PCT	COMP_PCT
C2T1	634.91	286.52	0	45.12	0
C2T120	575.31	286.24	0	49.75	0
C2T12	720.39	357.76	0	49.66	0
C2T119	719.39	357.46	0	49.68	0
C2T118	719.37	358.11	0	49.78	0

<...>

5.4 Operations Center and custom reports

The IBM Spectrum Protect Operations Center can use SQL select statements to query the server database to create custom email reports at a specified interval and time. For example, the SQL statements from the above sections, 4.2 and 4.3, can be used for custom email reports.



< End of Document>