

Version 1.0
October 2019

Beyond NAS

*SmartFiles: A Transformative Approach to
Unstructured Data*



ABSTRACT

This white paper looks at the world of files and objects where repositories are rapidly growing and the traditional scale-out NAS vendors are unable to keep pace with their proprietary hardware. The paper examines Cohesity's SmartFiles solution to answer this growth by providing a API-driven, software-defined architecture that provides global space efficiency along with integrated cybersecurity.

Table of Contents

Scale Up vs Scale Out3

Cohesity SmartFiles.....4

 Software-Defined Agility..... 4

Multiprotocol Support from the Start6

Global Space Efficiency7

 Deduplication — Fixed-length vs Cohesity’s Variable-length..... 7

Conventional Fixed-length Deduplication7

Cohesity Variable-length Deduplication.....7

 Compression 8

 Small File Storage Efficiency 8

 Limitless Scale 9

Cybersecurity Data Services and Applications..... 10

About the Authors 11

Document Version History..... 11

Your Feedback 11

Figures

Figure 1: SmartFiles Connects NAS to Cohesity DataPlatform’s Benefits4

Figure 2: Conventional Fixed-length Deduplication7

Figure 3: Cohesity Variable-length Deduplication8

Scale Up vs Scale Out

The International Data Corporation (IDC) predicts that 80 percent of worldwide data will be unstructured by 2025. For many companies, the unfettered growth of data has reached that critical mass already. Unstructured data creates a unique challenge for organizations wishing to use their information for analysis. The data can't easily be stored in a database, and has attributes that make it a challenge to search for, edit, and analyze.

If your organization is struggling to manage their unstructured data now, the problem is going to get exponentially harder over time. Unstructured data comes in the form of file shares, backups, archives, logs, media files, test/dev, and analytics. Traditional network-attached, or scale-up, storage was the most common form of file storage platform. Scale-Up architecture consists of a pair of controllers and several shelves of drives. When you run out of capacity, you have to add another shelf of drives.

With the advancement in applications, the traditional NAS platforms are not able to keep up with the needs of the ever-growing data. Over time, scale-out systems emerged to address the limitation of scale-up architectures. Scale-out storage provides the flexibility to respond to changing business needs and optimize resources. Scale-out storage systems use a group of nodes to form a clustered storage system and can provide file shares over a network — just like a traditional array. When additional capacity and/or performance is needed, additional nodes can be added non-disruptively.

Although some of the traditional NAS platforms can scale out in capacity, they are limited by their architecture. Over time, scale-out NAS platforms enhanced their storage services, which made them more storage-centric rather than data-centric.

At Cohesity, we believe that file services customers are ready for something that transcends the limitations of traditional scale-out NAS. Cohesity SpanFS is designed for web-scale architecture where all the components within Cohesity DataPlatform scale linearly. Cohesity SmartFiles is architected to meet your current and future data needs. The future of scale-out NAS must begin with a blank sheet of paper that provides a fresh solution to the growing need for unstructured data, a solution that is data-centric.

Cohesity SmartFiles

Cohesity file services differentiates itself from traditional scale-out NAS with “SmartFiles” as a next-gen product category. As a web-scale platform, this solution offers multiple benefits, including unlimited scale-out and unparalleled storage efficiency with global variable-length block deduplication and compression across the Cohesity cluster.

Figure 1: SmartFiles Connects NAS to Cohesity DataPlatform’s Benefits



Software-Defined Agility

Companies today have a common challenge: Addressing their growing business needs while building an agile, automated, and programmable data center. As a result, an increasing number of IT organizations are starting to adopt a software-defined architecture as a means of transforming their legacy data centers.

While traditional NAS vendors have optimized the performance of their software, that performance is tied to the NAS vendors’ proprietary hardware. As a result, traditional NAS vendors are limited by the hardware that they support.

A software-defined architecture decouples software from the underlying hardware and moves all the intelligence into the software layer. By doing so, users can eliminate their dependency on proprietary, monolithic hardware. In other words, a software-defined approach allows customers to manage their IT infrastructure easily via the software stack, in an automated fashion.

