

Technical Review

Cohesity Data Management

Date: October, 2019 Author: Tony Palmer, Senior Validation Analyst

Abstract

This ESG Technical Review examines the Cohesity data management platform, focusing on Cohesity’s solution to mass data fragmentation (MDF): the ability to collapse workload silos, eliminate copies, and share data efficiently.

The Challenges

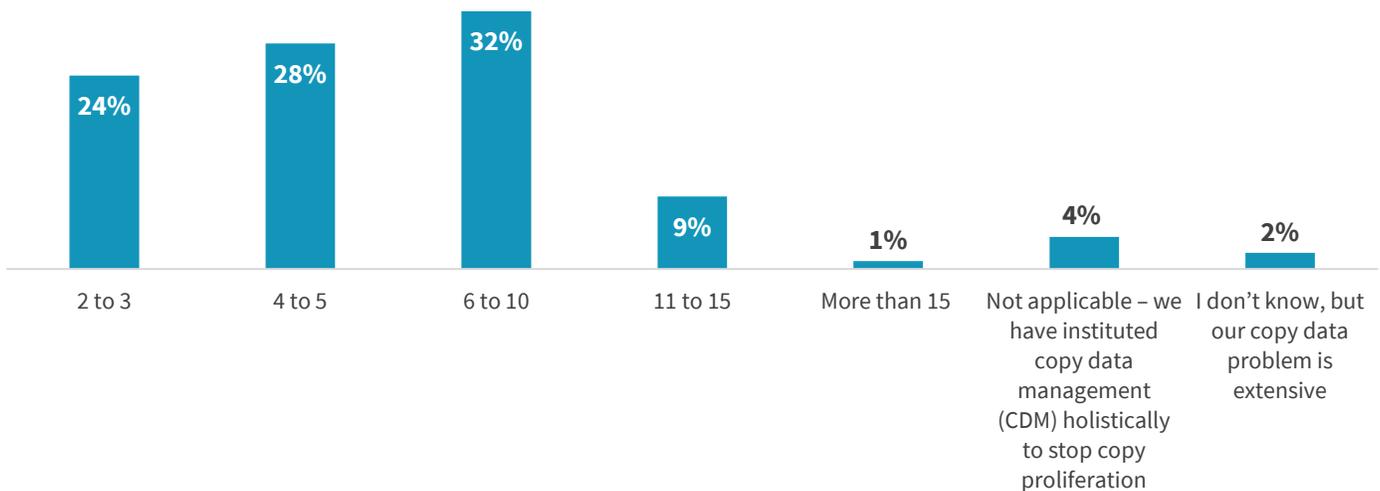
Organizations have made tremendous improvement in their ability to mine, organize, and create new value from their data. This ability enables companies to leverage their data to transform both IT and their business for the digital age and deliver a genuine competitive edge. Digital transformation leads to better productivity, better insight, and greater profit. Yet, as data grows and proliferates among and between different application silos, storage silos, geographic silos, operational silos, and various clouds, organizations’ ability to see, access, manage, and harness the power of that data is weakened.

ESG recently conducted research to better understand and quantify the extent to which data is siloed in organizations today and to what degree respondents believe that creates problems.¹ When ESG asked respondents the number of times a typical data set is stored and copied, two key trends emerged. First, in the aggregate, organizations are experiencing significant “copy sprawl,” reporting that the typical data set is copied and stored an average of six times (see Figure 1).

Figure 1. On Average, Organizations Store Six Copies of the Same Data

For the typical data set, please estimate how many copies of the same data are created in your organization (i.e., on average, how many times is a data set copied and stored separately across all locations)? (Percent of respondents, N=200)

Estimated mean: ~6



Source: Enterprise Strategy Group

¹ Source: ESG Research Insight Paper, [Mass Data Fragmentation Is Quietly Killing Digital Transformation Efforts](#), May 2019. All ESG research references and charts in this technical review were taken from this research insights paper, unless otherwise noted.

Second, CIOs are alarmed at the number of copies they have under management, especially considering that these copies are often spread across many locations: 73% of respondents report their organization stores data in multiple public clouds today in addition to their own data centers.

