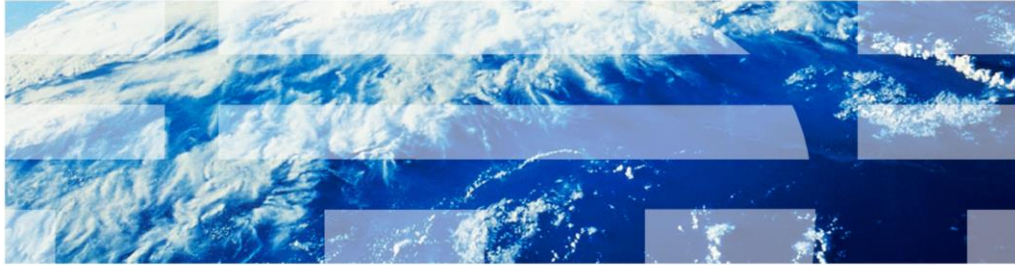


IBM Tivoli Storage Manager 6.2

Client-side data deduplication



© 2011 IBM Corporation

In this module, you learn to implement Tivoli Storage Manager client-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


When you complete this module, you can perform these tasks:

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

When you complete this module, you can perform these tasks:

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

Where data deduplication is performed

Approach	Advantages	Disadvantages
<u>Target-side (server-side)</u> Deduplication performed at the Tivoli Storage Manager Server or storage device after receiving backup data from client	<ul style="list-style-type: none"> ▪ No new deployment of Tivoli Storage Manager client software ▪ Possible use of direct comparison to confirm duplicates 	<ul style="list-style-type: none"> ▪ Deduplication uses CPU cycles on the Tivoli Storage Manager server or storage device ▪ Data might be discarded after being transmitted to the target
 <u>Source-side (client-side)</u> Deduplication performed at the Tivoli Storage Manager backup client, before transfer to Tivoli Storage Manager server	<ul style="list-style-type: none"> ▪ Deduplication before transmission conserves network bandwidth ▪ Awareness of data usage and format can provide more effective data reduction ▪ Processing at the source can facilitate scale-out 	<ul style="list-style-type: none"> ▪ Deduplication uses CPU cycles on the client ▪ Requires software deployment or upgrade at Tivoli Storage Manager client system

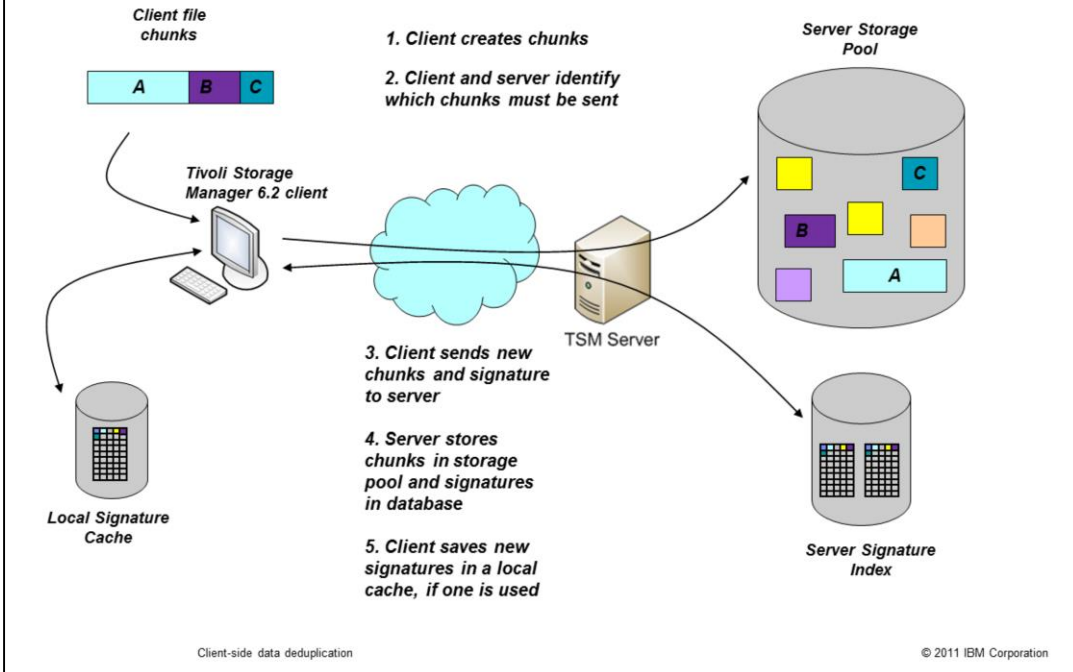
Note: Source-side and target-side deduplication are not mutually exclusive