As monolithic solutions, commonly known as the “black boxes,” these legacy network and storage solutions offered little-to-no choice and flexibility to users. As a result, IT organizations were forced to:

- Over-provision.
- Acquire vendor-specific skills to maintain and operate the solution.

- Create infrastructure and team silos to address growth.

All this created bottlenecks and inefficient workflows. IT teams were usually in catchup mode, unable to focus on any true innovation. Fast forward to today's mature technologies, and a software-defined model that delivers numerous benefits, including:

- **API-Driven.** Built to support an API-first approach, Cohesity DataPlatform integrates with leading automation solutions and custom business scripts to increase operational efficiency.
- **Flexibility.** Most software-defined solutions run on standard x86 architectures, which allows customers to select their own preferred hardware supply. IT administrators now have the flexibility to choose where to run their software. In the case of Cohesity, we can provide an integrated software+hardware solution, or our clients can choose to run our software on qualified Cisco UCS and HPE ProLine nodes.
- **Cloud Readiness.** By extension, software-defined data solutions can also run on-premises or in the cloud. As most enterprises will adopt and use both forms of infrastructure, technology solutions that can flexibly operate across both their capitalized data centers, or cloud service providers, give them the choice to optimize between lowest costs and the most agile infrastructures.
- **Simplicity.** By eliminating silos, and consolidating workloads on a single platform, which is especially true for a software-defined storage like Cohesity, customers can simplify their overall operations. What's more, Cohesity DataPlatform consolidates all your data management tasks in a single, browser-based dashboard.
- **Scalability.** Built for mega-scale data infrastructures, Cohesity DataPlatform is designed to help customers start small with a pay-as-you-grow model. This eliminates the need for overprovisioning and creating unnecessary silos down the road.
- **Lower cost.** This happens in two ways:
 - **CapEx.** Because most software-defined solutions run on x86 hardware, customers don't need to overspend on proprietary hardware, which is usually two to three times more expensive.
 - **OpEx.** A standardized and simplified infrastructure streamlines workflow, which increases operational efficiency, which helps reduce OpEx. And eliminating expensive renewals generates additional OpEx savings.

Multiprotocol Support from the Start

As organizations start adopting hybrid cloud environments, more and more applications are migrating from the file (NFS or SMB/CIFS) protocol to the open-source S3 protocol that allows application owners to easily migrate an application from on-premises to the cloud. However, the traditional NAS vendors only support file protocols, NFS and SMB. Customers are forced to adopt an additional Object storage platform that supports the S3 protocol.

Many traditional vendors took a shortcut with a gateway approach in front of the file storage platform, or they used open-source NFS or SMB libraries in front of the Object platform. Both of these approaches have limitations in terms of performance and scalability.

Unlike some NAS solutions that are implemented using Samba or NFS-Ganesha, Cohesity DataPlatform built support for SMB, NFS, and S3 protocols from the ground up.

Cohesity DataPlatform provides the ability to create file shares, called “Views,” that can be accessed via NFS or SMB/CIFS protocols with Unified Permissions. These Views are members of Cohesity DataPlatform “Storage Domains,” which are logical data pools with defined storage policies for efficiency (deduplication & compression), replication factor and/or erasure coding, encryption, and cloud tiering. Multiprotocol access to the same data allows support of applications across all major enterprise operating systems, including Microsoft Windows, Linux, and the S3 API.

Global Space Efficiency

Given Gartner’s prediction of 800% growth in enterprise data within the next five years, storage space efficiency capabilities, such as deduplication, are critical to any enterprise storage platform.

To deal with the expected data growth, enterprises need storage products capable of delivering the highest level of space efficiency — at optimal cost. Deduplication is one of the key storage technologies enterprises rely on to deliver optimal storage efficiency and reduced infrastructure costs. In addition, Cohesity DataPlatform has built-in compression and small file efficiency.