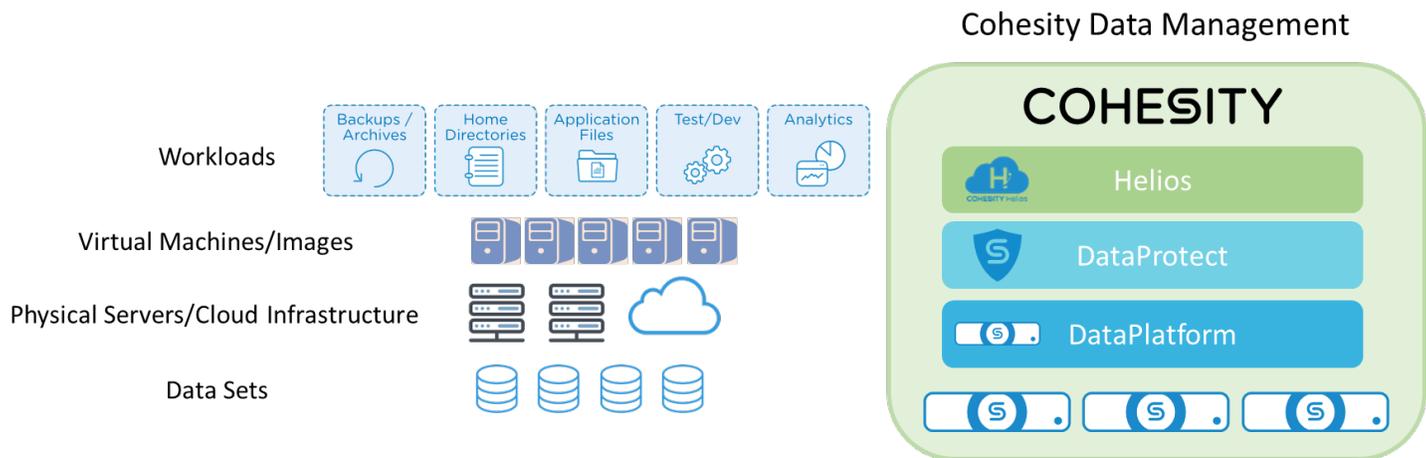
The Solution: Cohesity Data Management

Cohesity DataPlatform is an intelligent, scale-out, software-defined platform that runs on hyperconverged nodes with compute, flash, and storage capacity. Cohesity delivers an efficient general-purpose storage platform, which scales as organizations add nodes to the cluster. Cohesity’s variable-length deduplication engine runs globally across the cluster and can execute inline, post process, or not at all depending on the workload and data attributes. Cohesity’s patented SnapTree snapshotting technology uses a tree structure of pointers in contrast with traditional link-chain metadata journaling. Cohesity designed SnapTree to limit the number of hops to retrieve data blocks to two, to enable the platform to offer virtually unlimited snapshots without impact to performance.

Cohesity provides fully converged data protection with DataProtect software, a complete backup and recovery solution. Cohesity can also be used as a target for third-party backup applications including Veeam, Commvault, and NetBackup, as well as dedupe storage for database dumps and copies taken with native database tools such as Oracle RMAN. Administrators use DataProtect to manage backup jobs, establish policies, and apply these policies to groups of servers or VMs; DataPlatform itself manages global deduplication and compression.

Cohesity eliminates the complexity of deploying and scaling separate media servers, master servers, proxies, media agents, and target storage; DataPlatform scales simply by adding hyperconverged nodes to the cluster, each of which brings CPU and storage capacity, expanding the global filesystem. It is tightly integrated with VMware, including the VMware APIs for Data Protection (VADP) that provide Changed Block Tracking (CBT) so only incremental changes are transferred. Cohesity DataPlatform includes near-unlimited snapshots that can be created on demand; instant restore at scale; global deduplication across the cluster; sub-five-minute RPOs; a simple UI for end-to-end workflows; policy-based automation; linear scalability; granular VM-, file-, and object-level recovery; AES-256 encryption; and integration with tape libraries and the cloud for long-term archiving. Once data is backed up on Cohesity, it can be used for other use cases such as spinning up zero-cost clones for test/dev and analytics.

Figure 2. Cohesity Data Management Platform



Source: Enterprise Strategy Group

Data is intelligently tiered between HDD and flash, and the cluster maintains two or more copies for availability. Cohesity has multiple data redundancy options that administrators can select based on their needs. Data replicates automatically within the Cohesity cluster and is optimized for fault tolerance. Upgrades, node addition or removal, and other maintenance tasks are non-disruptive.

