rubrik | Zero Trust Data Management™

EBOOK

Protecting Unstructured Data

Table of Contents

Introduction

In today's digital economy, nearly every industry is finding that data is becoming increasingly critical to its core business success. Data enables innovation and discovery, powers advances in artificial intelligence and machine learning, and drives medical and earth-science breakthroughs. Most of the collective data fueling today's technological and scientific advances consists of unstructured data files – trillions of them. Large files, medium files, and small files are stored on NAS systems, Unix servers, Windows servers, and public cloud platforms – essentially anywhere and everywhere. And the number of individual files behind these innovations continue to increase at an exponential rate.

In order to safeguard your organization's investment in generating data, and protect the stake in the data you own, data-centric organizations like yours need to invest in actively managing those files through their entire lifecycle.

A key challenge in dataset lifecycle management is identifying data as it ages, and taking the appropriate action as each dataset moves into a new phase. Active files must be protected with regular backups. Less-active datasets need to be flagged proactively as usage rates taper off. And cold data needs to be archived to make room for the new data being generated every day.

One of the toughest hurdles for organizations whose unstructured data portfolio numbers billions of files, and/or petabytes of data, is maintaining an accurate, up-to-date count of those datasets and their usage patterns. Things like, how many files, where they are, how old they are, and whether or not they're still in active use.

Without reliable, up-to-date visibility into the full spectrum of critical business files, your organizations can easily be overwhelmed by the magnitude of your data footprint, not knowing where critical datasets are located, which datasets are still growing in size, or which datasets have aged out of use.

Data visibility — finding new and changed files, tracking ages and usage patterns — is the number-one prerequisite for successful management of unstructured data.

Defining and Organizing "Unstructured Data"

Generally, unstructured data refers to file-based data hosted on dedicated NAS (or Object storage) systems, as opposed to structured data (i.e. databases and virtual machine images) or files hosted locally on application servers. The key differentiators between unstructured data and other types of file-based data are in how the data was generated, and how the data is organized and managed during its lifecycle.

Data Types and Sources

Nearly all unstructured data at scale is generated via automated, machine-driven processes, e.g., medical imaging systems, cryomicroscopy devices, genetic sequencing platforms, electronic design systems, geospatial devices, and CGI/4K media platforms.

"Unstructured data" refers to machine-generated files and datasets, stored on enterprise NAS systems.

The type and volume of unstructured data that an organization must accommodate depends on the specific industry as well as the data source. Unstructured data can take any number of different forms, depending on its source. Medical imaging and microscopy platforms generally create image files; CGI/4K systems produce video files, while other platforms create either proprietary-format or text files. Some platforms, such as IoT ecosystems, can require huge amounts of specific data types for individual machine runs, and then generate huge amounts of additional data as part of the overall process.

Datasets

Unlike other types of file data, such as user home directories and shared corporate drives that are managed as individual files, unstructured data is managed at an aggregate level, with files from a particular machine run or generation cycle being managed – cataloged, protected, manipulated – as a single entity.

These management units, typically called "datasets", may each contain hundreds of thousands of individual files, and terabytes of disk space. In some organizations, dataset production can collectively add up to multiple terabytes of new, unstructured data on a daily basis.

Challenges and Opportunities of Data Protection at Scale

For many companies, unstructured data (i.e. files) was once generated and managed as a by-product of their primary focus – as a means for tracking customer information and order histories, or for internal employee documents around business development, etc. Most companies' unstructured data occupied a small, manageable footprint within both their enterprise environment and as a subset of their overall corporate asset portfolio.

In today's economy, whole industries have evolved around unstructured data in some form being the core of their business model. For these companies, data management, which encompasses data protection, data movement, and data tiering, is critical. Loss of any of that data, or even an inability to find any of that data, can mean loss of revenue, loss of business opportunity, even loss of the business itself.