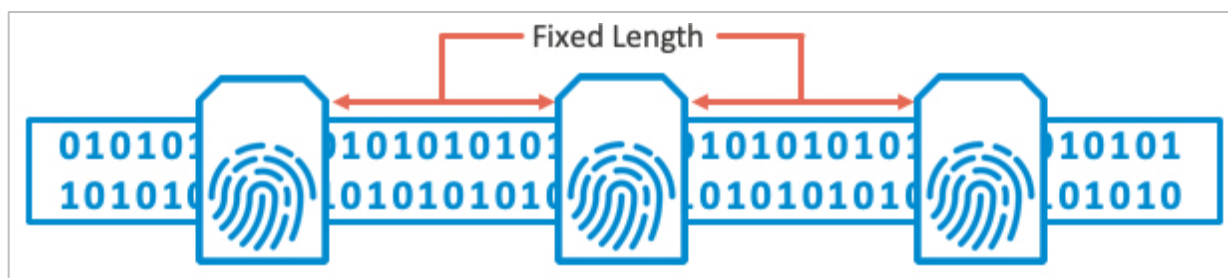
Deduplication — Fixed-length vs Cohesity’s Variable-length

Cohesity DataPlatform uses a unique, variable-length data deduplication technology that spans an entire cluster, resulting in significant savings across the entire storage footprint. With variable-length deduplication, the size is not fixed. Instead, the algorithm divides the data into chunks of varying sizes based on the data characteristics.

Conventional Fixed-length Deduplication

- Markers are placed at fixed-length intervals.
- Less efficient but easier to implement.
- Small file changes result in loss of efficiency.
- Best for file systems with small fixed block sizes.

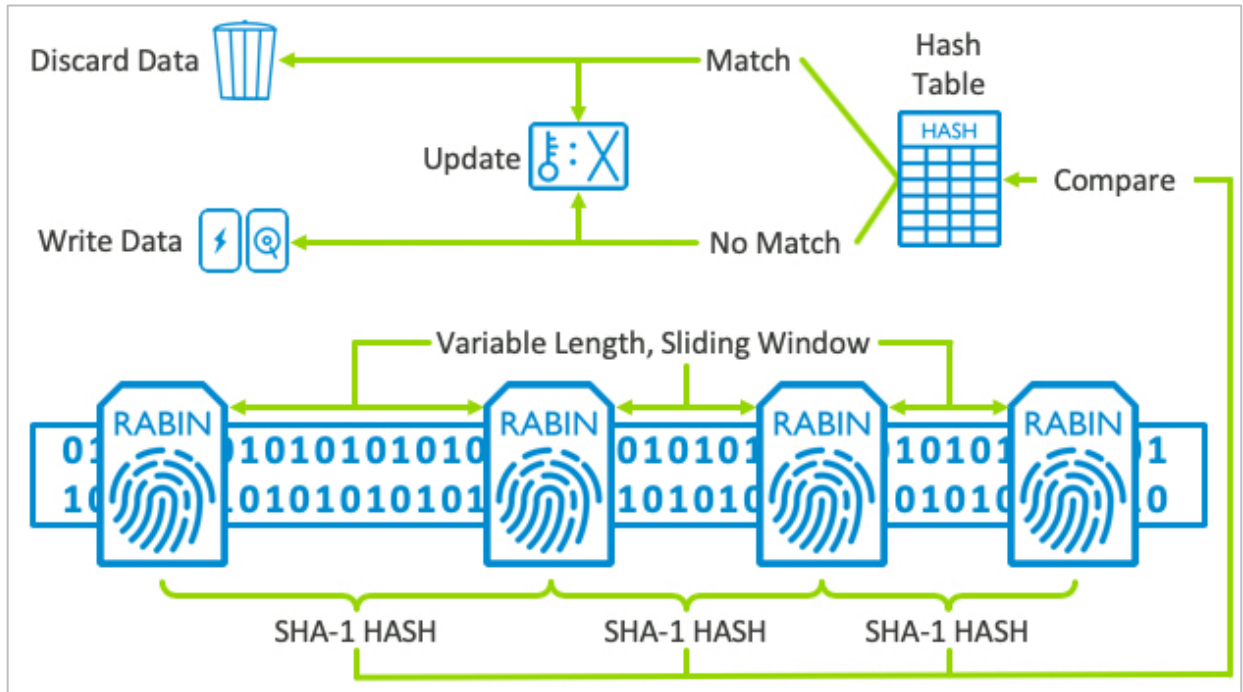
Figure 2: Conventional Fixed-length Deduplication



Cohesity Variable-length Deduplication

- Type of compression that identifies redundant segments of data and replaces duplicate segments with a pointer.
- Cohesity uses variable-length, sliding-window inline and post-process deduplication.

