White Paper

NEXT-GENERATION STORAGE EFFICIENCY WITH EMC ISILON SMARTDEDUPE

Abstract

Most file systems are a thin layer of organization on top of a block device and cannot efficiently address data on a large scale. This paper focuses on EMC Isilon OneFS, a modern file system that meets the unique needs of Big Data. Isilon OneFS includes EMC Isilon SmartDedupe, a native data reduction capability, which enables enterprises to reduce their storage costs and footprint, and increase data efficiency without sacrificing data protection or management simplicity.

October 2013

Part Number H12395

Table of Contents

Introduction

Information technology managers across most areas of commerce are grappling with the challenges presented by explosive file data growth, which significantly raises the cost and complexity of storage environments. Business data is often filled with significant amounts of redundant information. For example, each time an email attachment is stored by multiple employees, many of the same files are stored or replicated. This leads to multiple copies of the same data, which take up valuable disk capacity. Data deduplication is a specialized data reduction technique that allows for the elimination of duplicate copies of data.

Deduplication is yet another milestone in the industry-leading data efficiency of EMC® Isilon® solutions, and it is a key ingredient for organizations that wish to maintain a competitive edge.

EMC Isilon SmartDedupe

Overview

EMC Isilon SmartDedupe software maximizes the storage efficiency of a cluster by decreasing the amount of physical storage required to house an organization’s data. Efficiency is achieved by scanning the on-disk data for identical blocks and then eliminating any duplicates. This approach is commonly referred to as post-process, or asynchronous, deduplication.

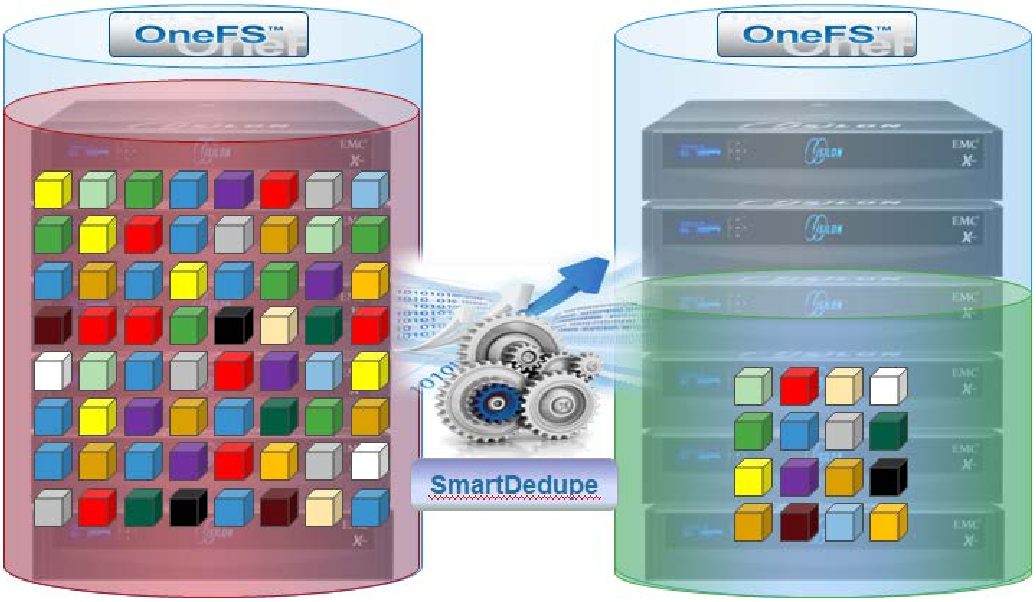


Figure 1. EMC Isilon storage efficiency with SmartDedupe

After duplicate blocks are discovered, Isilon SmartDedupe moves a single copy of those blocks to a special set of files known as shadow stores. During this process, duplicate blocks are removed from the actual files and replaced with pointers to the shadow stores.

With post-process deduplication, new data is first stored on the storage device and then a subsequent process analyzes the data looking for commonality. This means that initial file write or modify performance is not impacted because no additional computation is required in the write path.

EMC Isilon SmartDedupe architecture

The EMC Isilon SmartDedupe architecture comprises five principle modules:

- Deduplication Control Path
- Deduplication Job
- Deduplication Engine
- Shadow Store
- Deduplication Infrastructure

The SmartDedupe control path comprises the EMC Isilon OneFS Web Management Interface (WebUI), command line interface (CLI), and RESTful platform API, and is responsible for managing the configuration, scheduling, and control of the Deduplication Job. The job itself is a highly distributed background process that manages the orchestration of deduplication across all the nodes in the cluster. Job control encompasses file system scanning, detection, and sharing of matching data blocks, in concert with the Deduplication Engine. The Deduplication Infrastructure layer is the kernel module that performs the consolidation of shared data blocks into shadow stores, the file system containers that hold both physical data blocks and references (or pointers) to shared blocks. These elements are described in more detail in the text that follows.

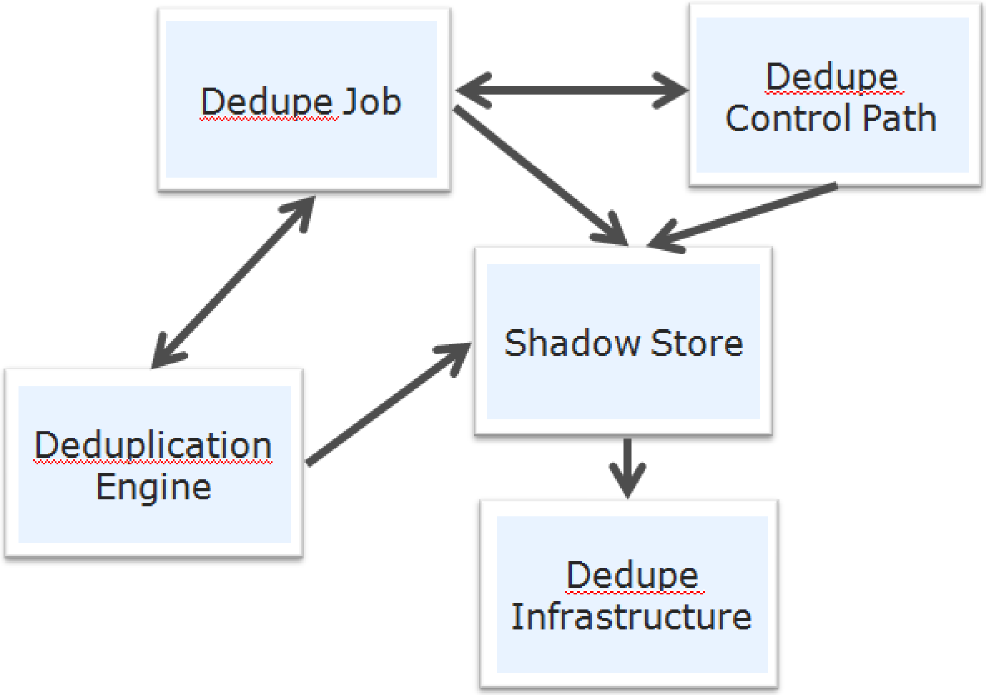


Figure 2. The Isilon SmartDedupe modular architecture

Deduplication Engine—Sampling, fingerprinting, and matching

One of the most fundamental components of OneFS SmartDedupe, and deduplication in general, is “fingerprinting.” In this part of the deduplication process, unique digital signatures, or fingerprints, are calculated using the SHA-1 hashing algorithm, one for each 8 KB data block in the sampled set.