As datasets continue to scale, some legacy NDMP solutions now require more than 24 hours to protect a day's worth of data.

The double-digit rate in year-over-year data growth that these organizations face – in many cases, their data footprint can double every three to four years, meaning that an enterprise with two billion separate files comprising 3PB of data in 2019 can expect to be responsible for a four-billion file, 6PB data footprint by 2023, and 8 billion files in 12PB of data by 2027.

It also means that organizations facing that level of annual growth have a series of challenges they need to address – how to prepare effectively for a petabyte (or more) of new data annually while simultaneously managing the data they already own. As a critical business asset, that unstructured data needs to be stored, accessed, protected, moved, and tiered. Which means having the right backup solution is vital.

Double-digit growth means that unstructured data will double every three to five years.

While backup as an objective can take any of a number of forms, a core set of distinguishing attributes of a backup solution, as opposed to an archive platform, requires that the solution do the following:

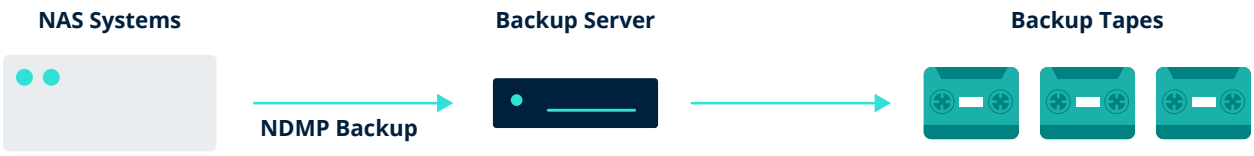
- Create a versioned replica of the production dataset
- Use a dedicated data platform, separate from the primary data source
- Retain data for a set period before being allowed to expire
- Store data in an "offline" state and not make if available for any active workloads

Historically, even large enterprises were able to rely on Network Data Management Protocol and disk-to-disk (D2D) backup solutions. Today, however, those same enterprises' data portfolios have grown beyond what Network Data Management Protocol can effectively protect. D2D for daily backups may offer larger-scale backup support, but can also be expensive and cause undue stress on production systems.

NDMP: An Old Solution for a New Problem

In the past, when all corporate data was centralized on a single NAS array and total enterprise data footprints were less than a few hundred gigabytes, a tape solution provided comprehensive data protection. Typically one full backup per week, supplemented with nightly incremental backups. Since then, unstructured data has grown at double-digit rates, and enterprise operations have gotten more complex, with heterogeneous storage platforms, geographically-dispersed business units, and the proliferation of new data sources. Despite advances in NAS performance, tape backup technology has failed to evolve to accommodate these changes.

Built on decades-old technology and practices, whether with tape or disk as the target, NDMP was engineered in conjunction with the idea of a backup window, which ran after daytime business operations had ended. It was assumed that the NAS system would be otherwise idle during backup cycles, which meant that NDMP could consume all available system resources during a backup run.



These fundamental assumptions about NDMP are still in place. Even today, despite decades of changes in NAS technology, NDMP still:

- Runs in single-threaded mode.
- Requires highest priority access to the data.
- Generates significant workload on the NAS array.

Nearly three decades after NDMP was created, even with today's more powerful NAS systems, users may still see performance issues during backup cycles, which affect their ability to read or write to the target storage, and which in turn can impact production workloads and key business services.

Limitations of Tape Backups

Tape backup requires complex infrastructure of hardware and software to operationalize, such as backup servers, tape silos, enterprise backup software, and backup tapes. All of these components add up to significant data-center space consumption, ongoing maintenance costs, and operational overhead for your organization.

- **Management Complexity.** Tape-based backups require two separate tape sets: one set of tapes for the actual backup job, and another set for the accompanying backup catalog. In larger environments, when both sets are factored in, a single full backup job may require hundreds of tapes.

  Additionally, the sheer complexity of every backup operation requires careful handling procedures to ensure the integrity of the backup data. Just one lost or damaged tape among those hundreds of tapes can ruin an entire backup set.

