COHESITY

Buyer's Guide for Modern Web-Scale Backup and Recovery

A Comprehensive List of Requirements for Evaluating a Modern Solution

Contents

Introduction: The Data Deluge is Underway

As organizational needs change, and workloads become increasingly distributed, a key realization is emerging: traditional approaches to backup and recovery are no longer sufficient for most organizations. Companies have discovered that their existing tools are not keeping pace with other advancements in their IT environment, such as hyperconverged systems and software-defined infrastructure, which seek to reduce data center complexity and help deter the surging total cost of ownership.

Today's backup and recovery landscape is littered with separate legacy point products for backups, target storage, and long-term data retention. It's a complex environment to manage since each of these silos is designed on proprietary hardware and/or software packages that typically have their own management tools, upgrade cycles, and maintenance and support contracts.

At the same time, as cracks have begun to appear in the backup and recovery foundation, organizations are creating and consuming more data than ever before. As the saying goes, "data is the new currency" for enterprises, and it's exploding all around us. Organizations across the globe are experiencing a data deluge, with the quantity of data increasing at an accelerating rate as new types of information are assimilated into existing data management systems.

No longer are companies storing only traditional routine business data. For some companies, the need to store human-generated data is paramount. But today, businesses are also absorbing data from a myriad of sensors and machines. And these devices are distributed across the enterprise.

Another significant part of the data growth challenge revolves around supporting a far broader set of applications than in the past. And as cloud becomes an integral part of the overall IT ecosystem, modern applications and their data are residing both on-premises and in the cloud, creating new silos that lead to mass data fragmentation—a digital transformation roadblock

The point is that data growth is real and will continue, as the adoption of an ever-increasing number of applications shows. But, to paraphrase the old cliché, with great data comes great responsibility. Regardless of the source, enterprises need to keep pace with this data growth as they consider backup and recovery capabilities.

The reality however is that oftentimes customers are already at their limit and at best only able to perform backup jobs once a day, with such jobs sometimes bleeding into their "working hours" production windows. Perhaps even more importantly, should rapid recovery be necessary—say as the result of a ransomware attack—it can take multiple hours to restore, which has serious business implications. This is a recipe for IT tragedy should disaster strike.

Time for a Change?

Historically, organizations have invested significant portions of their IT budgets on backup and recovery tools, which they have accumulated over time. They continue to protect their data with legacy backup and recovery solutions that were designed for a different era of computing.

With some products having their origins tracing back decades, many such tools are simply unable to keep pace with modern business requirements and constant advancements. As business needs evolve, organizations must strive to stay ahead of growing IT complexities. Enterprises need IT infrastructure, including backup and recovery tools, that's simple to manage, agile, and easy to scale.

If you're considering an upgrade to your existing legacy backup, and recovery capabilities, you're far from alone. Gartner predicts that "… by 2021, 50% of organizations will augment or replace their current backup application with another solution, compared to what they deployed at the beginning of 2017."

Choosing the right enterprise-grade, modern, web-scale backup and recovery solution involves understanding the full list of potential problems first, then comparing solutions to those problems. This guide is your handbook for selecting the product that best meets the needs of your organization, for today as well as tomorrow. As your business evolves, you need a backup and recovery platform that can evolve with you.

The primary objective of this Buyer's Guide is to help you determine criteria as you consider your next backup and recovery solution. Inside this guide, you will discover:

- The fundamental questions to ask about your current environment
- Common backup and recovery issues currently faced by enterprises
- What to look for in your next backup and recovery solution, to future-proof your investments
- RFP-ready questions you can use to ensure that your selection process is complete

As is the case with so many technologies, the community around the backup and recovery space is chock-full of often conflicting opinions. Some of the information out there is calculated to create fear, uncertainty, and doubt in the minds of buyers. This guide will help you cut through the noise and provide direction and confidence as you journey through your modernizing of your backup and recovery environment.

As you review the evaluation criteria for a modern backup and recovery solution, take particular note of the questions you should ask vendors. If you're talking with them in person, make sure you get complete, clear, concise answers to these questions. If you're preparing an RFP for a new backup and recovery solution, these questions will help you get the answers you need.

Assessment: Backup and Recovery Challenges Facing Enterprise IT

Before you get started with a selection process, it's important to understand any potential shortcomings in your existing backup and recovery environment. Table 1 outlines the 7 common backup and recovery challenges faced by IT. Take a few minutes to determine your starting point to help best discover where to kick off your search.