Client-side data deduplication

© 2011 IBM Corporation

You can use data deduplication to identify duplicate parts, or chunks, of a file or multiple files. You can store a single copy of that chunk during backup and reduce the amount of storage that is required for several backup files. You can perform both server-side deduplication and client-side deduplication in a single Tivoli Storage Manager installation. Deduplication of data can also occur on either the server side or on the client side of the backup process. On the server side, deduplication happens after the data is transmitted to the Tivoli Storage Manager server. Deduplication occurs in the Tivoli Storage Manager server software or in a storage device, such as a ProtecTIER virtual tape library. One advantage of performing deduplication at the server is that you do not need additional software on the client side. Also, deduplication can occur on data that has been stored previously. Another advantage is that the data can be compared directly to existing backup data to ensure that a perfect match is made in the deduplication process. However, this method does require significant resources. The server or storage device is deduplicating data from many clients. On the other hand, if the deduplication process occurs on the client before transmission to the server, the network requirements are reduced. The resource requirements at the server are also reduced because most of the processing is done across many client systems. The server needs only to manage the storing of this reduced amount of backup data. With this processing now occurring on the client side, the clients performing this function must have available resources to accomplish the task. Client-side deduplication also requires that these clients are updated and configured with new software.

Client-side data deduplication operation



During the backup process, the client breaks up the file into smaller chunks and calculates a unique signature for each chunk. The server is queried to see if a chunk with that unique signature exists in the inventory of that server. The local signature cache file of the client is used optionally to reduce the frequency of server queries. If that chunk does not already exist, the client sends both the new chunk and its signature to the server. If, however, there is a matching chunk on the server, the server notes that this chunk is a part of this newly backed up file. New chunks are stored in the storage pool, and new signatures are added to the index of chunks in the database for that server. This signature can also be stored in the local signature cache for the client for use in a later backup.

Tivoli Storage Manager server requirements

- Tivoli Storage Manager server version 6.2 or higher
- Server configuration
 - Storage pool
 - Use device class of FILE
 - Specify DEDuplicate=YES
 - Use the destination that is specified in the copygroup
 - Server options
 - CLIENTDEDUPTXNLIMIT: Specify the maximum size of a transaction when client-side deduplicated data is backed up or archived. The default size is 50 GB
 - SERVERDEDUPTXNLIMIT: Specify the maximum size of objects that can be deduplicated on the server. The default size is 300 GB
 - DEDUPREQUIRESBACKUP: Specify whether volumes in primary sequential-access storage pools that are set up for data deduplication can be reclaimed. Specify whether duplicate data can be discarded before the storage pools are backed up. The default setting is YES
 - Command: SET dedupverificationlevel
 - Percentage of duplicate chunks sent to the server for verification. The default percentage is zero