- **Cost.** While it's often assumed that "tape is cheap", the number of tapes for a full week's backup can quickly add up. At an aggregate level, assuming one full backup per month, a monthly change rate of 8% and a one-year retention schedule, a single petabyte of primary data will consume over 24PB of tape over the course of that year.

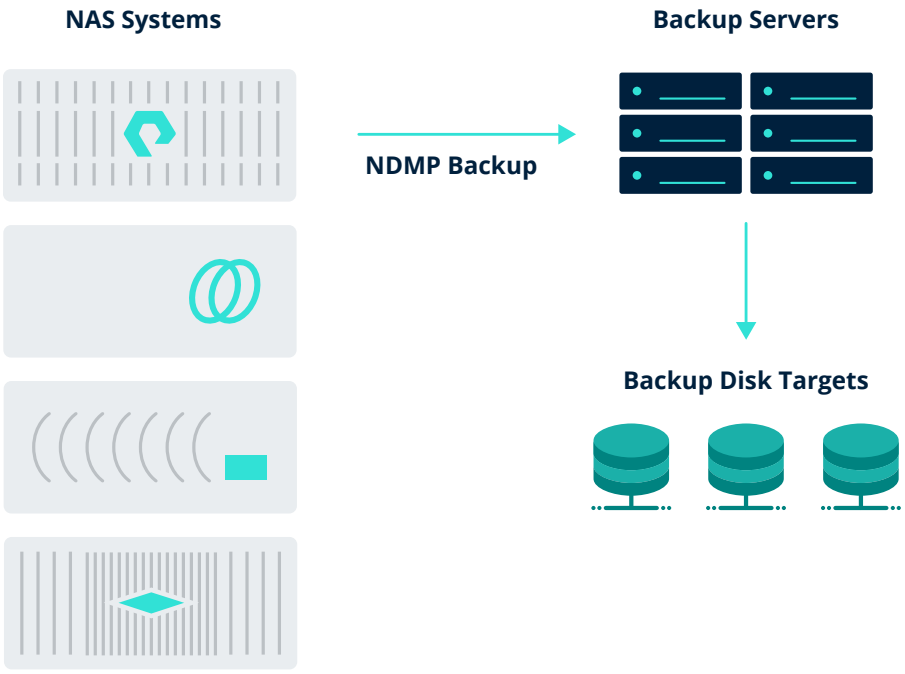
  If a second backup copy of that same dataset is required for offsite storage – a common industry practice – then the number of tapes required per year is doubled. That means IT will need to budget 48PB of tape capacity, per year, for every 1PB of primary data.

  Offsite tape rotation can drive costs up even further since it is usually outsourced to a third party. When all these factors are aggregated together – the combined cost of tape capacity, infrastructure and floor space, licensing, and staff – offsite service contracts can equal or even exceed the per-terabyte cost of primary-tier disk capacity.