Issue	Description	Present in Your Environment?
Siloed infrastructure with multiple, fragmented UIs to configure backup workflows	Does your current solution require the use of backup software along with separate media servers, dedicated storage targets and multiple user interfaces to create backup workflows? Do you need to operate separate environment for backing up physical workloads and another silo for virtual servers, and yet another product and console to help protect databases and applications?	☐
Bolt-on cloud gateways	If your current system provides any kind of support for the public cloud, does it require you to deploy separate bolt-on cloud gateways that act as intermediaries between your on-premises and public cloud-based backup and recovery environments?	☐
Forklift upgrades and disruptive updates	When the time comes to grow the backup and recovery environment, do you need to rip-and-replace your existing infrastructure with a new one, which is expensive and complex? Do you need to schedule downtime to perform software updates?	☐
Slow restores, last point in time only	Does your current solution suffer from performance problems at restore time, resulting in the potential for RTO misses? Or does it just allow you to recover from only the latest backup made?	☐
Variable and/or fixed block deduplication with compression	Does your current solution lack data deduplication features? Or, if it includes them, does it have restricted deduplication domains or fixed block deduplication capability only?	☐
Protection against ransomware attacks	Are you confident that your existing backup solution can effectively defend against a ransomware attack? Will your current solution help you detect the threat and respond quickly to reduce downtime?	☐
Backup data cannot be reused; infrastructure remains siloed	Does your current solution provide only backup and recovery services? Does it allow you to reuse the data for other use cases such as dev/test, security and compliance?	☐

Table 1. A checklist for profiling your current backup and recovery environment.

If you checked one or more of the boxes in Table 1, consider reviewing options for replacing your existing backup and recovery software. There are options on the market today that have rethought backup and recovery with a modern, web-scale point of view to address all these deficiencies in elegant, simple, and affordable ways.

Evaluation Criteria for a Modern Backup and Recovery Solution

The evaluation criteria you used for yesterday's backup and recovery solution no longer applies. A new way of thinking is necessary. It's time to stop thinking about backup and recovery as an expensive insurance policy that is siloed by services and responsibilities. Today, these capabilities need to be a core part of your infrastructure, not a bolted-on afterthought that adds more complexity to the environment.

Further, today's modern application-driven workloads require smooth data mobility between on-premises environments and the cloud. Organizations need backup and recovery solutions that are agile, offer infinite levels of flexibility, and can respond quickly to changing business requirements.

The sections that follow detail the critical attributes of a modern backup and recovery solution.

Unified, Software-Defined Platform

With IT under more pressure than ever before, and as organizations seek to complete critical digital transformation initiatives, there is decreasing tolerance for complex data center environments to support what are considered routine operations. As crucial as backup and recovery solutions are to ongoing enterprise health and business continuity, you can't afford solutions that introduce needless complexity.

From an architectural perspective, what does a modern, forward-thinking backup and recovery infrastructure look like? As you consider how your backup and recovery environment correlates with your production environment, you'll discover that you need a backup and recovery solution that isn't static. It needs to be able to grow with your production environment grows, and it needs to be flexible to address complex and evolving business requirements. For example, as you add new applications, expand to the public cloud, or address changing security and compliance requirements, your backup and recovery environment shouldn't take months of planning and weeks of downtime to expand to accommodate them.

When you review your backup and recovery options, consider what you need to buy. With some solutions, you need to procure multiple point products from different vendors, including:

- Backup software
- Media and master servers
- Target storage devices
- Compatible tape devices for long-term archiving
- Cloud gateway(s) to integrate with public cloud

A key goal in your backup and recovery journey should be simplicity. Modern data center architecture options—such as software-defined infrastructure—can help you to bring simplicity to your backup and recovery environment. Software-defined infrastructure has emerged as a leading contender for primary production workloads. Given the principles of software-defined and web-scale, it's an even better fit for backup and recovery needs. Why? What makes a unified, software-defined solution so well-suited to support comprehensive backup and recovery needs? First, it eliminates silos. You already have enough silos in IT. You shouldn't need to 1) break down your backup and recovery environment into compute, storage, and software silos to be managed separately and 2) maintain parallel environments to support multiple workloads.

The opportunity to consolidate disparate silos into a single, unified solution that integrates the backup software and target storage in a single solution makes it possible for you to quickly and easily deploy—and scale—your backup and recovery environment as needed, anywhere.

A modern backup solution needs to ensure that all the workloads you're operating can be protected on a single platform. As such, it needs to be able to protect your bare-metal servers as well as all your virtualized environment (VMs and containers), provide full support for your traditional and modern enterprise applications and databases; and protect your primary storage systems and NAS devices.