When SmartDedupe runs for the first time, it scans the dataset and selectively samples blocks from it, creating the fingerprint index. This index contains a sorted list of the digital fingerprints, or hashes, and their associated blocks. After the index is created, the fingerprints are checked for duplicates. When a match is found, during the sharing phase, a byte-by-byte comparison of the blocks is performed to verify that they are absolutely identical and to ensure that there are no hash collisions. Then, if they are determined to be identical, the block’s pointer is updated to the already existing data block, and the new duplicate data block is released.

Hash computation and comparison is utilized only during the sampling phase. The deduplication job phases are covered in detail below. For the block sharing phase, full data comparison is employed. SmartDedupe also operates on the premise of variable length deduplication, where the block matching window is increased to encompass larger runs of contiguous matching blocks.

Shadow stores

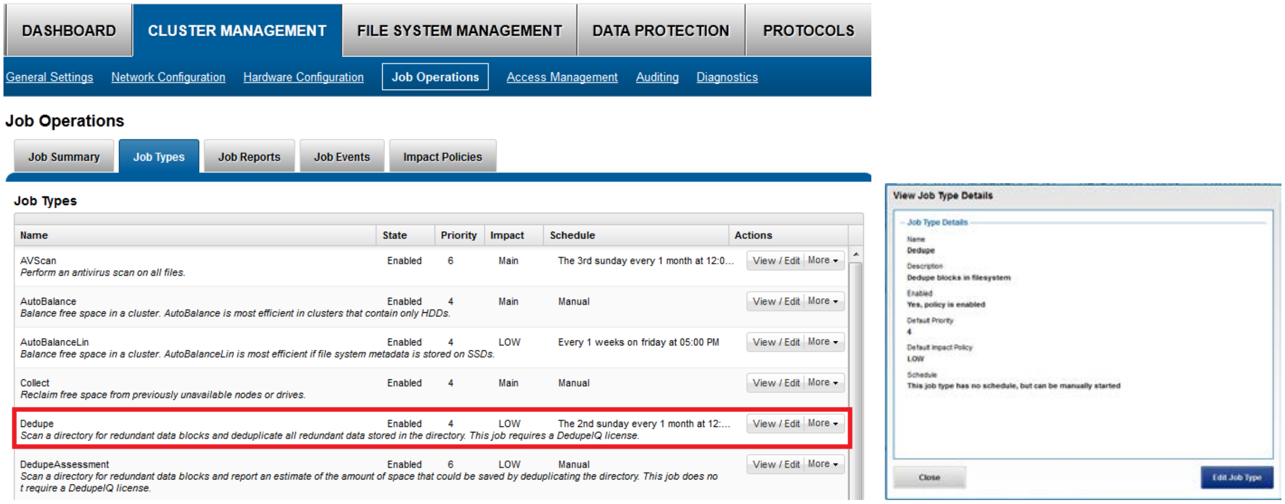
OneFS shadow stores are file system containers that allow data to be stored in a sharable manner. As a result, files on OneFS can contain both physical data and pointers, or references, to shared blocks in shadow stores. Shadow stores were introduced in OneFS 7.0, initially supporting Isilon OneFS file clones. And there are many overlaps between cloning and deduplicating files.

Shadow stores are similar to regular files, but typically don't contain all the metadata typically associated with regular file inodes. In particular, time-based attributes (creation time, modification time, and so on) are explicitly not maintained. Each shadow store can contain up to 256 blocks, with each block able to be referenced by 32,000 files. If this 32,000 reference limit is exceeded, a new shadow store is created. Additionally, shadow stores do not reference other shadow stores, and snapshots of shadow stores are not allowed because shadow stores have no hard links.

Deduplication Job and Deduplication Infrastructure

Deduplication is performed in parallel across the cluster by the OneFS Job Engine via a dedicated deduplication job, which distributes worker threads across all nodes. This distributed work allocation model allows SmartDedupe to scale linearly as an Isilon cluster grows and additional nodes are added.

The control, impact management, monitoring and reporting of the deduplication job is performed by the Job Engine in a similar manner to other storage management and maintenance jobs on the cluster.



Name	State	Priority	Impact	Schedule	Actions
AVScan *Perform an antivirus scan on all files.*	Enabled	6	Main	The 3rd sunday every 1 month at 12:0...	View / Edit More ▾
AutoBalance *Balance free space in a cluster. AutoBalance is most efficient in clusters that contain only HDDs.*	Enabled	4	Main	Manual	View / Edit More ▾
AutoBalanceLin *Balance free space in a cluster. AutoBalanceLin is most efficient if file system metadata is stored on SSDs.*	Enabled	4	LOW	Every 1 weeks on friday at 05:00 PM	View / Edit More ▾
Collect *Reclaim free space from previously unavailable nodes or drives.*	Enabled	4	Main	Manual	View / Edit More ▾
Dedupe *Scan a directory for redundant data blocks and deduplicate all redundant data stored in the directory. This job requires a DedupeIQ license.*	Enabled	4	LOW	The 2nd sunday every 1 month at 12:...	View / Edit More ▾
DedupeAssessment *Scan a directory for redundant data blocks and report an estimate of the amount of space that could be saved by deduplicating the directory. This job does no t require a DedupeIQ license.*	Enabled	6	LOW	Manual	View / Edit More ▾

View Job Type Details

Job Type Details

Name
Dedupe

Description
Dedupe blocks in filesystem

Enabled
Yes, policy is enabled

Default Priority
4

Default Impact Policy
LOW

Schedule
This job type has no schedule, but can be manually started

Figure 3. SmartDedupe job control via the OneFS WebUI

While deduplication can run concurrently with other cluster jobs, only a single instance of the deduplication job, albeit with multiple workers, can run at any one time. Although the overall performance impact on a cluster is relatively small, the deduplication job does consume CPU and memory resources.

The primary user-facing component of OneFS SmartDedupe is the deduplication job. This job performs a file system tree walk of the configured directory, or multiple directories, hierarchy.

Note: The deduplication job will automatically ignore (not deduplicate) the reserved cluster configuration information located under the /ifs/.ifsvar/ directory, as well as any file system snapshots.

Architecturally, the duplication job, and the supporting dedupe infrastructure, comprise the following four phases:

- Sampling
- Duplicate Detection
- Block Sharing
- Index Update

These phases are described in more detail below.

Because the SmartDedupe job is typically long running, each of the phases are executed for a set time period, performing as much work as possible before yielding to the next phase. When all four phases have been run, the job returns to the first phase and continues from where it left off. Incremental deduplication job progress tracking is available via the OneFS Job Engine reporting infrastructure.

Sampling phase

In the Sampling phase, SmartDedupe performs a tree walk of the configured dataset in order to collect deduplication candidates for each file. The rationale is that a large percentage of shared blocks can be detected with only a smaller sample of data blocks represented in the index table. By default, the Sampling phase selects one block from every 16 blocks of a file as a deduplication candidate. For each candidate, a key/value pair consisting of the block’s fingerprint (SHA-1 hash) and file system location (logical inode number and byte offset) is inserted into the index. Once a file has been sampled, the file is flagged and won’t be re-scanned until it has been modified. This dramatically improves the performance of subsequent deduplication jobs.

Duplicate Detection phase

During the Duplicate, or commonality Detection phase, the deduplication job scans the index table for fingerprints (or hashes) that match those of the candidate blocks. If the index entries of two files match, a request entry is generated. In order to improve deduplication efficiency, a request entry also contains pre- and post-limit information. This information contains the number of blocks in front of and behind the matching block that the block sharing phase should search for a larger matching data chunk, and typically aligns to a OneFS protection group’s boundaries.