NDMP-to-Disk Backups

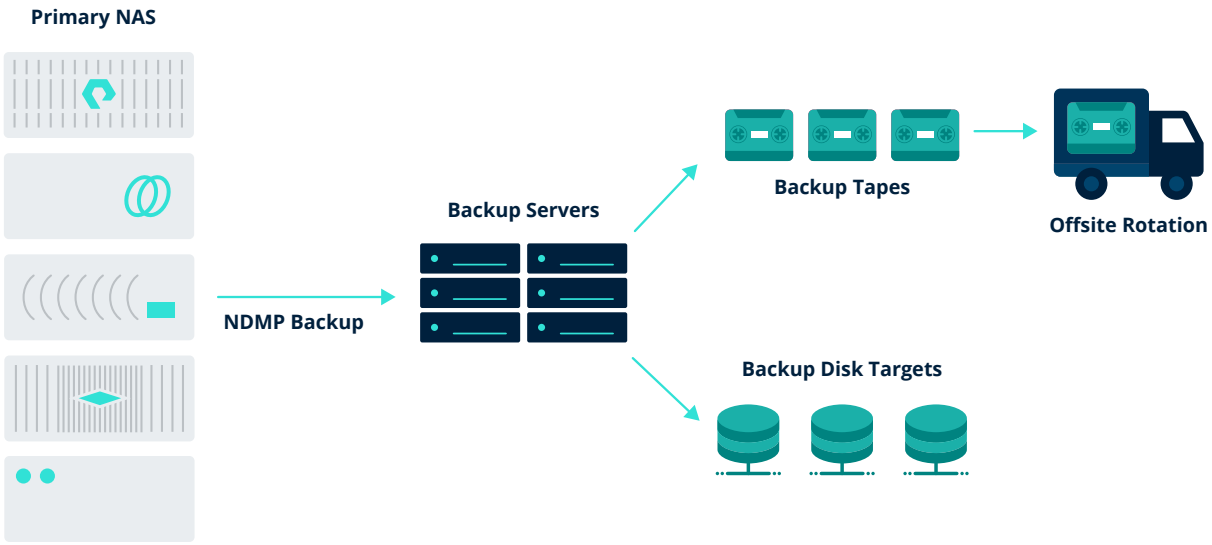
To address some of the cost and complexity factors of tape-based data protection, NDMP solution vendors have added support for disk-based backups, enabling some enterprises to replace their tape infrastructure. This change may have simplified backup operations, but it didn't address the core limitations of NDMP itself. Since it runs in single-threaded mode and requires the highest-priority access to NAS system resources, your production services can be potentially affected during backup windows.

Additionally, the move away from tape storage eliminated the ability to rotate backup sets offsite.



Hybrid NDMP

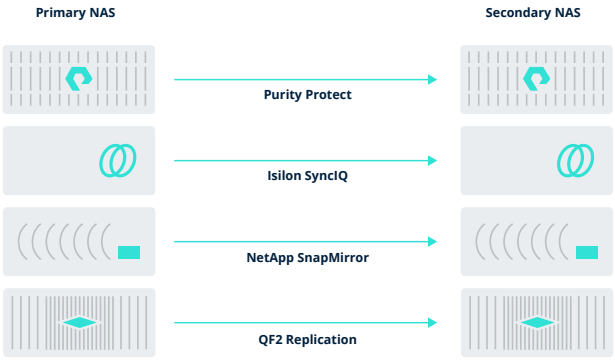
Enterprises who need the added protection of offsite storage have adopted a hybrid strategy, in which their most critical data is backed up via NDMP to tape, and the remaining data is streamed via NDMP to disk. The only real benefit that this solution offers is a slight reduction in overall operational complexity, stemming from the reduction in total number of tapes handled per day.



Disk-to-Disk Backup

Enterprises looking to move away from the constraints of NDMP have adopted disk-to-disk (D2D) backup instead. Originally engineered as a disaster-recovery (DR) replication solution, D2D replication has been adapted by the major NAS vendors into a backup platform.

- NetApp offers SnapMirror™ or SnapVault™ for D2D replication.
- Dell EMC Isilon uses SyncIQ™.
- Pure Storage offers Purity Protect™.
- Qumulo QF2 replication tools are included as part of its core storage platform.



Each vendor's D2D software works only with that vendor's hardware, meaning that D2D offers high-performance backup, but none of the interoperability of NDMP.

Disk-to-Disk Limitations

D2D's original purpose – disaster recovery – was to replicate new and changed data to a standby site. Data that changed on the primary storage was replicated to the target array, overwriting previous versions. For planning purposes in DR use cases, capacity and performance management are simple: just use the same platform and configuration on both sides of the replication pair.