Figure 3: Cohesity Variable-length Deduplication



Compression

Cohesity DataPlatform uses the Zstandard (zstd) compression algorithm. Zstandard offers high compression along with high performance. Compression shrinks consumed storage capacity by reducing the size of bit or byte strings in a data stream. Cohesity DataPlatform is designed to ensure that deduplication always happens before the data is compressed.

You can also toggle inline compression from the Cohesity DataPlatform interface.

- **On.** Compress inline for HDD and flash.
- **Off.** Compress inline for HDD only – meaning data in the flash tier will be compressed only after being down-tiered to HDD.

Small File Storage Efficiency

Cohesity DataPlatform provides small file efficiency. Writes less than 4 kilobytes go to the distributed flash subsystem by default. Writes less than 8 megabytes are stored in small, coalesced, disk-based objects, while files larger than 8 megabytes are stored in larger, coalesced, disk-based objects. Regardless of the file size, all files are eligible for deduplication and compression.

Limitless Scale

Cohesity DataPlatform scales limitlessly, seamlessly, and cost effectively. Cohesity performance scales with capacity. The web-scale architecture allows for unlimited scalability, while our SnapTree technology provides unlimited, distributed snapshots and clones with no impact on performance.

Cybersecurity Data Services and Applications

Critical file-associated apps (such as anti-virus, file audit, and content search) in the file ecosystem still require separate infrastructure and management, separate from NAS appliances. As a result, customers have additional infrastructure costs and management burdens. In a perfect world, the entire file services ecosystem would reside on a single data-centric file services platform and automatically scale out as the file environment expands.

Cohesity DataPlatform's file system provides full at-rest encryption based on the strong 256-bit Advanced Encryption Standard (AES-256). The Cohesity DataPlatform architecture provides this high data security while maintaining the flexibility to leverage the available hardware and software resources. Cohesity DataPlatform also supports multi-factor authentication (MFA).

Cohesity DataPlatform empowers organizations to run apps directly on the platform. By doing so, Cohesity brings compute power to the data, which allows businesses to derive insights from their data. The Cohesity Marketplace offers a growing list of Cohesity and third-party applications.

Currently, the Cohesity Marketplace includes:

- **Cohesity Spotlight.** Analyze file audit logs and find anomalous file-access patterns.
- **Cohesity Runbook.** Automate Data Management.
- **Cohesity Cyberscan.** Discover, assess, and respond to cyber exposure within production environments.
- **ClamAV.** Antivirus scans for your file data.
- **Splunk.** Collect, index, search, analyze, and visualize all your data in one place.

About the Authors

Scott Owens is a Technical Marketing Engineer at Cohesity. In his role, Scott focuses on file services.

Other essential contributors included:

- Adaikkappan Arumugam, Sr Manager, Tech Marketing, Solutions Engineering & Tech Pubs
- Vibhor Gupta, Product Management File Services

Document Version History

VERSION	DATE	DOCUMENT HISTORY
1.0	Oct 2019	First full release

Your Feedback

Was this document helpful? [Send us your feedback!](#)

ABOUT COHESITY

Cohesity makes your data work for you by consolidating secondary storage silos onto a hyperconverged, web-scale data platform that spans both private and public clouds. Enterprise customers begin by radically streamlining their backup and data protection, then converge file and object services, test/dev instances, and analytic functions to provide a global data store. Cohesity counts many Global 1000 companies and federal agencies among its rapidly growing customer base and was named to Forbes' "Next Billion-Dollar Startups 2017," LinkedIn's "Startups: The 50 Industry Disruptors You Need to Know Now," and CRN's "2017 Emerging Vendors in Storage" lists.

For more information, visit our [website](#) and [blog](#), follow us on [Twitter](#) and [LinkedIn](#) and like us on [Facebook](#).

© 2019. Cohesity, Inc.

Cohesity, the Cohesity logo, SnapFS, SnapTree, SpanFS, and SpanOS, are registered trademarks, and DataPlatform, DataProtect, and Helios are trademarks of Cohesity, Inc. All rights reserved.

2000024-001-EN