While evaluating your next modern backup and recovery solution, the IT practitioner should look at an offering that eliminates silos by consolidating backup software and target storage onto a single solution, while eliminating the need for unnecessary master and media servers. And that solution should seamlessly sync with long-term storage media, whether it's public cloud or tape.

And, while collapsing the hardware and software silos inherent in legacy backup and recovery products is a worthy goal by itself, there's also upside opportunity when you consider the potential of making backed up data productive to support other workloads, such as analytics, security, compliance – even application development and testing.

A key issue that results from a legacy approach to data management is *mass data fragmentation* and, without an appropriate infrastructure environment, it's practically impossible to fix. Mass data fragmentation refers to the growing proliferation of data across workloads, within workloads (e.g. backup and recovery), across organizational functions and locations (e.g. on-premises and public clouds). This fracturing of data elements into separate silos reduces the potential for deriving value from that data. Each further fragmentation makes it harder for an organization to be able to get a truly comprehensive understanding of which data elements are in inventory and therefore complicates efforts to analyze it.

Moreover, each silo often has its own infrastructure in some way, with the result being untold copies of data floating around different locations and systems. Such fragmentation increases costs due to the need for additional infrastructure and makes it more difficult to have confidence that one particular copy of a data set is the "right" one for analysis.

A unified, software-defined platform can provide you with the potential to consolidate a range of workloads into a single cohesive environment that enables high levels of scalability and visibility. By consolidating all copies of data into one platform, you significantly reduce your overall capacity requirement while increasing your ability to exploit data for competitive advantage due to increased confidence in the completeness and reliability of remaining data sets. The ability for a unified software-defined platform to solve the problem of mass data fragmentation should be a key consideration for organizations seeking to undertake digital transformation initiatives that require efficiency and consistency across organizational data.

Questions to ask vendors about their solution's architecture

- Does your solution offer the flexibility of a software-centric architecture that can be deployed on-premises, in the public cloud, or edge locations?
- Can you provide details on how your solution scales and how that impacts performance?
- Can your solution help in the reusing of backup data for other workloads beyond recovery?

Single, Global Management Interface and Dashboard

Once any new system is up and running, management becomes the next potential pain point. And unlike deployment, management is something you do every day, making ease of use not just a "nice to have" feature, but an absolute necessity.

Simplicity in every aspect of IT infrastructure is an increasingly important requirement. Modern backup and recovery solutions need to go beyond user-friendliness, however. They must also provide comprehensive support for all data protection use cases from a single console, including disaster recovery, archival/long-term data retention, storage tier and replication, spanning from on-premises, to public cloud, or the edge.

At first glance, it may seem okay to have different consoles to back up different kinds of data, but backup and recovery isn't just about backup. It's all about recovery, too.

When you're in recovery mode, minutes count. Confusion can be high, and answers can be in frustratingly short supply. You don't need an overly complex administrative experience further impacting your recovery efforts.

From beginning to end, a backup and recovery tool needs to provide a simple, intuitive, single interface that can manage all aspects of the environment globally, including initial deployment of the backup and recovery hardware and software, ongoing management of backups, and eventual recovery across locations.

Of course, a fully converged data protection and management environment goes far beyond backup and recovery, meaning that the management paradigm should encompass both data protection as well as myriad other uses while retaining a focus on simplicity. In modern terms, simplicity includes the assumption that the infrastructure environment will make efforts to learn what's important to you through machine learning-based (ML), personalization and automated recommendations around adherence to best practices based on your industry. An ML-assisted approach means that operations can be performed more quickly and more accurately, which can help you meet your organization's service level agreements (SLAs).

In other words, the management layer should act not just as a console, but as an intelligent global assistant, helping you identify anomalies, including detecting potential ransomware attacks, and making orresponding remediation recommendations, such as for additional capacity planning. This too should fit within the confines of a single global management interface.

Questions to ask vendors about their solution's manageability

- How many consoles are required to provide comprehensive backup and recovery support for bare metal servers, virtual machines, traditional and modern databases, container environments, and SaaS-based applications?
- Can your solution help me with comprehensive capacity planning across my environment?
- Does your solution provide machine learning capabilities to help me make more efficient use of my time and improve adherence to SLAs?

Support for Both Traditional and Modern Data Sources

Today's data sources are vastly different than the ones that administrators supported just a few short years ago. You're likely more than aware of the fact that most organizations maintain some kind of hybrid cloud environment, with many applications residing on-premises and many more operating in the cloud (SaaS and/or IaaS). But as you examine the next level down in each of these deployment scenarios, the data sources themselves have evolved. There are still virtual machines, file servers, database servers and the like, but these have been supplemented with modern data sources like containerized applications based on Kubernetes or Docker Swarm, distributed NoSQL databases, and Hadoop. These types of applications and data sources didn't really exist all that long ago, so the need to protect them and support them is also new.

