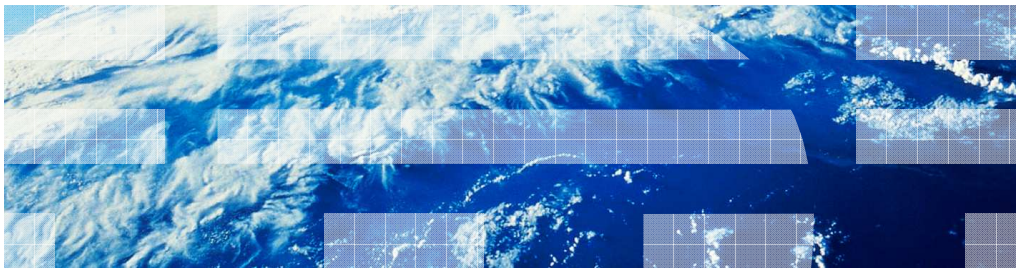
IBM Tivoli Storage Manager 6.2

Server-side data deduplication



Tivoli software

IBM Tivoli® Storage Manager 6.2 Server-side data deduplication.

Assumptions

You are familiar with Tivoli Storage Manager version 5.5 or higher

You are familiar with Tivoli Storage Manager version 5.5 or higher.

Objectives

After completing this module, you should be able to:

- Describe the deduplication process
- List the benefits of deduplication
- Explain the difference between server and client-side methods
- Set client and server options
- Configure primary and copy storage pools for deduplication

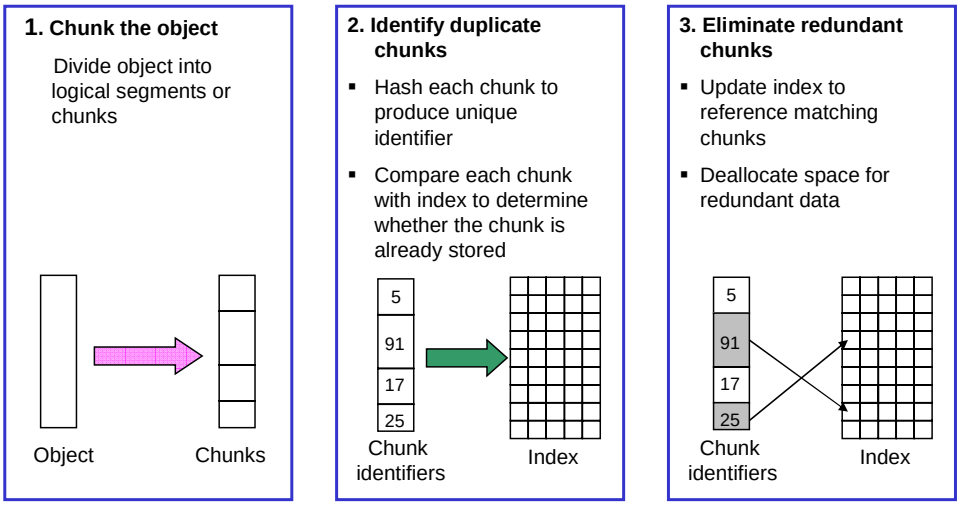
After completing this module, you should be able to describe the deduplication process, list the benefits of deduplication, explain the difference between server and client-side methods, set client and server options, and configure primary and copy storage pools for deduplication.