Adapting D2D into a backup platform complicates both the replication process and capacity planning. The 1:1 relationship of primary-to-secondary capacity isn't applicable to backup use cases, since backup requirements mean that the secondary storage array needs to have enough disk capacity to store multiple versions of files as they change. Any cost savings realized by using cheaper, slower disk in the secondary array may be canceled out by the amount of capacity needed to satisfy corporate retention policies.

Management Complexity.

Enterprises with multiple primary NAS systems – both single and multi-vendor environments – quickly find that D2D as a backup solution increases the complexity of their operating model while increasing administrative overhead.

Vendor Lock-In.

For all its inherent limitations, NDMP offers near-universal compatibility (Qumulo QF2 does not offer NDMP-based backups). D2D backups, on the other hand, are vendor-specific. Each D2D replication pairing means another instance of vendor lock-in. SyncIQ will not replicate from Dell EMC Isilon storage to NetApp FAS, and Qumulo QF2 will not work with an Isilon target.

This particularly complicates operations in multi-vendor enterprises. In a heterogeneous environment, D2D requires separate hardware, software, and support for every replicated pair of NAS systems.

Resource Management.

When D2D is used for backup, the secondary storage system needs enough capacity to host multiple versions of the production data. Assuming the same 12-month retention requirement as with the above tape example, along with a change rate of 8% per month, every 1PB of primary data will require 2PB of capacity on the secondary array.

This means that an IT storage administrator looking to budget for production capacity and data protection will need to plan for and purchase three petabytes of disk capacity for every petabyte of primary data. And unless the per terabyte cost of secondary storage is less than half that of tier-one capacity, D2D solution costs can be even higher than that of primary disk.

Disk cost is only part of the challenge that enterprises need to plan for. In addition to planning and budgeting for the necessary storage space, IT must also:

- Track and manage resource utilization on both source and target arrays to protect production availability and defined replication objectives.
- Monitor network bandwidth and latency between source and target arrays.
- Configure, implement, and monitor all necessary replication jobs for success.
- Identify and resolve snapshot and replication failures as they occur.

With D2D backups, every petabyte of data requires two or more petabytes of secondary capacity to protect it.

In a large, multi-vendor environment, with multiple replication relationships from each vendor, D2D can quickly become operationally unsustainable under the weight of its own complexity.

Net Costs.

As a backup solution, D2D does not scale. Every replication relationship must be configured, managed, and monitored individually. In multi-vendor environments, IT must also:

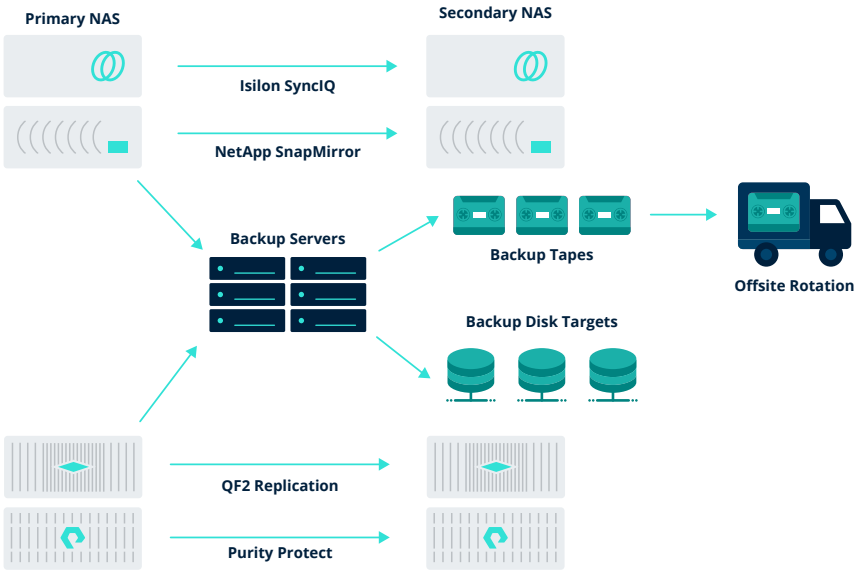
- Maintain sufficient storage capacity well ahead of need, factoring in each vendor's unique purchasing processes and turnaround times.
- Ensure licensing compliance for all NAS vendors and replication pairs
- Maintain administrative expertise and support for every managed platform

Additionally, three of the four major NAS vendors license their D2D software separately from their storage capacity, with the total licensing cost tied to the amount of data being replicated.