There are other services that didn't exist all that long ago either, including unstructured data repositories such as Microsoft Exchange Online and OneDrive, part of the Office 365 SaaS offering. Increasingly customers are discovering that the SaaS applications they rely on provide data availability but not enterprise-grade backup and recovery. Consider Office 365. Microsoft keeps the service up and running but only provides minimal protection in terms of recovery in the event of a failure. Most organizations, though, want more than the minimum when it comes to something as key as Office 365.

More and more, organizations need to add better protection for SaaS-based workloads such as Office 365, Salesforce, and Workday. Backup and recovery for SaaS offerings needs to be able to simply make a copy of SaaS application data and store it either into an on-premises cluster, or perhaps into another cloud-based target such as Azure or AWS.

Your modern backup and recovery solution needs to support all the applications you've always run, but it also needs to support new constructs as they are introduced to the market, including popular SaaS tools, container-centric services, NoSQL services, and much more. These are the tools that make the modern enterprise tick and they need to have the same level of support as more traditional workloads. Backup and recovery associated with these modern workloads needs to be as seamless as it is for legacy ones.

Questions to ask vendors about their solution's workload support

- Can your current solution support data sources on-premises and in the public cloud?
- Can your current solution support traditional and modern data sources, including SaaS, containerized applications, and distributed databases?
- Does your current solution leverage a single policy framework and user interface to protect traditional and modern data sources?

Predictable and Flexible Recovery

At the end of the day, the entire purpose of a backup environment is to enable recovery. That's it. That singular focus doesn't mean that every product on the market makes this a simple, flexible and predictable process, but that's exactly what's needed. Whether your recovery moment comes when a user has lost a single file or it comes due to a disaster that requires a mass recovery, it seems like the restoration process is often far too stressful.

Your recovery process should be the simplest and most predictable part of the equation and shouldn't take days to complete. Moreover, locating the right data to recover shouldn't be an exercise in frustration. Imagine at your fingertips a backup solution architected around recovery. This solution features a comprehensive global search and recovery feature that indexes every protected file so that you can find any file you need, no matter where it lives in the production environment for rapid restore. That's the kind of capability that is critically important to today's enterprises.

Recovery isn't always a simple matter of recovering a single file or folder. Often, you need to recover entire 3-tier virtual environments, or you've gotten hit with ransomware and need to recover hundreds or thousands of files and their dependencies as a result. Traditional backup tools can get in the way, requiring long waits and offering no visibility into the recoverability or vulnerabilities while data restores, a situation that isn't always palatable in a world in which near-zero RTO is an objective.

Your backup and recovery solution needs to have the ability to instantly and completely recover the last known good version of your environment, whether you need to restore a single large file, a single VM, hundreds of VMs, or everything. The solution also needs to be able to ingest and recover large files easily.

In this section, it's also important to note that the same solution service needs to be able to go deep – right down to the file or VM level – as well as go broad. You shouldn't have to compromise on granularity. Your selected solution should enable whatever level of granularity you need so that an already stressful recovery operation doesn't become an impossible one..

Questions to ask vendors about their solution's recovery capabilities

- Can your solution provide a global search function that spans my entire backup environment (on-premises, the public cloud, and edge locations)?
- Does your solution support instant mass recovery of objects, including the entire application environment?
- What are my options in terms of recovery granularity?
- Can your solution do backup verifications to ensure recoverability?

Cloud Native

Use of the public cloud is maturing with enterprises moving traditional on-premises applications to the public cloud. As a result, today's organizational IT infrastructure extends well beyond the four walls of the local data center. Your data backup and recovery environment needs to account for this evolution as well.

The cloud has emerged as a key enabler in terms of data backup and recovery. More companies than ever are turning to the public cloud as their backup destination of choice rather than using an on-premises data center. In many cases, cloud has become the new tape. The TCO cloud offers for archival storage is hard to beat. Of course, it's also faster to get up and running and allows for an OpEx for CapEx swap out. The ability to leverage the elasticity and economics of the public cloud for backup and recovery and other workloads is enticing and should be a native capability of modern platforms.

Regardless of your current state of adoption of public cloud technologies, your next backup and recovery product needs to include native integration with the public cloud. You may tell yourself that your current legacy backup and recovery product does fine providing support for the public cloud. Perhaps you've purchased a license or appliance that bolts public cloud functionality onto the product or added a gateway appliance that adds cloud support.