Block Sharing phase

During the Block Sharing phase, the Deduplication Job calls into the shadow store library and Deduplication Infrastructure to perform the block sharing. Multiple request entries are consolidated into a single sharing request, which is processed by the block sharing phase and ultimately results in the deduplication of the common blocks. The file system searches for contiguous matching regions before and after the matching blocks in the sharing request; if any such regions are found, they too will be shared. Blocks are shared by writing the matching data to a common shadow store and creating references from the original files to this shadow store.

Index Update phase

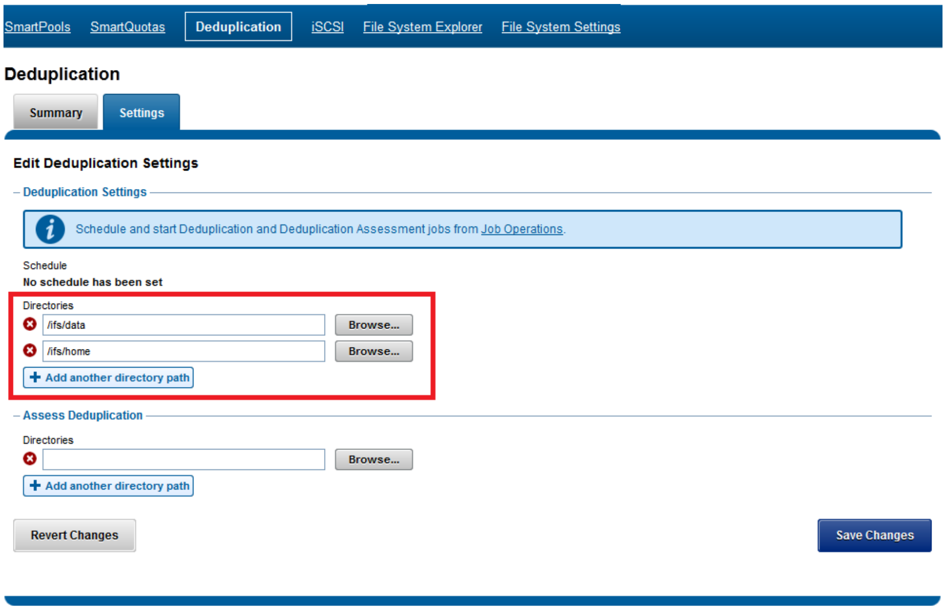
This phase populates the index table with the sampled and matching block information gathered during the previous three phases. After a file has been scanned by the deduplication job, OneFS may not find any matching blocks in other files on the cluster. If, after a number of other files have been scanned, and a file continues to not share any blocks with other files on the cluster, OneFS will remove the index entries for that file. This helps prevent OneFS from wasting cluster resources by searching for unlikely matches. SmartDedupe scans each file in the specified dataset once, after which time the file is marked, preventing subsequent deduplication jobs from rescanning the file until it has been modified.

SmartDedupe management

There are two principal elements to managing deduplication in OneFS. The first is the configuration of the SmartDedupe process itself. The second involves the scheduling and execution of the deduplication job. Both elements are described in the text that follows.

Configuring SmartDedupe

SmartDedupe works on datasets that are configured at the directory level, targeting all files and directories under each specified root directory. Multiple directory paths can be specified as part of the overall deduplication job configuration and scheduling.



Deduplication

Edit Deduplication Settings

Deduplication Settings

Schedule and start Deduplication and Deduplication Assessment jobs from Job Operations.

Schedule
No schedule has been set

Directories
/ifs/data
/ifs/home

Assess Deduplication

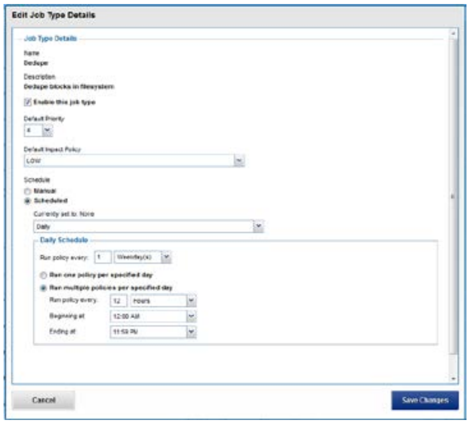
Directories

Figure 4. SmartDedupe configuration via the OneFS WebUI

Note: The permissions required to configure and modify deduplication settings are separate from those needed to run a deduplication job. For example, a user's role must have Job Engine privileges to run a deduplication job. However, in order to configure and modify deduplication configuration settings, the person must have the deduplication role privileges.

Running SmartDedupe

SmartDedupe can be run either on demand (started manually) or via a predefined schedule. This is configured via the cluster management "Job Operations" section of the WebUI.


Edit Job Type Details

Job Type Details

Name
Dedupe

Description
Dedupe blocks in filesystem

☑ Enable this job type

Default Priority
4

Default Impact Policy
LOW

Schedule
○ Manual
● Scheduled

Currently set to: None
Daily

Daily Schedule

Run policy every: 1 Weekday(s)
○ Run one policy per specified day
● Run multiple policies per specified day

Run policy every: 12 Hours
Beginning at: 12:00 AM
Ending at: 11:59 PM

Figure 5. SmartDedupe job configuration and scheduling via the OneFS WebUI

Note: The deduplication job will always run at a low impact level. This value cannot be reconfigured from the default setting.

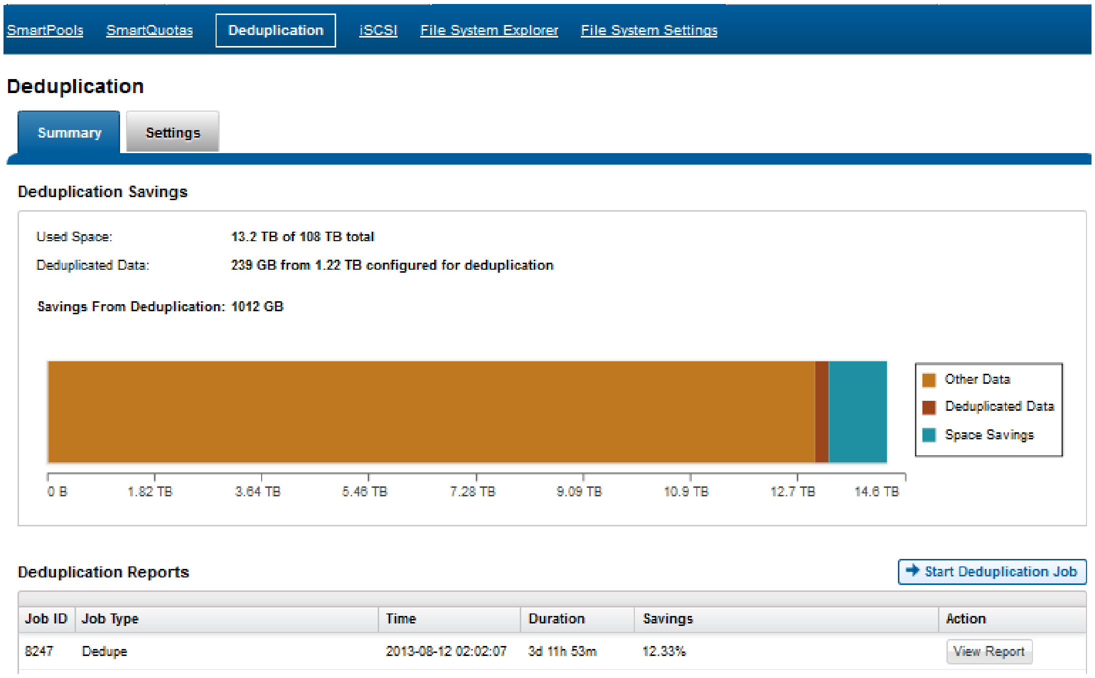
EMC recommends scheduling and running deduplication during off hours, when the rate of data change on the cluster is low. If clients are continually writing to files, the amount of space saved by deduplication will be minimal because the deduplicated blocks are constantly being removed from the shadow store.