When all these factors come into play – the cost of secondary storage capacity, D2D licensing and support fees, the required rack and floor space in the data center, and employee staffing – IT may find that the cost-per-terabyte of primary-tier storage is tripled by the additional costs necessary to protect it.

Backups using NDMP and D2D

To work around the inherent limits of NDMP and D2D, enterprises may use a hybrid configuration in which some unstructured data is replicated to secondary storage using D2D, some data is streamed via NDMP to disk for performance, and critical data is still backed up to tape so it can be rotated offsite.



For many enterprises, this is the current state of their operations: a complex, multi-source, multi-target backup platform that leverages both D2D replication and tape-based solutions, with all the disadvantages and costs associated with each.

In the modern data-center environment, backup as a core function often provides some of the lowest return on investment of any IT operation: consuming data center space for backup infrastructure, consuming NAS resources that could be better used on production workloads, and requiring significant ongoing IT budget commitments for infrastructure, software, and staff.

Data protection is often the most operationally complex service that IT provides, yet it usually offers the lowest return on investment.

Backup to Cloud

As data centers run out of available space for new equipment, and as storage capacity and licensing costs continue to accrue in response to ever-increasing unstructured data footprints, enterprises are increasingly turning to cloud service providers, such as Amazon Web Services, Google Cloud Platform, or Microsoft Azure, for overflow backup and/or archive storage capacity.

This offers a significant amount of flexibility relative to earlier data-protection solutions. Cloud storage is available to anybody with an internet connection, and offers virtually unlimited capacity for even the largest enterprises. Before adopting cloud-hosted backup or archive services, however, there are a number of considerations that need to be addressed.

Cloud Costs

Any organization looking to leverage cloud storage for either backup or archive needs to factor in the costs associated with using any public-cloud platform before adopting a cloud-based storage strategy, since every component of their data-protection and archive strategy – including the cloud provider, storage tier, access frequency, and upload/download solution – will have a direct bearing on the overall cost they will need to absorb.

- **Consumption and Transaction Fees.** All cloud resource usage is metered, meaning that, in addition to the cost of the capacity used, customers also pay an ingress charge for every uploaded file or object, and an egress charge for every file or object downloaded. "Hot" storage tiers generally offer lower per-transaction ingress/egress fees, but significantly higher costs per terabyte of storage consumed, while "cold" tiers (e.g. Amazon's Deep Glacier service) offer very low consumption costs but much higher ingress/egress charges per transaction.

  A large-scale enterprise looking to host a one-billion file backup on cloud storage must factor in not just the cost of the capacity consumed, but the per-upload fee for every one of those billion files. The per-transaction fee is very small – generally a few cents per hundred thousand transactions – but a single upload of a billion files will incur a one-time fee in the tens of thousands of dollars.

Additional access fees and consumption charges apply for every subsequent incremental backup job and restore request. Even as backup solutions have evolved over the years – from NDMP-to-tape, NDMP-to-disk, disk-to-disk – they have not kept up with the needs and operational practices of modern data-centric organizations. In an era of double-digit data growth, enterprises are finding it difficult to protect a day's worth of data in less than 24 hours. What's more, even the newest, fastest NAS systems can be impacted by the resource demands of NDMP and D2D backup tools.

And while public cloud providers can ease the pressure on overbuilt data centers and offer additional data protection options, none of the legacy backup solutions offer simple, economical access to cloud storage.