Beware, though: these kinds of bolt-on accessories are often incomplete, expensive, and add significant complexity to the operations of your backup and recovery environment. Given the importance of the cloud, can a solution built in the pre-cloud era really address your evolving business needs? Unlike these legacy backup products, a modern backup and recovery solution eliminates the need for bolt-on cloud gateways and supports cloud integration and native format data mobility.

A modern backup solution allows enterprise IT owners to take advantage of policy-based automation to seamlessly move their data between on-premises and cloud infrastructure. Use cases include not only cost-effective storage, but once the data is there – such mobility can leverage cloud computing to accelerate application development (e.g. spin up of a new dev/test environment) and analytics.

Disaster recovery is another compelling cloud use case. Recovery is all about getting workloads back into an operational state following an incident of some kind. An incident can be as simple as an accidentally deleted file, or as severe as the complete destruction of your data center. In the case of the former, recovery isn't that painful. For the latter, however, recovery can be incredibly challenging without the right tools in place.

To that end, the backup and recovery solution you choose should enable provisioning of workloads to either the original cluster or an entirely new location, be it in the cloud or to an on-premises facility across town. In this way, immediately following a disaster, you'll be able to spin up your workloads while you work on your longer-term recovery efforts.

As you consider a new backup and recovery solution, make sure that you consider the entirety of your application environment, including any applications that are operating in the cloud.

Questions to ask vendors about their solution's cloud native features

- Does your solution support easy data and application mobility into the public cloud?
- Can your solution help me protect my cloud-native applications?
- Once again, can you protect SaaS application data?

Limitless Horizontal Scale-Out

Just like your production environment, your backup and recovery environment isn't static. It's constantly changing in lockstep with the changes you make in your production environment.

An inflexible backup and recovery solution leads to increased costs and delays in deploying new production systems as IT operators are forced to contend with adjusting the backup and recovery environment to accommodate new workloads.

In an era where speed-to-market impacts the bottom line, it's unacceptable for your backup and recovery environment to hold a business hostage. It's even more intolerable when you consider the fact that unified, software-defined solutions exist on the market. These kinds of solutions are purpose-built to make scaling such environments easy.

This type of scale-out architecture to address business needs should also be true of modern backup and recovery solutions. With legacy backup and recovery products, you're forced to make critical sizing decisions at the start of your deployment, meaning that you have to attempt to predict your backup and recovery capacity needs three to five years in advance. Who can do that with any real degree of confidence (and potentially tie up valuable CapEx in the process)?

Similar to the way you build modern primary environments, your backup and recovery should also be designed on web-scale principles. Most importantly, the solution you choose should allow you to start small at the beginning and scale out as your needs dictate. That means that you should be able to start with whatever you want, without the need to overprovision infrastructure from the outset.

Then, as your business requirements evolve, your production environment will also grow, which should be followed by growth in your backup and recovery needs. Make sure you can easily scale that environment on demand without disruption, complications and huge cost to address growing business requirements. A modern solution that limitlessly scales linearly, without any impact on workload performance.

Questions to ask vendors about their solution's scaling features

- Does the product use a scale up or scale out expansion methodology?
- To what level can you scale the backup and recovery environment?
- Will we experience any performance impact at scale?

Non-Disruptive Online Upgrades and Expansion

Fact: Today's enterprises expect solutions to be up 24/7. Unlike days past, unplanned outages are unacceptable and even planned outages must be very few and far between, if they happen at all. The old method of taking systems down to patch, upgrade, or augment them is no longer palatable to organizations. At the same time, it's important to keep up with the latest and greatest software. Modern backup and recovery solutions allow backup admins to upgrade their clusters without any downtime, using rolling upgrades. A "rolling upgrade" means that each node in the cluster is upgraded individually, leaving all services operational on the remaining nodes for the duration of the maintenance.

This rolling upgrade paradigm also extends to eventual node replacement, as refresh cycles come and go. The concept of "forklift upgrades" doesn't exist in modern backup and recovery solutions. The administrators can introduce or retire nodes on the fly. Never again should you have to rip and replace a backup and recovery environment just because a depreciation schedule demands it.

Questions to ask vendors about their non-disruptive Online Upgrades and Expansion capabilities

- What process do you use to patch or upgrade the software components in the solution?
- Are your solution's hardware and software updates non-disruptive to the production environment?
- Will my lifecycle hardware upgrades require "forklift" replacement?

Guaranteed Data Resiliency

Organizations are not just backing up their data once a day anymore. Modern applications are accessing and modifying data continuously, so organizations back up their data more often.