Overview of deduplication

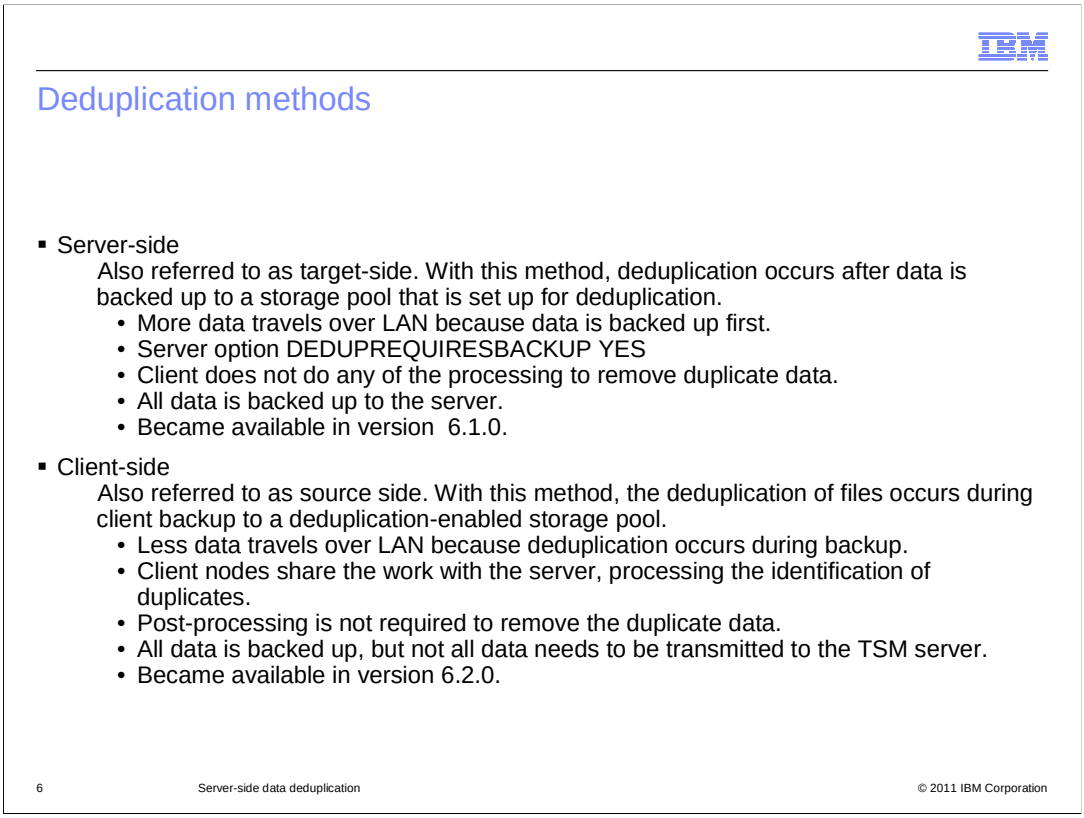
- Data deduplication is a method of eliminating redundant data in sequential-access disk (FILE device class) primary, copy, and active-data storage pools.
- One unique instance of the data is retained on storage media. Redundant data is replaced with a pointer to the unique data copy.
- The goal of deduplication is to:
  - Reduce the amount of data you need to store.
  - Reduce the overall amount of time that is required to retrieve data by letting you store more data on disk, rather than on tape.
- Deduplication is available with Tivoli Storage Manager Extended Edition V6.1 and later.

IBM Tivoli Storage Manager has always done a very good job of reducing the amount of data you need to store by providing progressive incremental backup of client data. This process works at a file level to reduce the amount of space required for storing client data. With Tivoli Storage Manager Extended Edition version 6.1, IBM introduced server-side data deduplication. This process works at a block level to identify redundant blocks of data and replace those with pointers to a single unique instance of the data. This is intended to not only further reduce the amount of storage space required, but to reduce the recovery time by keeping more data on disk. You can deduplicate any type of data except encrypted data. You can deduplicate client backup and archive data, Tivoli Data Protection data, and so on. Tivoli Storage Manager can deduplicate whole files and files that are members of an aggregate. You can deduplicate data that has already been stored, and no additional backup, archive, or migration is required.