Data Management Requires Data Visibility

For organizations with massive-scale unstructured enterprise data, effective data management requires a comprehensive suite of solutions that can discover, recognize, and track files and datasets throughout your environment. You need to be able to move data across storage systems, tiers, and sites, while protecting data against loss in the process.

None of these data-management objectives can be effectively met without the ability to see and understand their entire unstructured data environment, including answering the following questions:

- How much data does the enterprise own?
- Where is the data stored?
- How is it used?
- Who uses it?

For effective stewardship of both infrastructure and data, IT needs to have insight into the depth and breadth of your organization's unstructured datasets and the workflows that use them. If the full suite of an enterprise's unstructured data can't be fully indexed and analyzed, then it can't be properly managed.

The challenge that enterprises and organizations must contend with is the scale of their unstructured data. While IT may have spent significant time and resources in pursuit of effective data visibility, the available tools either don't work at scale, don't cover the full scope of the unstructured data portfolio, or don't provide outputs that lead to actionable insights.

Even with enormouse infrastructure and administrative investment, most data-discovery tools are still inadequite to the task.

At an enterprise level, data visibility specifically requires the ability to:

- Scan at scale: discover, collect and analyze metadata for all unstructured data across the entire organization, then collate it into a single, centralized index.
- View at scale: analyze up-to-date, interactive insights to develop a holistic view of the entire hierarchy of unstructured data, and the ability to browse and search file-system tree structures for specific datasets that can be consolidated, protected, archived, replicated, or moved.
- Search at scale: query the indexed metadata for deep data analysis across the enterprise.

When an organization's data footprint scales to petabytes or more, full visibility becomes exponentially more critical for effective management. At the same time, however, it becomes exponentially more difficult, particularly in multi-site, multi-vendor environments where unstructured data can span across multiple locations, platforms, and protocols. At this scale of operations, legacy data-discovery tools lose their effectiveness. Standard Linux admin utilities like grep and find, provide only limited visibility even in small-scale environments. In today's enterprises, any insight they can deliver is out-of-date by the time the outputs are available, since a full, end-to-end scan can take weeks or even months to complete.

Data Visibility Tools

Even with data visibility and data analytics at the file, folder, and NAS levels as critical as they are, there are only a limited number of additional tools that can deliver some of the insights needed for data at scale.

Vendor-Specific Storage Analytics

Most major NAS vendors offer proprietary data analytics tools for their own platforms, either as a modular add-on feature (e.g., Dell EMC Isilon InsightIQ™ and NetApp OnCommand Insight™), or integrated directly into the storage operating system (as with Qumulo QF2™), all of which are platform specific, but all of which come with their own inherent limitations.

Even within the context of their own platform, these tools may offer limited functionality, e.g. delivering only high-level information rather than in-depth file analysis, or providing only per-system visibility, i.e., analytics from individual systems can't be aggregated or centralized, even when all NAS systems come from the same vendor. None of these utilities offer comprehensive platform analytics that span the entire unstructured data environment.

Open Source and Third-Party Utilities

As with other components of a holistic data-management strategy (backup, archive, migration, and replication), there are dedicated open-source and third-party tools that attempt to meet this need. Third-party tools are generally licensed based on the organization's total managed-data footprint. At multi-petabyte scales, even with significant volume discounting, annual licensing costs alone can total millions of dollars and still account for only a portion of the total cost. As the company's data footprint continues to grow, high–performance compute and storage infrastructure, along with licensing, must be built out to keep pace.

Additionally, having originally been written for home directories and workgroup datasets, most such tools were never intended to scale to the size that modern enterprises need. In addition to their hefty infrastructure and licensing investment requirements, they may also be agent-driven, require a lot of administrative overhead (e.g. a manual mount or mapping for every in-scope export or share, repeated after every system restart or new export creation), and limited actionable output.