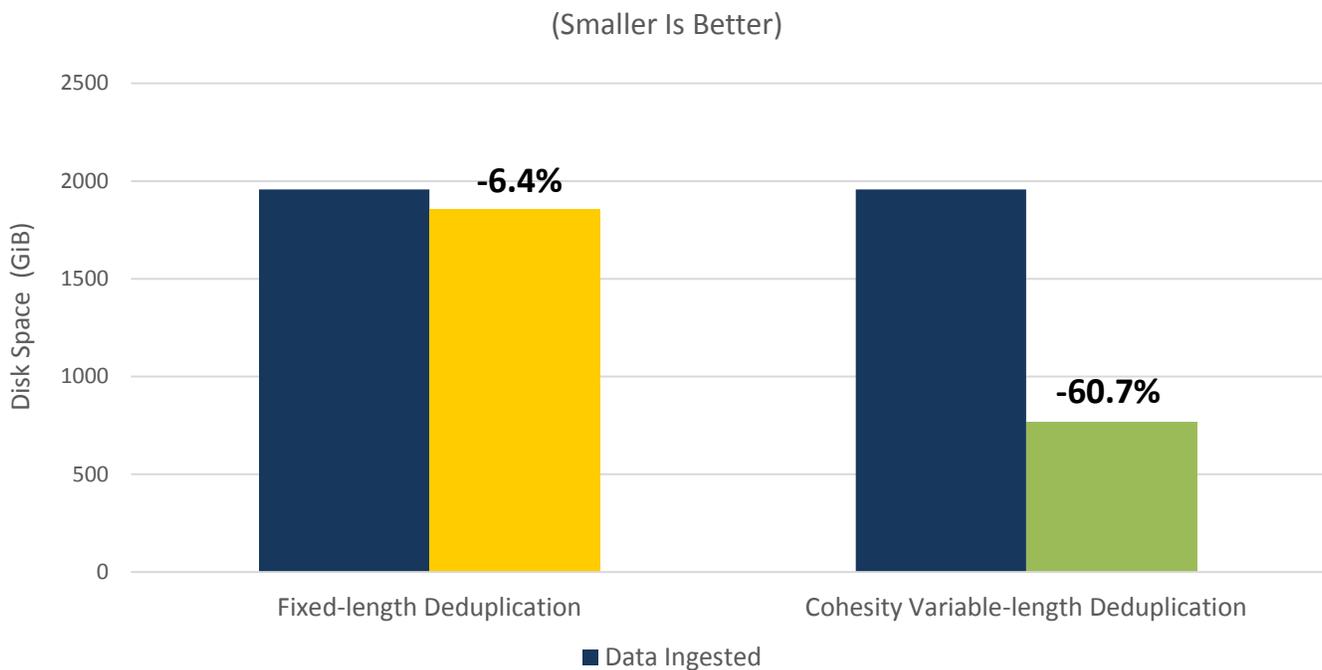
Cohesity Helios is a SaaS-based application that works in concert with Cohesity DataPlatform to provide a single view and global management of all of an organization’s data and workloads wherever it resides—on-premises, public cloud, or the edge. Cohesity Helios utilizes machine learning algorithms to proactively assess IT needs and automate infrastructure resources.

ESG Tested

ESG examined data reduction and small file efficiency. We compared Cohesity DataPlatform global variable-length deduplication and small file optimization to a system that utilizes fixed-length deduplication and does not deduplicate globally across views—Vendor X. We used synthetic data sets generated specifically for these tests: A 1TB unstructured data set with a duplication factor of 2:1 for the deduplication test, and a set of 1 million 1KB files for the small file efficiency test. Erasure Code (EC) redundancy was set to 2:1 in both systems.