Duplicate data chunk hashing

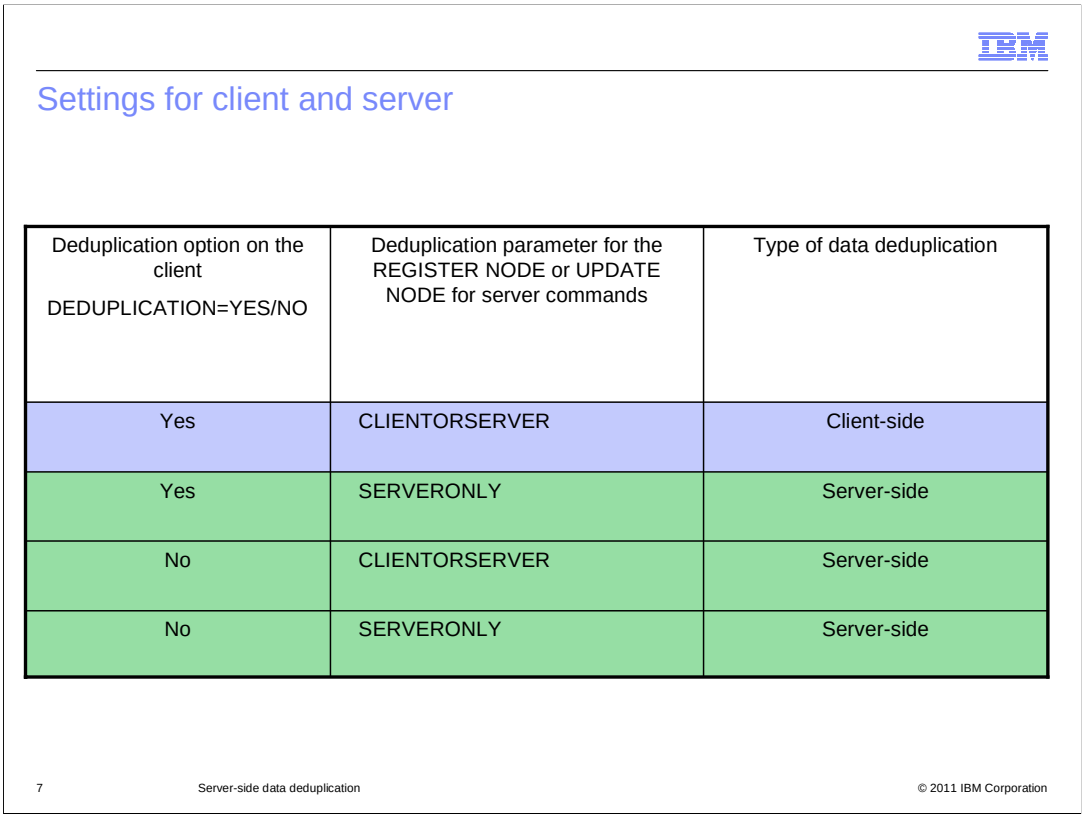


Chunk is a term used to describe data objects that have been divided into logical segments. Each chunk is hashed to produce a unique identifier. These chunks are compared to the index to identify duplicates. The redundant chunks are eliminated, and the index is updated. Only files larger than 2 KB are deduplicated.