During a backup job, if the target device fails, most legacy backup products have the intelligence to continue the backup on another target device from the failed point; but that process isn't as smooth if it involves third-party backup software. Take Oracle RMAN, for instance. In that case, a failed node requires the administrator to restart the entire backup job.

This need results in extended backup windows, and can potentially bleed into production time, impacting Recovery Point Objectives (RPOs) due to the eventually consistent data state the legacy solution supports. The impact of eventual consistency can be far more significant if the node fails during a recovery process. While the data is being recovered, some applications/users might be writing to the same data volume, and a node failure at that point can result in inconsistent and corrupt data.

To meet business SLAs, avoid "file not found" errors, and achieve guaranteed data resiliency, your next backup and recovery solution needs to be strictly consistent. In a strictly consistent model, the application or client only receives an acknowledgement of the write once the data propagates to multiple nodes. Thus, in the event the ingest node fails, the application or client won't receive a false acknowledgment that its data is protected or written. This results in keeping the data strictly consistent. Insist on a strictly consistent model, whether you're backing up directly or using a third-party application like Oracle RMAN or running a recovery process.

Questions to ask vendors about their solution's data resiliency

- Can your existing solution withstand the loss of two nodes, two disks, an entire chassis, or even a cluster?
- Can your solution help me meet my SLAs and guarantee data resiliency at scale?
- Does your solution only propagate the data, or does it also include the metadata?

How to Guarantee Data Resiliency

There are several methods by which data resiliency can be guaranteed. The first is at a hardware level. Hardware can fail or become otherwise unusable. The cluster needs to be able to handle such failures. Many solutions are architected to support the loss of a single node or a single disk in a node. Here's the critical problem with that limitation: During the outage of a node or disk, *any subsequent failure will result in data loss*. For true high availability, the environment needs to be able to withstand the loss of multiple hardware elements.

You may say to yourself, "We're never down, so this doesn't apply to us." Consider this: Your environment needs updates and upgrades from time to time. During these periods, nodes will be down while they're being updated, which can leave you in a vulnerable state until the process completes. Solutions that support only one node outage make updates a risky proposition.

A particularly resilient solution can support the loss or failure of two complete nodes or two disks in a storage domain. If you do experience a failure of a node or disk, this additional level of resilience ensures that your environment remains protected even while you're taking steps to mitigate the original problem.

There are two ways that this level of resilience is achieved: erasure coding and replication. Erasure coding is a method by which data is broken up into small chunks before being written to storage. An algorithm imbues each chunk with redundant pieces of data from other chunks before data is then written to various locations in the cluster. This method allows a complete data set to be reconstructed even when it's not possible to recover all of the individual chunks. Erasure coding provides both more resiliency than typical RAID systems and is also much more capacity efficient, requiring less overhead than RAID.

Replication is a second method that provides additional redundancy, particularly at the full node level. By making a complete copy of data, a cluster can withstand the loss of one or more nodes, depending on how many replicas you decide to enable.

To achieve the necessary levels of availability in modern enterprises, a combination of these two approaches provides both node-level and cluster-level protection in the event of a hardware failure.

Maximized Storage Capacity and Reduced Data Footprint

Data copies are exactly what they sound like: copies of your data that may exist in different parts of the environment. The problem of multiple copies has had such an impact on businesses that an entirely new class of solutions was formerly introduced to address it: data copy management. When you consider your backup environment, you may not even realize that there are dozens of copies of your data strewn about, particularly if you're using a bare-bones backup product. And that doesn't even take into consideration the data copies that you might make to facilitate other needs, such as data analytics.

***Data copies** in this context refers to the actual copies of data you or others in your company make to do work and can include backups, copies of data to be used for analysis, copies of data used for archiving, and many more. This is different than the data replication process discussed previously. Data replication is a critical part of your overall data resiliency solution whereas data copies are often inefficiently using storage capacity.*

No matter how many copies of your data are in use or in your backup environment, the total capacity consumed by that data should not grow appreciably. Automatic data deduplication is used to significantly reduce the number of copies of your data so that your storage capacity goes much further than it otherwise would.

Many organizations rely on separate deduplication appliances that create yet another silo and don't scale. In fact, most of these solutions only dedupe at the node level, which isn't efficient given that the same data set might be stored in another node within the same cluster.

There are many ways to achieve the outcomes that are desired with data deduplication, but many products support just one of them. To maximize the impact of data deduplication, combining different methodologies has a beneficial impact on both speed and capacity.