For most clusters, after the initial deduplication job has completed, the recommendation is to run an incremental deduplication job once every two weeks.

SmartDedupe monitoring and reporting

Deduplication efficiency reporting

The amount of disk space currently saved by SmartDedupe can be determined by viewing the cluster capacity usage chart and deduplication reports summary table in the WebUI. The cluster capacity chart and deduplication reports can be found by navigating to **File System Management > Deduplication > Summary.**



Deduplication

Deduplication Savings

Used Space: 13.2 TB of 108 TB total

Deduplicated Data: 239 GB from 1.22 TB configured for deduplication

Savings From Deduplication: 1012 GB

Deduplication Reports

Job ID	Job Type	Time	Duration	Savings	Action
8247	Dedupe	2013-08-12 02:02:07	3d 11h 53m	12.33%	View Report

Figure 6. SmartDedupe Cluster Capacity Savings WebUI chart

In addition to the bar chart and accompanying statistics shown in Figure 6, which graphically represents the dataset and space efficiency in actual capacity terms, the deduplication job report overview field also displays the SmartDedupe savings as a percentage.

SmartDedupe space efficiency metrics are also provided via the “isi dedupe stats” CLI command:

```
# isi dedupe stats
      Cluster Physical Size: 676.8841T
          Cluster Used Size: 236.3181T
  Logical Size Deduplicated: 29.2562T
             Logical Saving: 25.5125T
Estimated Size Deduplicated: 42.5774T

  Estimated Physical Saving: 37.1290T
```

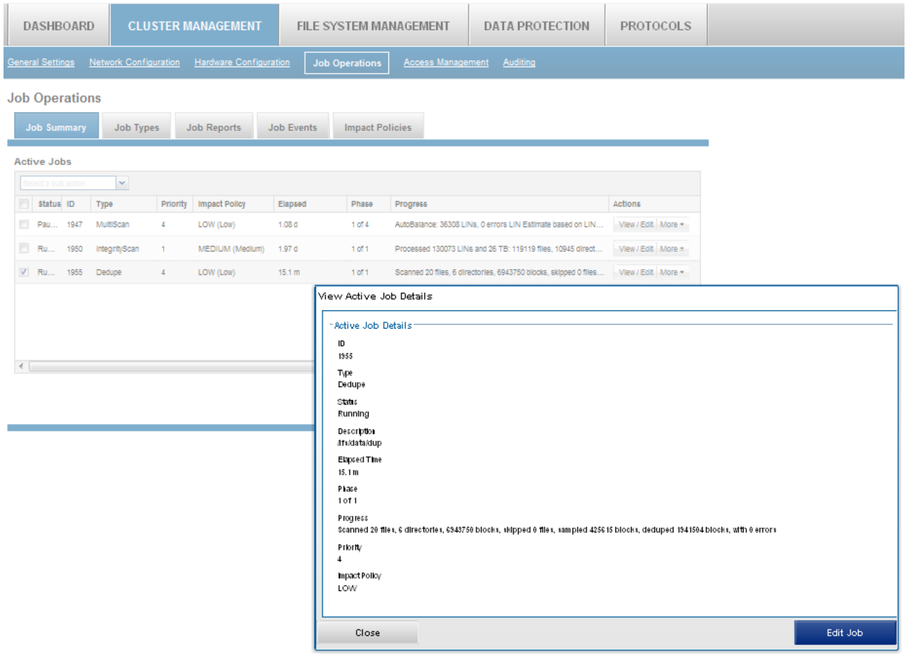
Figure 7. SmartDedupe efficiency statistics via the CLI

SmartDedupe job progress

The Job Engine parallel execution framework provides comprehensive runtime and completion reporting for the deduplication job.

While SmartDedupe is underway, the job status is available at a glance via the Progress column in the Active Jobs table. This information includes the number of files, directories, and blocks that have been scanned, skipped, and sampled, as well as any errors that may have been encountered.

Additional progress information is provided in an Active Job Details status update, which includes an estimated completion percentage based on the number of logical inodes (LINs) that have been counted and processed.



	Status	ID	Type	Priority	Impact Policy	Elapsed	Phase	Progress	Actions
□	Pau...	1947	MultiScan	4	LOW (Low)	1.08 d	1 of 4	AutoBalance: 36308 LINs, 0 errors LIN Estimate based on LIN...	View / Edit More ▾
□	Ru...	1950	IntegrityScan	1	MEDIUM (Medium)	1.97 d	1 of 1	Processed 130073 LINs and 26 TB: 119119 files, 10945 direct...	View / Edit More ▾
☑	Ru...	1955	Dedupe	4	LOW (Low)	15.1 m	1 of 1	Scanned 20 files, 6 directories, 6943750 blocks, skipped 0 files...	View / Edit More ▾

ID
1955

Type
Dedupe

Status
Running

Description
/ifs/data/dup

Elapsed Time
15.1 m

Phase
1 of 1

Progress
Scanned 20 files, 6 directories, 6943750 blocks, skipped 0 files, sampled 425615 blocks, deduped 1941504 blocks, with 0 errors

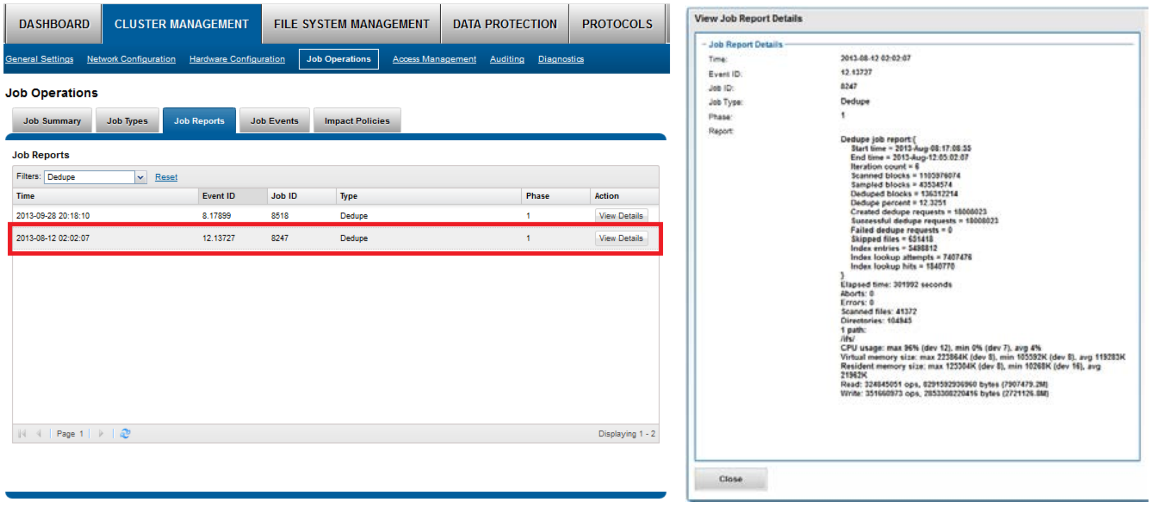
Priority
4

Impact Policy
LOW

Figure 8. Example Active Jobs status update

SmartDedupe job reports

Once the SmartDedupe job has run to completion, or has been terminated, a full deduplication job report is available. This can be accessed from the WebUI by navigating to **Cluster Management > Job Operations > Job Reports**, and selecting the “View Details” action button on the desired deduplication job line item.