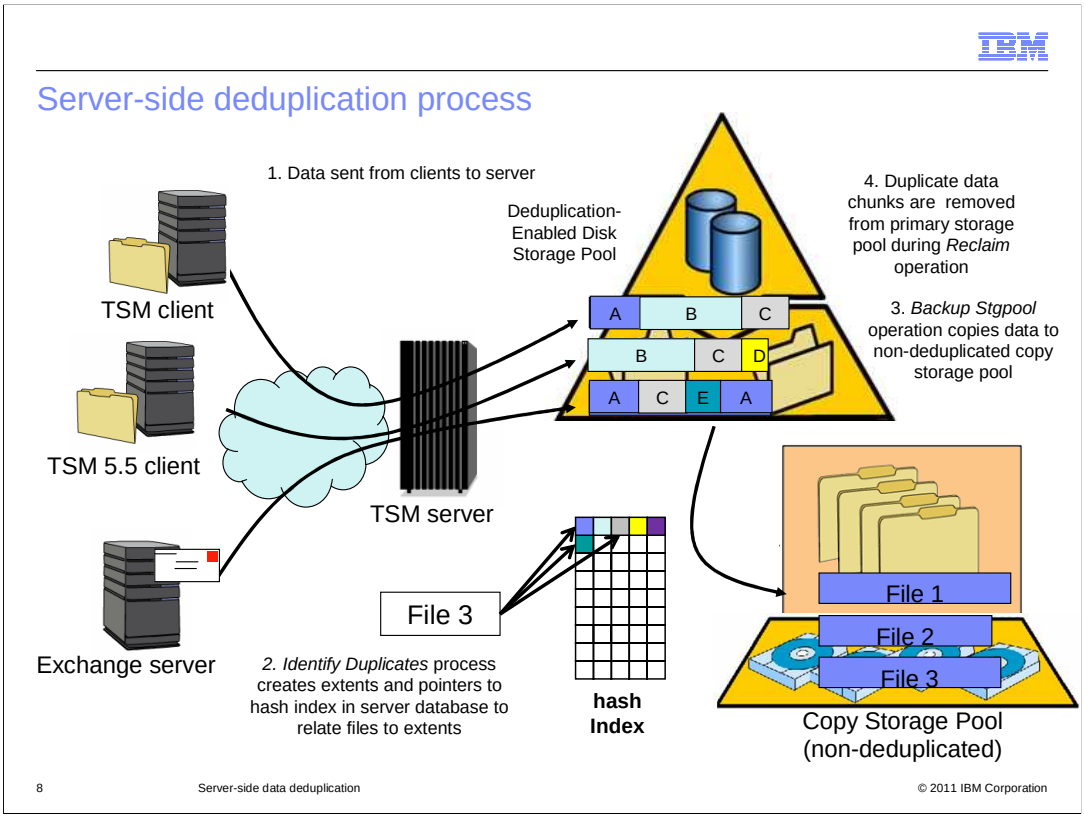


There are two data deduplication methods, server-side which was introduced with 6.1 and client-side which was introduced at 6.2. The main difference between these two methods is where the duplicate identification process occurs. With the server-side method, the process occurs after data is backed up to a deduplication enabled storage pool, and the server does all the processing. The primary storage pool is backed up before the duplicate identification process runs. This backup will occur only if the server option DEDUPREQUIRESBACKUP is set to YES. This is the default setting.

With the Client-side method, duplicate blocks are identified during client backup processing and the client shares the workload. So with the client-side method, the data is already reduced before it travels over the LAN.



You can enable deduplication using a combination of settings on the client node and the server. If there is a problem with client-side deduplication, the client data backup will still run, but the deduplication will occur at the storage pool or server-side.



With server-side deduplication, all of the processing is done by the server. The server identifies the duplicate chunks of data after client backup processing has copied the data to a disk storage pool that has been set up for deduplication. If a user restores a file that exists in a deduplicated storage pool, the server does the work of reconstructing the file.

The process begins when the Data is sent from the clients to the server.

The server creates a hash index in the server database which has pointers to chunks in the storage pool. The pointers are used to relate chunks to the actual backup files. In this scenario deduprequiresbackup has been set to yes, so the storage pool must be backed up before deduplication can proceed.

A backup storage pool operation copies data to a non-deduplicated copy storage pool.

Duplicate data chunks are removed from primary storage pool during reclamation.

With this approach, deduplication is performed out-of-band, and at least one copy of non-deduplicated data exists.

Removing duplicates

Duplicate data is removed by any of the following processes:

- Reclamation of volumes in the primary storage pool, copy storage pool, or active-data pool.
- Backing up of a primary storage pool to a copy storage pool that is also set up for deduplication.
- Copying of active data in the primary storage pool to an active-data pool that is also set up for deduplication.
- Migrating data from the primary storage pool to another primary storage pool that is also set up for deduplication.
- Moving data from the primary storage pool to a different primary storage pool that is also set up for deduplication, moving data within the same copy storage pool, or moving data within the same active-data pool.

After the duplicates have been identified, the duplicate data can be removed by reclamation or when you back up, copy, migrate or move data from the primary storage pool to a storage pool that is configured for deduplication.

Planning for server-side deduplication

Before setting up storage pools:

- Determine which client nodes will use server-side deduplication.
- Decide whether you want to define a new storage pool exclusively for deduplication or update an existing storage pool.
- Create a policy that points to the correct storage pool
- Decide how you want to control duplicate-identification processes:
  - Automatically
  - Manually

Before setting up storage pools for deduplication you need to decide which clients will use deduplication and which method to use for which client. The clients that use deduplication will need to be associated with a policy that has the storage destination in the backup copygroup set to the deduplication enabled storage pool. You also need to decide how you want to control duplicate-identification processes.