Modern backup solutions help IT administrators reduce their data center footprint with efficient data reduction techniques like global variable-length sliding window dedupe. In this model, the deduplication is performed using variable-length data deduplication technology that spans an entire cluster and dedupes across all workloads, including data stored on physical systems, inside VMs, databases, and more. It also uses efficient protocols that result in significant savings across the storage footprint.

Let's start with an overview of fixed length vs. global variable-length deduplication. Deduplication works by analyzing chunks of data to see if they match other chunks of data that already exist somewhere in the system. If a match is found, the new data chunk isn't written again. Instead a small pointer is inserted with a reference to the matching chunk of data that already exists. As the name implies, fixed length deduplication algorithms divide data inside fixed-size chunks, all of equal length.

The key problem with fixed length deduplication is what happens when there are minor offsets in otherwise common data. In the image below, it's clear that there is a lot of data commonality, but the minor offset means that none of the data in this example will be included in the deduplication process and, instead, it will all be written to disk as-is, which wastes disk space.

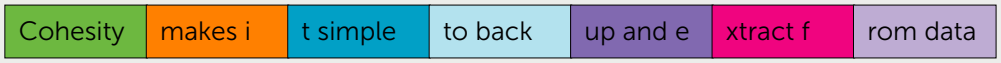
Adding this variable length "sliding window" component to deduplication means that your backup data can be squashed down into far less disk space than is possible with fixed length deduplication schemes. Making this capability even more powerful happens when the deduplication domain spans the entire cluster. In many deduplication algorithms, capacity savings only comes inside each node. If you have data copies that span nodes, those copies don't get reduced at the cluster level. By extending variable block length deduplication to the entire cluster, you get the potential to multiply your capacity savings.

Original Statement
Cohesity makes it simple to back up and extract from data

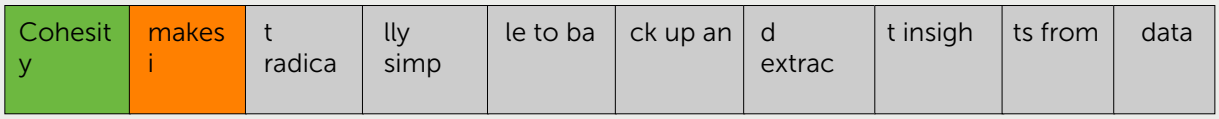
Edited Statement
Cohesity makes it radically simple to back up and extract insights from data

Fixed-Length Deduplication (8-character blocks)

Original



Edited (shows 8 new data blocks)

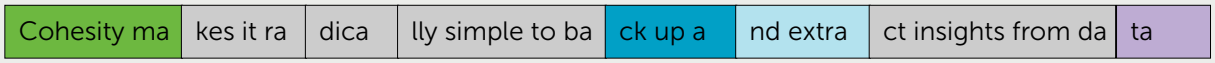


Vs.

Variable-Length Deduplication (after "a")

Original

Edited (shows 4 new data blocks)



** Gray blocks denote new data sets*

Variable-length deduplication is a method by which the analyzed block size isn't fixed in the algorithm. Instead, the algorithm studies the underlying data and dynamically imposes a block size based on the characteristics of the data being analyzed. This way, the minor offsets in the data stream won't result in a failed deduplication effort. The data that is actually common will be deduplicated, even if there's an offset in the way. In the image above, you can see the results of variable-length deduplication on the same data set as the previous image.

Adding this variable-length "sliding window" component to deduplication means that your backup data can be squashed down into far less disk space than is possible with fixed-length deduplication schemes. Making this capability even more powerful happens when the deduplication domain spans the entire cluster. In many deduplication algorithms, capacity savings only comes inside each node. If you have data copies that span nodes, those copies don't get reduced at the cluster level. By extending variable block-length deduplication to the entire cluster, you get the potential to multiply your capacity savings

Questions to ask vendors about their solution's storage capacity and efficiency

- What data reduction methods are supported by your solution?
- What is the average data reduction ratio across your install base?
- Does your solution provide a configurable replication factor?

Defense Against Ransomware Attacks

Security can be seen in multiple dimensions: one, the security of the backed up data and another, how backed up data can be leveraged to help organizations improve their overall security and risk posture. With increasing cyber threats, your next backup solution should have a built-in security-first approach that ensures your company stays out of the news, out of court, and continues to please customers.

While the legacy approach to backup offers some level of encryption and role-based access control (RBAC) mechanisms, modern threats, including sophisticated ransomware, incorporate numerous techniques to bypass and go undetected for some time to spread throughout an environment, including the backup data. Additionally, the legacy backup approach that relies on multiple point products exposes organizations to multiple entry points with no way to detect or rapidly respond to an attack. To defend against modern cyber threats, your next backup and recovery solution needs to go beyond AES-256 encryption and RBAC. The solution needs to defend backup data and the environment with multi-factor authentication, ensuring only authorized users gain access to the mission-critical data. In addition, the solution should include immutability and WORM on backup data, ensuring no application or user can modify the "gold" backup copy. And if the worst happens, your backup solution should offer flexible and rapid recovery at scale to reduce downtime.