Time	Event ID	Job ID	Type	Phase	Action
2013-09-28 20:18:10	8.17899	8518	Dedupe	1	View Details
2013-08-12 02:02:07	12.13727	8247	Dedupe	1	View Details

Time:	2013-08-12 02:02:07
Event ID:	12.13727
Job ID:	8247
Job Type:	Dedupe
Phase:	1
Report:	

```
Dedupe job report:{
    Start time = 2013-Aug-08:17:08:35
    End time = 2013-Aug-12:05:02:07
    Iteration count = 6
    Scanned blocks = 1105976074
    Sampled blocks = 43534574
    Deduped blocks = 136312214
    Dedupe percent = 12.3251
    Created dedupe requests = 18008023
    Successful dedupe requests = 18008023
    Failed dedupe requests = 0
    Skipped files = 631418
    Index entries = 5498812
    Index lookup attempts = 7407476
    Index lookup hits = 1840770
}
Elapsed time: 301992 seconds
Aborts: 0
Errors: 0
Scanned files: 41372
Directories: 104845
1 path:
/ifs/
CPU usage: max 96% (dev 12), min 0% (dev 7), avg 4%
Virtual memory size: max 223864K (dev 8), min 105592K (dev 8), avg 119283K
Resident memory size: max 125304K (dev 8), min 10268K (dev 16), avg
21962K
Read: 324845051 ops, 8291592936960 bytes (7907479.2M)
Write: 351660973 ops, 2853308220416 bytes (2721126.8M)
```

Figure 9. Example WebUI Deduplication Job report

The job report contains the relevant deduplication metrics shown in Table 1.

Report field	Description of metric
Start time	When the deduplication job started.
End time	When the deduplication job finished.
Scanned blocks	Total number of blocks scanned under configured path(s).
Sampled blocks	Number of blocks that OneFS created index entries for.
Created deduplication requests	Total number of deduplication requests created. A deduplication request is created for each matching pair of data blocks. For example, if three data blocks all match, two requests are created: One request to pair file1 and file2 together, and the other request to pair file2 and file3 together.
Successful deduplication requests	Number of deduplication requests that completed successfully.
Failed deduplication requests	Number of deduplication requests that failed. If a dedupe request fails, it does *not* mean that the job also failed. A deduplication request can fail for any number of reasons. For example, the file might have been modified since it was sampled.
Skipped files	Number of files that were not scanned by the deduplication job. The primary reason this occurs is that the file has already been scanned and hasn't been modified since. Another reason for a file to be skipped is that it is less than 32 KB in size. Such files are considered too small and don't provide enough space savings benefit to offset the fragmentation they will cause.
Index entries	Number of entries that currently exist in the index.
Index lookup attempts	Cumulative total number of lookups that have been performed by prior and current deduplication jobs. A lookup occurs when the deduplication job attempts to match a block that has been indexed with a block that hasn't been indexed.
Index lookup hits	Total number of lookup hits that have been performed by earlier deduplication jobs plus the number of lookup hits performed by this deduplication job. A hit is a match of a sampled block with a block in index.

Table 1. Deduplication Job report statistics table

Deduplication job reports are also available from the CLI via the " `isi job reports view <job_id>` " command.

Note: From an execution and reporting stance, the Job Engine considers the "dedupe" job to consist of a single process or phase. The Job Engine events list will report that Dedupe Phase1 has ended and succeeded. This indicates that an entire SmartDedupe job, including all four internal deduplication phases (Sampling, Duplicate Detection, Block Sharing, and Index Update), has been successfully completed. For example:

```
# isi job events list --job-type dedupe

Time                Message

------------------------------------------------------

2013-08-01T13:39:32 Dedupe[1955] Running

2013-08-01T13:39:32 Dedupe[1955] Phase 1: begin dedupe

2013-08-01T14:20:32 Dedupe[1955] Phase 1: end dedupe

2013-08-01T14:20:32 Dedupe[1955] Phase 1: end dedupe

2013-08-01T14:20:32 Dedupe[1955] Succeeded
```

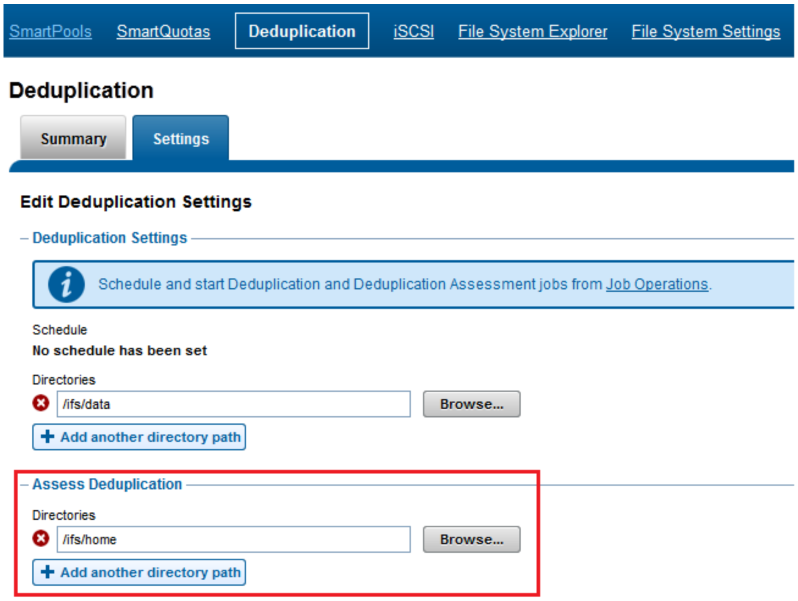
Figure 10. Example command line (CLI) Deduplication Job events list

For deduplication reporting across multiple OneFS clusters, EMC Isilon SmartConnect™ is also integrated with the EMC Isilon InsightIQ® cluster reporting and analysis product. A report detailing the space savings delivered by deduplication is available via the InsightIQ File Systems Analytics module.

Space savings estimation with the SmartDedupe Assessment Job

To complement the actual deduplication job, a dry-run SmartDedupe Assessment Job is also provided to help estimate the amount of space savings that will be realized by running deduplication on a particular directory or set of directories. The SmartDedupe Assessment Job reports the total potential space savings. The dedupe assessment does not differentiate a fresh run from a previous deduplication job that has already performed some sharing on the files in that directory. The SmartDedupe Assessment Job does not provide the incremental differences between instances of this job. Isilon recommends that users run the assessment job once on a specific directory prior to starting an actual deduplication job on that directory.

The assessment job runs similarly to the actual deduplication job, but uses a separate configuration. It also does not require a product license and can be run prior to purchasing SmartDedupe in order to determine whether deduplication is appropriate for a particular dataset or environment.



Deduplication

Edit Deduplication Settings

Deduplication Settings

Schedule and start Deduplication and Deduplication Assessment jobs from Job Operations.