While open-source utilities may not incur the same licensing costs as proprietary solutions, customers quickly discover their limits when scaling up to the full scope of managed data. Inevitably, the scan rate falls behind the rate of data growth, and the index service's open-source database bogs down under even moderate-sized queries, quickly leaving enterprises with out-of-date information about their unstructured data.

Whether proprietary or open-source, most existing data-discovery tools offer limited (or nonexistent) analytics or movement capabilities. Once the full set of unstructured files, folders, and file systems has been initially inventoried, any potential actions based on analytical insights – e.g. data protection, archive, migration, replication, etc. – require more tools, more investment, and more administrative complexity.

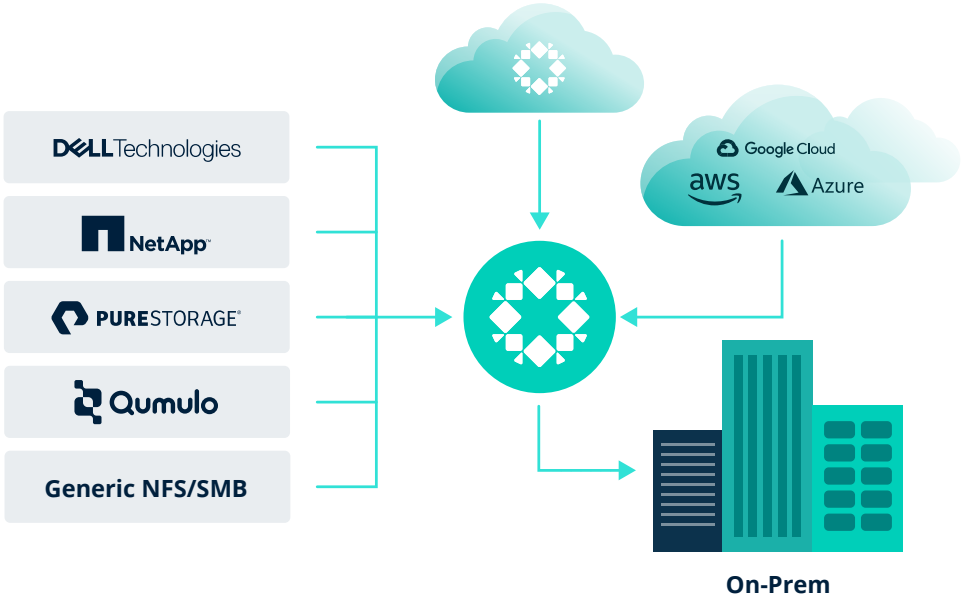
Open-source tools offer only limited visibility and limited functionality that quickly breaks down at scale.

How Rubrik Can Help

Enterprises who use Rubik have the flexibility to easily scale their backup capacity and service levels as their data needs grow.

Rubrik NAS Cloud Direct delivers a comprehensive data-management solution for unstructured data at scale. NAS Cloud Direct delivers high-performance, policy-based backup protection and data archive services, either on-premises or direct to any public cloud target, even at petabyte scales.

To enable data visibility and simplify the process of identifying and archiving cold data, Data Discover's powerful index-and-search engines enable users to see and analyze petabytes of unstructured data quickly and easily.



Purpose-built to index, protect, and archive billions of files and petabytes of data and offering API-level integration with all major NAS platforms, NAS Cloud Direct also provides native support for all public-cloud providers. Additionally, NAS Cloud Direct also offers as-a-Service simplicity and universal interoperability with all NAS architectures and protocols for frictionless data management.

To learn how Rubrik helps with unstructured data visibility, protection, and mobility at petabyte scale, read the Unstructured Data Management with NAS Cloud Direct white paper and visit www.rubrik.com/solutions/nas

rubrik

Global HQ
3495 Deer Creek Road
Palo Alto, CA 94304
United States

1-844-4RUBRIK
inquiries@rubrik.com
www.rubrik.com

20210929_v1