Another dimension to the security discussion can be around leveraging backed up data to gain visibility and detect potential threats, like ransomware attacks on the primary environment or for discovering vulnerabilities in the IT infrastructure that can be proactively addressed, without impacting the production environment.

Questions to ask vendors about their solution's security and anti-ransomware capabilities

- What does your solution offer to keep my backed up data safe from accidental deletion as well as cyber threats, including ransomware attacks?
- Does your solution have a machine learning algorithm to detect a ransomware attack in progress, and help identify a clean copy to perform predictable recovery?
- After an attack, does your solution offer recovery at scale to reduce downtime?

Upside Opportunities: Extensibility and Compliance

As you consider your next backup and recovery solution, think bigger. Your backup doesn't need to be just an expensive insurance policy but can also support other business requirements. For example, the opportunity exists to use the backed up data and the backup infrastructure to stand up dev/test environments, or as a place to deploy an analytics application.

No longer does an agile developer have to wait on a data administrator to laboriously restore data to build out such environments. The solution can do so instantly, improving time to value and reducing overall costs as a result. The same improvements can be applied to analytics, enabling you to gain deep insight into the backup data without having to figure out where to store it all. As you seek out a new backup and recovery solution, look for one that goes beyond backup and enables value-add in other aspects of your operations.

Accelerate Application Development

Digital transformation is being driven by modern applications and the people who build them. Yet legacy approaches to backup make it challenging to empower developers and test teams with rapid access to high-quality data. Provisioning data to support dev/test should not take enterprises multiple days or weeks and multiple approval cycles, prolonging the time to build and deliver an application. Because it typically does, the legacy paradigm of provisioning data for dev/test does not align with today's modern development needs.

Your next backup solution must allow IT operators to instantly and securely provision high-fidelity data for developers and testing teams via zero-cost clones within the same environment. This will not only accelerate the development of high-quality applications within a single environment but also reduce data copies, which heighten your organization's security and compliance risk exposure.

Compliance

Today, compliance in backup and recovery products is becoming progressively more important, particularly as we continue down the path of cloud in an increasingly regulated world.

Backup administrators have to worry about data sovereignty, information governance, and governmental regulations such as the EU's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Data sovereignty is a regulatory requirement for some types of data, stipulating that certain kinds of data cannot leave a country's borders. In a cloud-centric world, it's often difficult to figure out where particular data is residing. GDPR is an overarching set of data protection requirements that aims to bring control of personal data back to the people. It also carries stiff penalties for companies that fail to follow its framework.

Related to GDPR, but also a much broader issue is the storage of personally identifiable information (PII) in any storage environment. Storage of unnecessary or improperly secured PII is one of the leading causes of non-compliance and fines today. Every backup and recovery solution, should have the ability to automatically scan the environment for PII and alert administrators when found. It may be perfectly valid for PII to exist on a system, but administrators need a clear understanding of where it all resides so that appropriate mitigation steps can be taken. After all, if you don't know where it is, you can't protect it.

Questions to ask vendors about their solution's extensibility and compliance features

- What capabilities does your solution offer to help me better support other business functions?
- Can I reuse my backed up data to meet compliance requirements, like locating PII data?
- Does your solution support zero-cost clones supporting application development within the same environment?

What Modern Backup and Recovery Looks Like

As you journey down the road toward implementation of a modern backup and recovery solution, it becomes apparent rather quickly that the solution landscape in this area has changed in a dramatic way in recent years.

No longer are you confined to thinking of your backup environment as a collection of silos. You can now have a unified and comprehensive platform, which can be easily managed through a single, global console. No longer do you have to dread the day that a ransomware attack strikes. You have a defense in place. No longer do you need to worry about the day when you max out your solution's capabilities and need to perform a forklift upgrade to rip and replace it. Modern solutions scale out, making expansion, upgrades, and node replacements a breeze.

But that's just the beginning. A modern backup and recovery solution provides capabilities that are either impossible or prohibitively expensive to achieve with legacy products. Native cloud features, mass instant recovery of VMs, and a comprehensive API to enable advanced workflows are all part of what you should look for going forward.

To learn much more about a leader in modern backup and recovery, visit Cohesity at www.cohesity.com/solution/backup-and-recovery/