Schedule
No schedule has been set

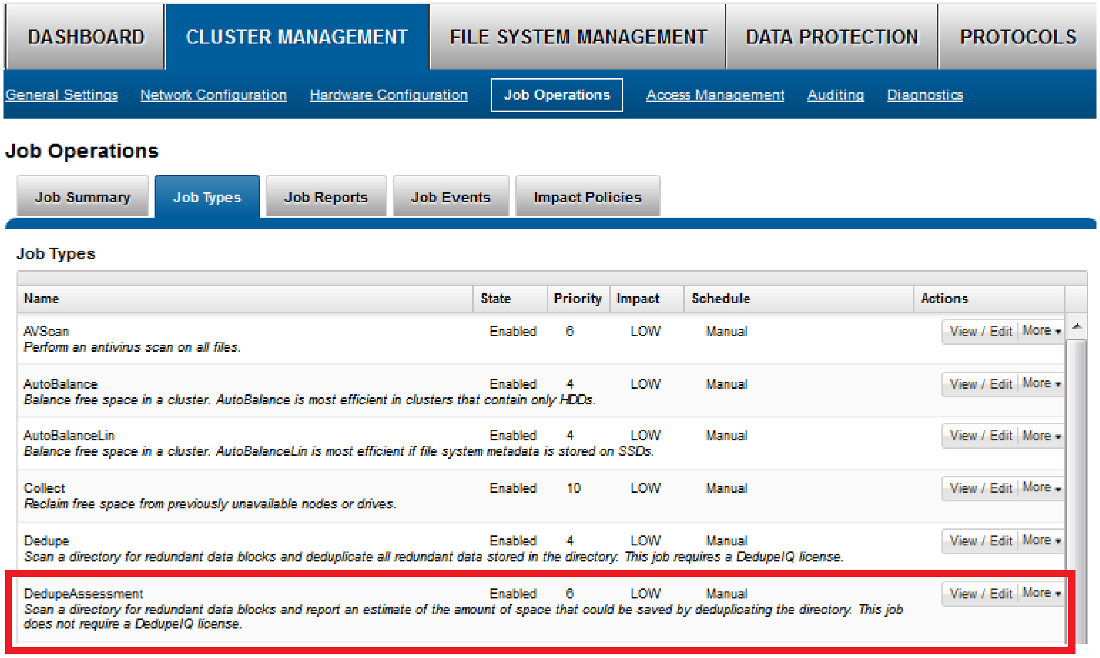
Directories
/ifs/data

Assess Deduplication

Directories
/ifs/home

Figure 11. SmartDedupe Assessment Job configuration

The SmartDedupe Assessment Job uses a separate index table. For efficiency, it also samples fewer candidate blocks than the main deduplication job, and does not actually perform deduplication. Using the sampling and consolidation statistics, the job provides a report that estimates the total deduplication space savings in bytes.



Job Operations

Job Types

Name	State	Priority	Impact	Schedule	Actions
AVScan *Perform an antivirus scan on all files.*	Enabled	6	LOW	Manual	View / Edit More ▾
AutoBalance *Balance free space in a cluster. AutoBalance is most efficient in clusters that contain only HDDs.*	Enabled	4	LOW	Manual	View / Edit More ▾
AutoBalanceLin *Balance free space in a cluster. AutoBalanceLin is most efficient if file system metadata is stored on SSDs.*	Enabled	4	LOW	Manual	View / Edit More ▾
Collect *Reclaim free space from previously unavailable nodes or drives.*	Enabled	10	LOW	Manual	View / Edit More ▾
Dedupe *Scan a directory for redundant data blocks and deduplicate all redundant data stored in the directory. This job requires a DedupeIQ license.*	Enabled	4	LOW	Manual	View / Edit More ▾
DedupeAssessment *Scan a directory for redundant data blocks and report an estimate of the amount of space that could be saved by deduplicating the directory. This job does not require a DedupeIQ license.*	Enabled	6	LOW	Manual	View / Edit More ▾

Figure 12. SmartDedupe Assessment Job control via the OneFS WebUI

Performance with SmartDedupe

As with most things in life, deduplication is a compromise. In order to gain increased levels of storage efficiency, additional cluster resources (CPU, memory, and disk I/O) are utilized to find and execute the sharing of common data blocks.

Another important performance impact consideration with deduplication is the potential for data fragmentation. After deduplication, files that previously enjoyed contiguous on-disk layout will often have chunks spread across less optimal file system regions. This can lead to slightly increased latencies when accessing these files directly from disk, rather than from cache. To help reduce this risk, SmartDedupe will not share blocks across node pools or data tiers, and it will not attempt to deduplicate files smaller than 32 KB in size. On the other end of the spectrum, the largest contiguous region that will be matched is 4 MB in size.

Because deduplication is a data efficiency product rather than performance-enhancing tool, in most cases the consideration will be around cluster impact management. This is from both the client data access performance front because, by design, multiple files will be sharing common data blocks, and also from the deduplication job execution perspective, as additional cluster resources are consumed to detect and share commonality.

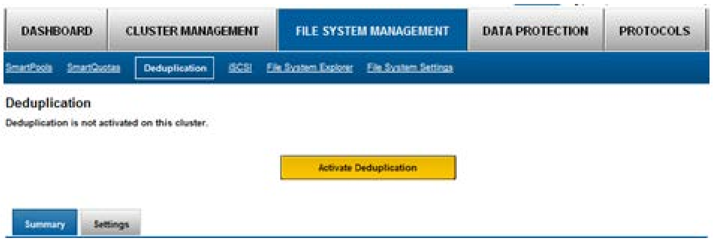
The first deduplication job run will often take a substantial amount of time to run because it must scan all files under the specified directories to generate the initial index and then create the appropriate shadow stores. However, deduplication job performance will typically improve significantly on the second and subsequent job runs (incrementals), once the initial index and the bulk of the shadow stores have already been created.

If incremental deduplication jobs do take a long time to complete, this is most likely indicative of a dataset with a high rate of change. If a deduplication job is paused or interrupted, it will automatically resume the scanning process from where it left off.

As mentioned previously, deduplication is a long-running process that involves multiple job phases that are run iteratively. SmartDedupe typically processes around 1 TB of data per day, per node.

SmartDedupe licensing

SmartDedupe is included as a core component of EMC Isilon OneFS but requires a valid product license key in order to activate it. This license key can be purchased through your EMC Isilon account team. An unlicensed cluster will show the SmartDedupe warning displayed in Figure 13 until a valid product license has been purchased and applied to the cluster.

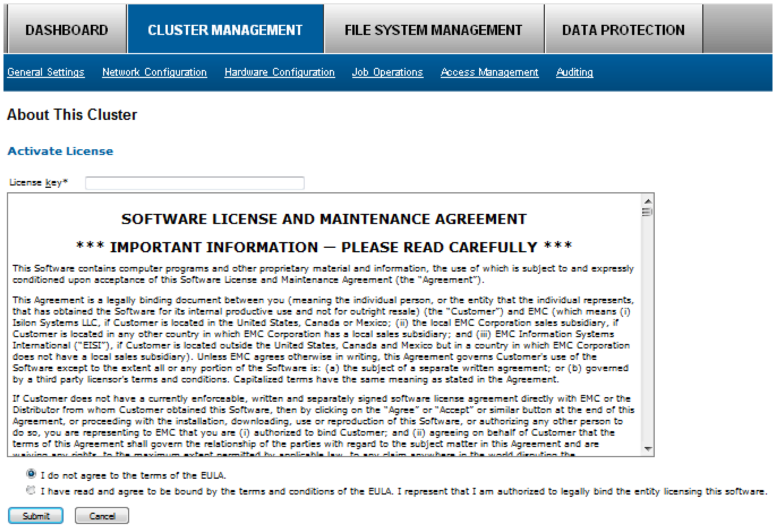


Deduplication

Deduplication is not activated on this cluster.