Storage pool considerations

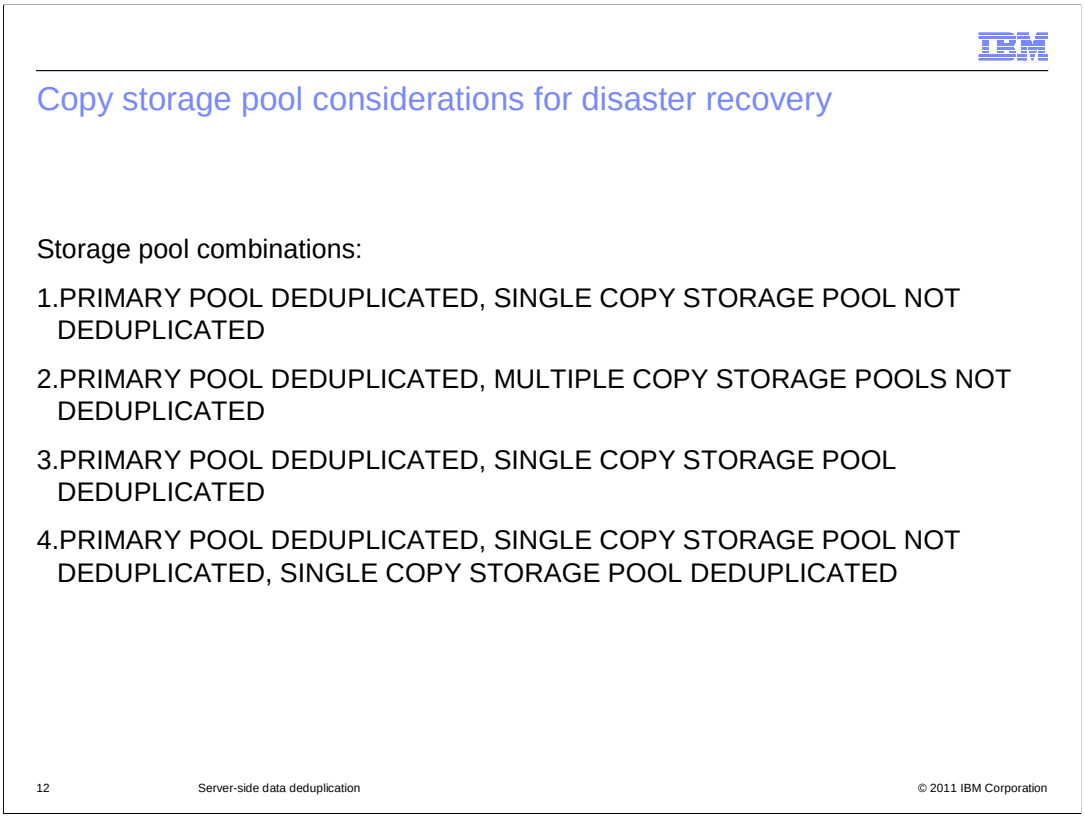
Decide which storage pools will run the deduplication-identification process.

- If you have a primary sequential-access disk storage pool and a copy sequential-access disk storage pool, and both pools are set up for deduplication, you might want to run duplicate-identification processes for the primary storage pool only. In this way, only the primary storage pool will read and deduplicate data.
- When the data is moved to the copy storage pool (that is deduplication enabled), the deduplication is preserved. No duplicate identification is required.
- If you plan to use the new Simultaneous Write Migration, that data cannot go to a deduplication-enabled storage pool.
  - AUTOCOPY option on the storage pool will be disabled.
  - Warning message will be issued.

DEDUPlicate = No/Yes
IDENTIFYPRocess = *number*

Because the process of identifying duplicates requires extra disk I/O and CPU resources, Tivoli Storage Manager lets you control when the identification processing begins and the number and duration of processes. You can run duplicate-identification processes automatically all the time, or you can start and stop processes manually. By default, a duplicate-identification process begins automatically after you define a storage pool for deduplication. If you specify a value for IDENTIFYPRocess when you update a storage pool, it also starts automatically. You can also start duplicate-identification processes automatically and then increase or decrease the number of processes depending on your server workload. In the storage pool definition, you can specify as many as 20 duplicate identification processes.