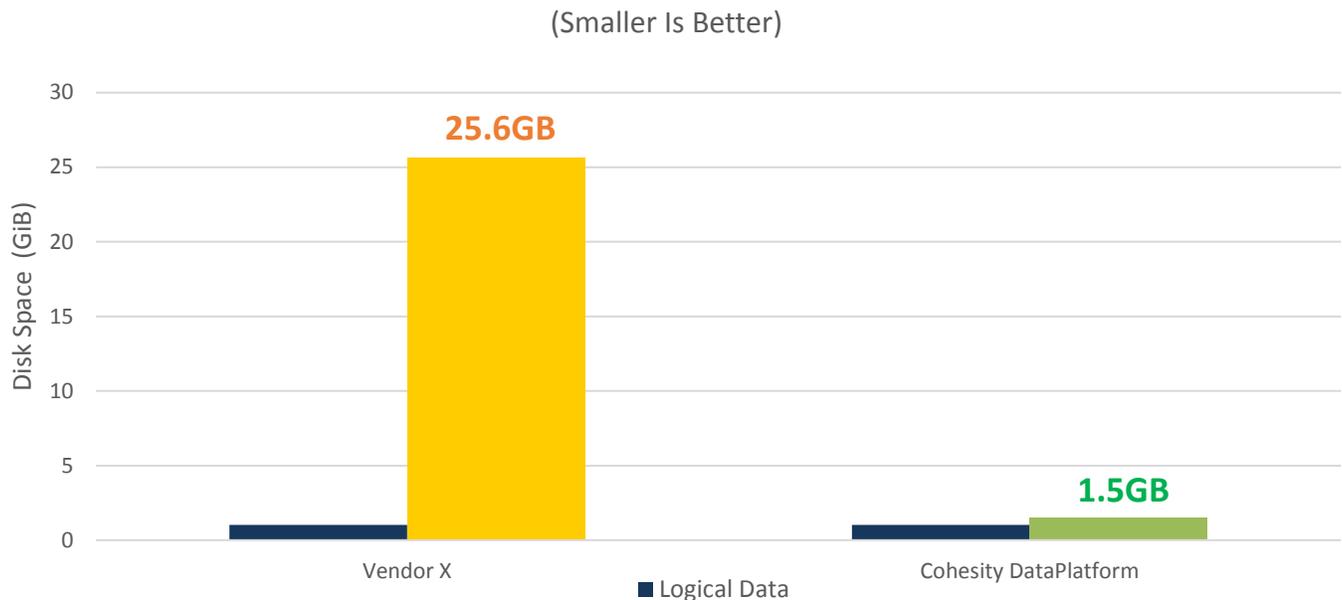
In the first series of tests, we copied the 1TB unstructured data set to two views—think volumes—in the Cohesity platform and two views in the alternative system. As seen in Figure 3, Cohesity’s global deduplication reduced the data footprint by 60.7% while the non-global deduplication system was only able to achieve 6.4% data reduction.

Figure 3. Cohesity Variable-length Deduplication Versus Fixed-length Deduplication



Source: Enterprise Strategy Group

Next, we looked at small file efficiency. We copied 1 million 1KB files to a view on the Cohesity DataPlatform and a view on the Vendor X system. To isolate the effect of small file optimization, we disabled deduplication for this test and set EC redundancy again to 2:1 in both systems. Due to its 8KB block size and requirement to store three copies of the data for redundancy, Vendor X inflated the data to 25.6GB. Cohesity was able to store the data on disk using only 1.5GB of capacity (see Figure 4). It is important to note that this test was designed to highlight the small file efficiency of the Cohesity platform and does not represent a real-world application by itself. Organizations that have a sizable number of small (sub 8KB) files should see some benefit, depending on the workload and how the files are written.

Figure 4. Small File Efficiency

Source: Enterprise Strategy Group

Cohesity Data Management in the Real World

ESG spoke with a customer using Cohesity to understand how it has performed in their environment. The customer is a service provider with approximately 70 clients in a six petabyte multi-tenant environment. This customer selected Cohesity to replace a backup storage environment, but quickly expanded its use to cover network shares and some production workloads as well. The consolidation of NAS and production workloads with the backup environment extended the reach of capacity efficiency beyond backup, improved staff productivity, and enabled instant restore. The size and scope of their environment has brought the usual challenges. They purchased six petabytes of storage over the course of nine months and were an early adopter of multi-tenancy. This customer is quite happy with Cohesity, citing their excellent support and ease of ongoing management.



Why This Matters

ESG research clearly shows that organizations today are dealing with extremely fragmented data assets, but what are the ramifications? Respondents told ESG that 42% of administrators' day-to-day tasks—on average—concern managing their organization's secondary data, applications, and copies across on-premises and cloud environments. If 42% of a typical IT admin's job is managing fragmented data, then 42% of their cost is ultimately wasted in non-productive, non-profitable endeavors. Worse, 49% of respondents believe that MDF leads directly to overworked employees. Since we know that employees seek intelligent, meaningful tasks as a condition of employment and job satisfaction, organizations are making their businesses undesirable places for the best talent to work by continuously demanding skilled employees to perform menial, mundane jobs.