Figure 13. Unlicensed SmartDedupe warning

License keys can be easily added via the "Activate License" section of the OneFS WebUI, which can be accessed by navigating via **Help > About This Cluster**.



About This Cluster

Activate License

License key*

SOFTWARE LICENSE AND MAINTENANCE AGREEMENT

*** IMPORTANT INFORMATION — PLEASE READ CAREFULLY ***

This Software contains computer programs and other proprietary material and information, the use of which is subject to and expressly conditioned upon acceptance of this Software License and Maintenance Agreement (the "Agreement").

This Agreement is a legally binding document between you (meaning the individual person, or the entity that the individual represents, that has obtained the Software for its internal productive use and not for outright resale) (the "Customer") and EMC (which means (i) Isilon Systems LLC, if Customer is located in the United States, Canada or Mexico; (ii) the local EMC Corporation sales subsidiary, if Customer is located in any other country in which EMC Corporation has a local sales subsidiary; and (iii) EMC Information Systems International ("EISI"), if Customer is located outside the United States, Canada and Mexico but in a country in which EMC Corporation does not have a local sales subsidiary). Unless EMC agrees otherwise in writing, this Agreement governs Customer's use of the Software except to the extent all or any portion of the Software is: (a) the subject of a separate written agreement; or (b) governed by a third party licensor's terms and conditions. Capitalized terms have the same meaning as stated in the Agreement.

If Customer does not have a currently enforceable, written and separately signed software license agreement directly with EMC or the Distributor from whom Customer obtained this Software, then by clicking on the "Agree" or "Accept" or similar button at the end of this Agreement, or proceeding with the installation, downloading, use or reproduction of this Software, or authorizing any other person to do so, you are representing to EMC that you are (i) authorized to bind Customer; and (ii) agreeing on behalf of Customer that the terms of this Agreement shall govern the relationship of the parties with regard to the subject matter in this Agreement and are waiving any rights, to the maximum extent permitted by applicable law, to any claim anywhere in the world disputing the

● I do not agree to the terms of the EULA.

○ I have read and agree to be bound by the terms and conditions of the EULA. I represent that I am authorized to legally bind the entity licensing this software.

Figure 14. OneFS WebUI license activation page

Note: The SmartDedupe dry-run estimation job can be run without any licensing requirements, allowing users to assess the potential space savings that a dataset might yield before making the decision to purchase the full product.

Deduplication efficiency

Deduplication can significantly increase the storage efficiency of data. However, the actual space savings will vary depending on the specific attributes of the data itself. As mentioned earlier, the deduplication assessment job can be run to help predict the likely space savings that deduplication will provide on a given dataset.

Virtual machines files often contain duplicate data, much of which is rarely modified. Deduplicating similar OS type virtual machine images (for example, VMware VMDK files, and so on) that have been block aligned can significantly decrease the amount of storage space consumed. However, as noted previously, the potential for performance degradation as a result of block sharing and fragmentation should be carefully considered first.

SmartDedupe can also deduplicate iSCSI LUNs because deduplication does not treat a LUN (a directory containing a group of extent files) differently from regular files.

Isilon SmartDedupe does not deduplicate across files that have different protection settings. For example, if two files share blocks, but file1 is parity protected at +2:1, and file2 has its protection set at +3, SmartDedupe will not attempt to deduplicate them. This ensures that all files and their constituent blocks are protected as configured. Additionally, SmartDedupe won't deduplicate files that are stored on different EMC Isilon SmartPools® storage tiers or node pools. For example, if file1 and file2 are stored on tier 1 and tier 2, respectively, and tier1 and tier2 are both protected at 2:1, OneFS won't deduplicate them. This helps guard against performance asynchronicity, where some of a file's blocks could live on a different tier, or class of storage, than the others.

Following are some examples of typical space reclamation levels that have been achieved with SmartDedupe.

Note: These deduplication space saving values are provided solely as general guidance. Because no two datasets are alike (unless they're replicated), actual results can vary considerably from these examples.

Workflow/data type	Typical space savings
Virtual machine data	35%
Home directories/file shares	25%
Email archive	20%
Engineering source code	15%
Media files	10%

Table 2. Typical workload space savings with SmartDedupe

SmartDedupe best practices

For optimal cluster performance, Isilon recommends that users observe the following SmartDedupe best practices (some of which may be covered elsewhere in this paper):

- Deduplication is most effective when applied to datasets with a low rate of change—for example, archived data.
- Enable SmartDedupe to run at subdirectory level(s) below /ifs.
- Avoid adding more than 10 subdirectory paths to the SmartDedupe configuration policy.
- SmartDedupe is ideal for home directories, departmental file shares, and warm and cold archive datasets.

- Run SmartDedupe against a smaller sample dataset first to evaluate performance impact versus space efficiency.
- Schedule deduplication to run during the cluster’s low usage hours—that is, overnight, weekends, and so on.
- After the initial deduplication job has been completed, schedule incremental deduplication jobs to run every two weeks or so, depending on the size and rate of change of the dataset.
- Run the deduplication assessment job on a single root directory at a time. If multiple directory paths are assessed in the same job, you will not be able to determine which directory should be deduplicated.
- When replicating deduplicated data, to avoid running out of space on a target cluster, it is important to verify that the logical data size (that is, the amount of storage space saved plus the actual storage space consumed) does not exceed the total available space on the target cluster.
- Run a deduplication job on an appropriate dataset prior to enabling a snapshot schedule.
- Where possible, perform any snapshot restores (reverts) before running a deduplication job. Also, run a deduplication job directly after restoring a prior snapshot version.

SmartDedupe considerations

As discussed earlier, deduplication isn’t free. There’s always trade-off between cluster resource consumption (CPU, memory, disk), the potential for data fragmentation, and the benefit of increased space efficiency.

- Because deduplication trades cluster performance for storage capacity savings, SmartDedupe is not ideally suited for heavily trafficked data or high-performance workloads.
- Depending on an application’s I/O profile and the effect of deduplication on the data layout, read/write performance and overall space savings can vary considerably.
- SmartDedupe will not permit block sharing across different hardware types or node pools to reduce the risk of performance asymmetry.
- SmartDedupe will not share blocks across files with different protection policies applied.
- OneFS metadata, including the deduplication index, is not deduplicated.
- OneFS will not deduplicate redundant information within a file; it only does so across different files.
- Deduplication is a long-running process that involves multiple job phases that are run iteratively. The deduplication job typically processes around 1 TB of data per day, per node.
- SmartDedupe will not attempt to deduplicate files smaller than 32 KB in size.
- Deduplication job performance will typically improve significantly on the second and subsequent job runs, once the initial index and the bulk of the shadow stores have already been created.

- SmartDedupe will not deduplicate the data stored in a snapshot. However, snapshots can certainly be created of deduplicated data.
- If deduplication is enabled on a cluster that already has a significant amount of data stored in snapshots, it will take time before the snapshot data is affected by deduplication. Newly created snapshots will contain deduplicated data, but older snapshots will not.
- SmartDedupe will automatically run with a “low-impact” Job Engine policy. This cannot be manually reconfigured.