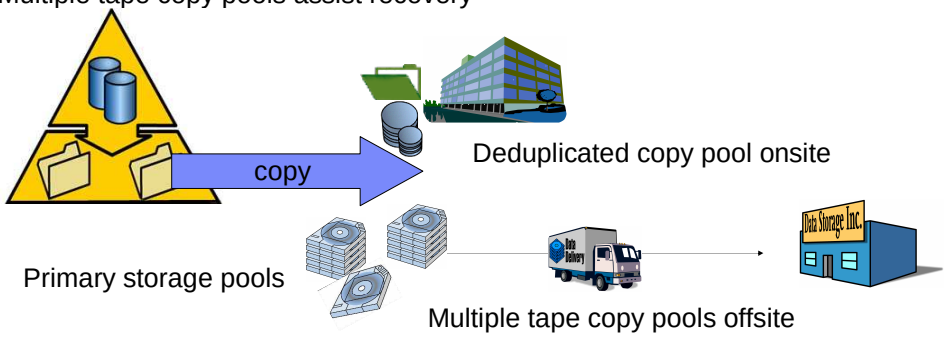
If you plan to use Simultaneous Write Migration, you cannot use it with a deduplication-enabled storage pool.



The storage pools you define or update for deduplication must be a sequential-access disk (FILE device class) storage pool. Because of this, you might want to consider the implications for disaster recovery. The data will be reconstructed if it goes to tape, but having offsite copies is critical step to protect your server and your data from a site disaster.

Copy storage pool customization

- Create custom copy pool structures for:
  - Storage pools with deduplicated data
  - Active data storage pools
  - For primary pools that use Simultaneous Write Migration
- Deduplicated for onsite
- Non-deduplicated to tape for offsite
- Multiple tape copy pools assist recovery



13 Server-side data deduplication

When creating copy storage pools it is important to consider the settings and function of the primary storage pool you are copying. Consider creating a separate copy pool structure for deduplicated data pools, and using multiple copy pools for your business critical/important copy pool data. This will reduce the tape contention, and during a site disaster recovery it will reduce the number of copy pool tapes required within the library during restore operations.

Query to display information about deduplication

- QUERY STGPOOL
  - Storage pool settings
  - Number of processes specified
  - Amount of data removed during reclamation
- QUERY PROCESS
  - Number of bytes and files processed
- QUERY CONTENT with the FOLLOWLINKS parameter

You can use the query stgpool command to obtain important statistics about data deduplication. This storage pool information will show you if the storage pool has been enabled for deduplication, the number of processes specified in the storage pool definition, and the amount of data that was removed by reclamation processing.

You can use the query process command to display the total number of bytes and total number of files processed.

Finally, you can query a volume for information about client files that link to files on other volumes. This information is useful when file extents created by data deduplication are distributed on different volumes. You can display information only about files that are linked to a volume or only about files that are stored on a volume. You can also display information about both stored files and linked files. To display information about files on a volume, issue the QUERY CONTENT command and specify the **FOLLOWLINKS** parameter.

Summary

You should now be able to:

- Describe the deduplication process
- List the benefits of deduplication
- Explain the difference between server and client-side methods
- Set client and server options
- Configure primary and copy storage pools for deduplication

You should now be able to describe the deduplication process, list the benefits of deduplication, explain the difference between server and client-side methods, set client and server options, and configure primary and copy storage pools for deduplication.

Feedback

Your feedback is valuable

You can help improve the quality of IBM Education Assistant content to better meet your needs by providing feedback.

- Did you find this module useful?
- Did it help you solve a problem or answer a question?
- Do you have suggestions for improvements?

Click to send email feedback:

mailto:iea@us.ibm.com?subject=Feedback_about_server-side_dedup.ppt

This module is also available in PDF format at: ../server-side_dedup.pdf

You can help improve the quality of IBM Education Assistant content by providing feedback.

Trademarks, disclaimer, and copyright information