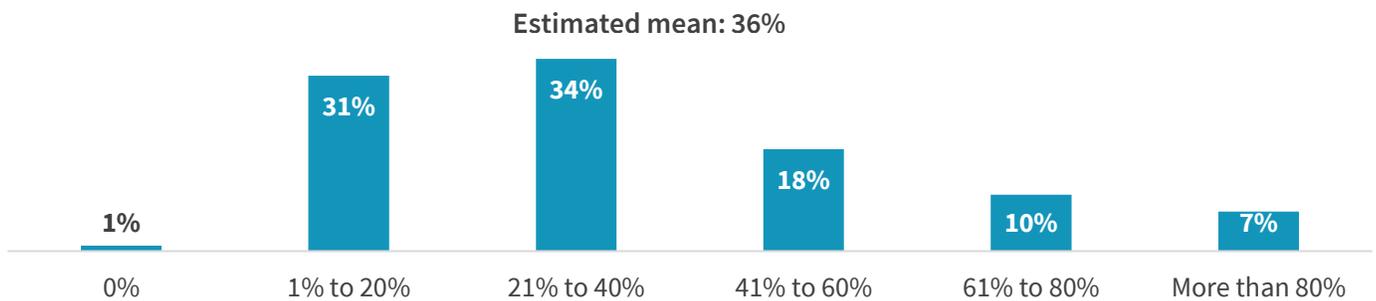
Cohesity global, variable-length deduplication was able to provide nearly 60% data reduction in testing audited by ESG, compared with just 6.4% reduction for the tested non-global deduplication system. Cohesity also clearly demonstrated greater small file efficiency. Cohesity was able to store one million 1KB files on disk using only 1.5GB of capacity, where the alternative system tested increased the footprint of the files on disk to more than 25GB. This translates to easier integration into an organization's existing environment and substantial savings in disk capacity, network bandwidth, and administration.

The Bigger Truth

Data growth is the root cause of all data fragmentation, but it is also a universal truth among digitally enabled organizations. Among all organizations surveyed by ESG, expected growth in secondary storage capacity for the remainder of the calendar year is significant: 36% (see Figure 5). Interestingly, the largest organizations ESG surveyed also expect to experience the fastest growth. Enterprises as a cohort expect a mean growth of 38%, while midmarket organizations expect 29%. This means organizations with the biggest data management jobs to do today are going to be under the greatest data management pressure as time passes.

Figure 5. Organizations Expect Storage Capacity to Increase in 2019

By approximately how much do you expect your organization’s total secondary storage capacity to increase between now and the end of 2019? (Percent of respondents, N=200)



Source: Enterprise Strategy Group

Cohesity has created a way to eliminate not only costly redundant data copies, but also the redundant infrastructure silos organizations buy and manage to act on that data. Cohesity’s DataPlatform is a single, highly scalable, intelligent appliance that first protects data, and then uses the copy you already have for production and to create snapshots for secondary uses. There’s nothing wrong with integrated backup appliances, deduplication target appliances, or copy data management solutions. The question you should be asking yourself is: Why have separate data copies, hardware, and software, when you can have them all in a single, scalable appliance that leverages one copy of data?

ESG validated that the Cohesity DataPlatform can house all these copies of an organization’s data, deduplicating across views, applications, and workloads, with a highly efficient mechanism for storing small files. The platform is expandable and flexible to meet the needs of each customer. The ability to scale incrementally means the environment can grow organically as needed, with predictable cost and performance. It offers greater insight into data, with simpler management.

Organizations need to consider how to store, protect, access, govern, and retire data based on each data operational area—and stop just assuming “someone will do it.” ESG believes that a more intelligent plan can save incalculable time and money, and lead to much better knowledge worker productivity, all of which mean a better bottom line. If your organization is ready to address the challenges of MDF, a look at Cohesity’s data management platform would be a smart first step.

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of The Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at 508.482.0188.

The goal of ESG Validation reports is to educate IT professionals about information technology solutions for companies of all types and sizes. ESG Validation reports are not meant to replace the evaluation process that should be conducted before making purchasing decisions, but rather to provide insight into these emerging technologies. Our objectives are to explore some of the more valuable features and functions of IT solutions, show how they can be used to solve real customer problems, and identify any areas needing improvement. The ESG Validation Team’s expert third-party perspective is based on our own hands-on testing as well as on interviews with customers who use these products in production environments.