SmartDedupe and OneFS feature integration

EMC Isilon SyncIQ Replication and SmartDedupe

When deduplicated files are replicated to another Isilon cluster via EMC Isilon SyncIQ®, or backed up to a tape device, the deduplicated files are inflated (or rehydrated) back to their original size because they no longer share blocks on the target Isilon cluster. However, once replicated data has landed, SmartDedupe can be run on the target cluster to provide the same space-efficiency benefits as on the source.

Shadows stores are not transferred to target clusters or backup devices. Because of this, deduplicated files do not consume less space than non-deduplicated files when they are replicated or backed up. To avoid running out of space on target clusters or tape devices, it is important to verify that the total amount of storage space saved and storage space consumed does not exceed the available space on the target cluster or tape device. To reduce the amount of storage space consumed on a target Isilon cluster, you can configure deduplication for the target directories of your replication policies. Although this will deduplicate data on the target directory, it will not allow SyncIQ to transfer shadow stores. Deduplication is still performed post-replication via a deduplication job running on the target cluster.

Backup and SmartDedupe

Because files are backed up as if the files were not deduplicated, backup and replication operations are not faster for deduplicated data. You can deduplicate data while the data is being replicated or backed up.

Note: OneFS Network Data Management Protocol (NDMP) backup data won’t be deduplicated unless deduplication is provided by the backup vendor’s direct memory access (DMA) software. However, compression is often provided natively by the backup tape or virtual tape library (VTL) device.

Snapshots and SmartDedupe

SmartDedupe will not deduplicate the data stored in a snapshot. However, snapshots can be created of deduplicated data. If a snapshot is taken of a deduplicated directory, and then the content of that directory is modified, the shadow stores will be transferred to the snapshot over time. Because of this, more space will be saved on a cluster if deduplication is run prior to enabling snapshots.

If deduplication is enabled on a cluster that already has a significant amount of data stored in snapshots, it will take time before the snapshot data is affected by deduplication. Newly created snapshots will contain deduplicated data, but older snapshots will not.

It is also good practice to revert a snapshot before running a deduplication job. Restoring a snapshot will cause many of the files on the cluster to be overwritten. Any deduplicated files are reverted back to normal files if they are overwritten by a snapshot revert. However, once the snapshot revert is completed, deduplication can be run on the directory again and the resulting space savings will persist on the cluster.

EMC Isilon SmartLock and SmartDedupe

SmartDedupe is also fully compatible with EMC Isilon SmartLock®, Isilon's data retention and compliance product. SmartDedupe delivers storage efficiency for immutable archives and write once, read many (or WORM) protected datasets.

EMC Isilon SmartQuotas and SmartDedupe

EMC Isilon SmartQuotas™ accounts for deduplicated files as if they consumed both shared and unshared data. From the quota side, deduplicated files appear no differently than regular files to standard quota policies. However, if the quota is configured to include data protection overhead, the additional space used by the shadow store will not be accounted for by the quota.

EMC Isilon SmartPools and SmartDedupe

SmartDedupe will not deduplicate files that span SmartPools node pools or tiers, or that have different protection levels set. This is to avoid potential performance or protection asymmetry, which could occur if portions of a file live on different classes of storage.

EMC Isilon InsightIQ and SmartDedupe

InsightIQ, Isilon's multicluster reporting and trending analytics suite, is integrated with SmartDedupe. Included in the data provided by the File Systems Analytics module is a report detailing the space savings efficiency delivered by deduplication.

SmartDedupe use cases

As mentioned earlier, an enterprise's data typically contains substantial quantities of redundant information. And home directories, file shares, and data archives are great example of workloads that consistently yield solid deduplication results. Each time a spreadsheet, document, or email attachment is saved by multiple employees, the same file is stored in full multiple times, taking up valuable disk capacity. OneFS SmartDedupe is typically used in the ways described in the examples that follow.

Example A: File shares and home directory deduplication

By architecting and configuring home directory and file share repositories under unifying top-level directories (for example, /ifs/home and /ifs/data, respectively), an organization can easily and efficiently configure and run deduplication against these datasets.

In terms of performance, home directories and file shares are typically midtier workloads, usually involving concurrent access with a reasonable balance of read/write and data and metadata operations. As a result, they make great candidates for SmartDedupe.

SmartDedupe should ideally be run during periods of low cluster load and client activity (nights and weekends, for example). Once the initial job has completed, the deduplication job can be scheduled to run every two weeks or so, depending on the data's rate of change.

Example B: Storage-efficient archiving

SmartDedupe is an ideal solution for large, infrequently accessed content repositories. Examples of these include digital asset management workloads, seismic data archives for energy exploration, document management repositories for legal discovery, compliance archives for financial or medical records, and so on.

These are all excellent use cases for deduplication because the performance requirements are typically low and biased toward metadata operations, and there are typically numerous duplications of data. As a result, trading system resources for data efficiency produces significant, tangible benefits to the bottom line. SmartDedupe is also ideal for OneFS SmartLock-protected immutable archives and other WORM datasets, and typically delivers attractive levels of storage efficiency.

For optimal results, where possible, ensure that archive data is configured with the same level of protection. For data archives that are frequently scanned or indexed, metadata read acceleration on solid-state drives (SSDs).

Example C: Disaster recovery target cluster deduplication

For performance-oriented environments that would prefer not to run deduplication against their primary dataset, the typical approach is to deduplicate the read-only data replica on their target, or disaster recovery (DR), cluster.

Once the initial deduplication job has successfully completed, subsequent incremental deduplication jobs can be scheduled to run soon after the completion of each SyncIQ replication job, or as best fits the rate of data change and frequency of cluster replication.

SmartDedupe and OneFS storage utilization

SmartDedupe is one of several components of OneFS that enables EMC Isilon to deliver a very high level of raw disk utilization. Another major storage efficiency attribute is the way that OneFS natively manages data protection in file systems. Unlike most file systems that rely on hardware RAID, OneFS protects data at the file level and, using software-based erasure coding, allows most customers to enjoy raw disk space utilization levels in the 80 percent range or higher. This is in contrast to the industry mean of around 50 to 60 percent raw disk capacity utilization. SmartDedupe serves to further extend this storage efficiency headroom, bringing an even more compelling and demonstrable TCO advantage to primary file-based storage.

Conclusion

Up until now, traditional deduplication implementations have typically been expensive, limited in scale, confined to secondary storage, and administratively complex.

The integration of EMC Isilon SmartDedupe with the industry's leading scale-out network-attached storage (NAS) architecture delivers on the promise of simple data efficiency at scale by providing significant storage cost savings without sacrificing ease of use or data protection.

With its simple, powerful interface and intelligent default settings, OneFS SmartDedupe is easy to estimate, configure, and manage, and provides enterprise data efficiency within a single, highly extensible storage pool. Scalability to petabytes and the ability to add new capacity and new technologies while retaining older capacity in the same system means strong investment protection. Integration with OneFS core functions eliminates data risks and gives the user control over what system resources are allocated to data movement.

To learn more about SmartDedupe and other EMC Isilon products, please see www.emc.com/domains/isilon/index.htm.

About EMC

EMC Corporation is a global leader in enabling businesses and service providers to transform their operations and deliver IT as a service. Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping IT departments to store, manage, protect and analyze their most valuable asset—information—in a more agile, trusted and cost-efficient way. Additional information about EMC can be found at www.EMC.com.