Client-side deduplication requires interaction between both the Tivoli Storage Manager client and the Tivoli Storage Manager server. The server must be at version 6.2 or higher to provide support for this interaction. In the server configuration, you must provide a storage pool that is prepared for deduplication. The client backups are directed to this pool. A storage pool must be on disk storage, use a device class of type=file, and have the DEDuplicate parameter set to 'YES'. This storage pool must be the destination for either the backup or archive copygroup of the management class that the client is using. The client queries the server for information about the management class to determine if deduplication is possible with this initial destination storage pool. Optionally, you can set some options on the server to help control the deduplication process. For client-side deduplication, you can set client dedup transaction limit. You can use this option to limit the size of an incoming client transaction. You minimize long transactions that use server resources for a long time. This server-wide setting applies to all clients that perform the deduplication function. The 'server dedup transaction limit' parameter similarly limits the size of server-based transactions for server-side deduplication. One other option is the 'dedup requires backup' parameter. While this parameter does not affect the deduplication process itself, it has an effect on copy storage pools. If you want to create deduplicated copy storage pools, the parameter must be set to 'NO'. If the parameter is set to Yes, the non-unique chunks cannot be deleted from the Tivoli Storage Manager server until the file has been copied to a non-deduplicated copy storage pool. This parameter is ignored for client-side deduplication. You can use one SET command with client-side deduplication, which is SET dedup verification level. This command specifies what percentage of incoming nonunique objects are sent to the server. You can verify that the deduplication on the server has the same result as the deduplication on the client. If the results of the client process and the server process do not match, the originating client is prohibited from performing deduplication.

Tivoli Storage Manager client requirements

- Tivoli Storage Manager client version 6.2 or higher
- Client is registered on the server
 - DEDUPLICATION=Clientorserver
- Client options
 - DEDUPLICATION YES specifies that you want to enable client-side data deduplication for backup and archive processing
 - The default setting is NO
- Management class assignment
 - Initial destination of management class must be a storage pool that is set up for deduplication
- Client-side data deduplication is not used with these items:
 - Encrypted files
 - Subfile backups
 - Operations without LAN

The Tivoli Storage Manager client must also be installed at version 6.2 or later. The client node definition for this client must have the deduplication parameter set to 'CLIENTORSERVER'. When this information is on the server, the client can perform client-side deduplication. The default setting for this parameter is 'server only'. In the option file for the client, you must set the deduplication parameter to 'YES'. The server and client must specify that client-side deduplication can be performed by this client. The default setting for this parameter is 'NO'. As mentioned, a deduplication storage pool must be the initial destination for a backup or archive copygroup. Likewise, the client must direct its backups or archives to a management class that has such a destination set. In some cases, the Tivoli Storage Manager client does not attempt to deduplicate data. For example, the deduplication does not occur when backing up encrypted files, using the subfile backup technique, or when the backup process is being performed without LAN.

Tivoli Storage Manager client caching customization

- **ENABLEDEDUPCACHE YES**
 - A local cache is preferred during client-side data deduplication to reduce network traffic between the Tivoli Storage Manager server and the client
 - The cache is shared among client sessions
 - The default setting is YES for the backup/archive client

- **DEDUPCACHEPATH path**
 - You can specify the location where the client-side data deduplication cache database is created
 - The default location is client or API installation directory

- **DEDUPCACHESIZE mb**
 - You can specify the maximum size, in megabytes, of the data deduplication cache file
 - The range of values is 1 - 2048
 - The default value is 256

While a signature is calculated for each chunk of data, it is compared to signatures of chunks that are stored on the Tivoli Storage Manager server. This server query uses a separate session with the server so that it does not disrupt other ongoing backup transmissions. These unique signatures can also be cached to avoid too many of these server queries. This cache file is shared among all the concurrent client sessions within one Tivoli Storage Manager client process. However, it is not shared across multiple processes. The default setting is to enable this local caching function on the backup/archive client. The DEDUPCACHEPATH is, by default, created within the directory where the backup/archive client is installed. However, you can change the location by using this parameter in the options file for the client. The size of this cache file is 256 megabytes by default. In the option file, you can change the size to as much as two gigabytes.

Deduplication filtering

- Include/exclude processing
 - Include.dedup to include matching objects
 - Exclude.dedup to exclude matching objects
- IEObjtype parameter
 - Filter deduplication processing for additional items
 - File (default)
 - Image
 - SYSTEMObject
 - SYSTEMState
 - ASR
- Examples
 - Include.dedup c:\mydata\.*
 - Exclude.dedup ?:\.**.zip
 - Exclude.dedup * ieobjtype=asr

By default, all objects are candidates for deduplication. However, with the include/exclude processing, you can selectively deduplicate certain objects or exclude them from the deduplication process altogether. The syntax for deduplication filtering parameters is similar to other filtering options. You can specify an object type with the IE object type setting. The default object type is file. You can specify several other types, such as image, system object, system state, or ASR. In the examples shown here, all the objects in the mydata directory on the C:\ drive are to be deduplicated. All archive files are to be excluded wherever they exist. In the last example, you see how to exclude all objects that are part of the automated system recovery backup.

Deduplication with API programs

- You must use the Tivoli Storage Manager V6.2 API run time library
 - No requirement to compile with V6.2 source
 - Add DEDUPLICATION YES to the API option file
- Deduplication is transparent to API programs
 - Effectiveness of the deduplication process is available to the calling program
 - Not all data protection products display this information
- ENABLEDEDUPCache defaults to NO for the API
 - You can set to YES
 - Each separate Tivoli Storage Manager process requires its own cache
 - Data protection clients can create multiple processes rather than multiple sessions within a single process

Client-side deduplication is also useful for API programs, such as the Tivoli Data Protection products. You must be able to start these programs with the version 6.2 API run time library. However, you do not need to recompile them with the newer API source code. The only requirement is to add the deduplication parameter to the option file with a setting of 'YES'. The deduplication process is transparent to these programs. The data can potentially be deduplicated before being sent to the Tivoli Storage Manager server. However, these programs cannot display the results of the underlying deduplication activity without using updated version 6.2 API calls. You can use local caching of generated signatures with API programs. Unlike the backup/archive client, the default for the API is 'NO' because the local cache cannot be shared across multiple processes. These API programs create multiple processes as compared to the backup/archive client, which generates multiple sessions within a single process. The local caching can be enabled by setting the 'enable dedup cache' parameter to 'YES'. An individual cache file is available only to the first process to use it. Setting up multiple cache files is possible, but it is complex in most cases.

Considerations for Tivoli Storage Manager client deduplication

- Considerations for using deduplication
 - Network bandwidth is limited
 - Data recovery can improve by storing more data objects on a limited number of disks and avoiding tape processing delays
 - Data must remain on a disk for an extended time
 - Large redundancy in data stored by Tivoli Storage Manager
 - Similar clients can have identical operating system files
 - Clients can each store the same business documents
 - Tivoli Storage Manager client resources are available for intensive processing to identify duplicate chunks
- Considerations for not using deduplication
 - You have mission-critical data, whose recovery can be delayed by accessing chunks that are not stored contiguously
 - Tivoli Storage Manager clients do not have sufficient resources
 - Business requirements are for pristine backup images

Implementing Tivoli Storage Manager client-side deduplication has advantages and disadvantages. One advantage is that client-side deduplication can reduce the effective network bandwidth consumption by sending less data during the backup process. It also increases the effective amount of backup data that is stored in a given amount of disk storage. The client-side deduplication can improve performance for single file restores and extend the time that objects remain on a disk. The effectiveness of either client-side or server-side deduplication ultimately depends on the amount of duplication in the backup data. For instance, with similar clients backing up to the same Tivoli Storage Manager server, many of the system files can be identical across these clients. Also, if these clients keep the same or similar business documents, you can save space. If these clients have the resources to handle the intensive processing that is necessary for identifying duplicate data, they might be good candidates for client-side deduplication. On the other hand, clients with limited available resources might not benefit individually from client-side deduplication. You might have business requirements for the data that you are backing up. Mission critical data that needs a single stream to restore performance might be delayed by having the chunks of a single file stored discontinuously. You might also have business or legal reasons to maintain the backup copies in their exact, original state, and not broken into chunks that are shared with other data objects.

Summary

Now that you have completed this module, you can perform these tasks:

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

Now that you have completed this module, you can perform these tasks:

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

Trademarks, disclaimer, and copyright information

IBM, the IBM logo, ibm.com, Approach, ProtecTIER, and Tivoli are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of other IBM trademarks is available on the web at "[Copyright and trademark information](http://www.ibm.com/legal/copytrade.shtml)" at <http://www.ibm.com/legal/copytrade.shtml>

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS OR SOFTWARE.

© Copyright International Business Machines Corporation 2011. All rights reserved.