Igneous

WHITEPAPER

Igneous Unstructured Data Management

Part II: Unstructured data backup and archive with Igneous DataProtect

Abstract

This whitepaper is the second of a four-part series that outlines Igneous' solutions portfolio for managing massive amounts of unstructured enterprise data.

As a purpose-built management platform for enterprise unstructured data, Igneous DataProtect integrates at an API level with all network-attached storage systems and public- cloud platforms, delivering a high-performance, high-density, resilient data movement engine that expedites and simplifies the process of protecting and archiving unstructured data in multi-petabyte environments.

Igneous
2401 Fourth Ave., Suite 200
Seattle, WA 98121
(844) IGNEOUS
igneous.io

Table of Contents

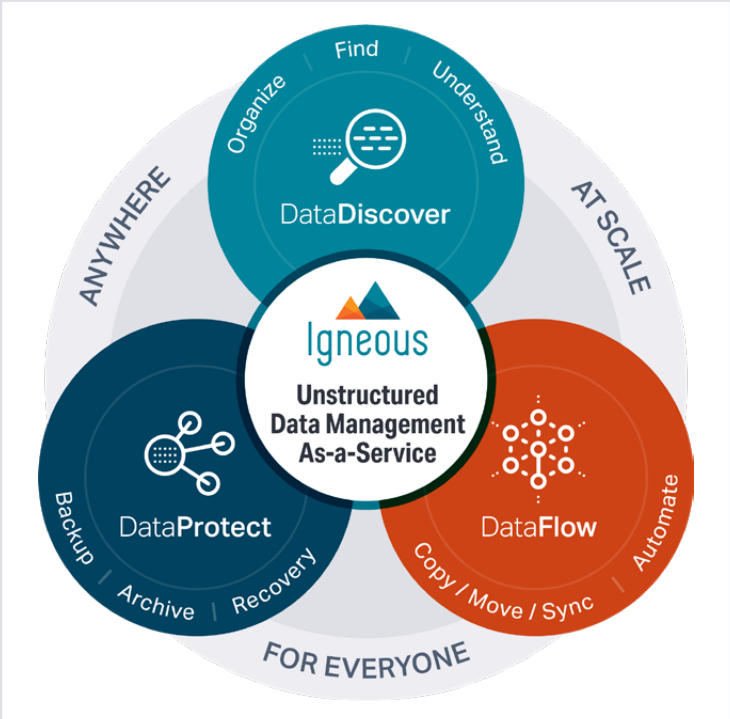
Introduction

Enterprise environments that include unstructured data – particularly unstructured data at petabyte-plus scales – often find that a legacy approach to data protection is inadequate for effective protection of their network-attached storage (NAS) platforms. Backup administrators may find that legacy tape and disk-based backup solutions are too slow to keep pace with the rate of data generation. Conversely, storage administrators may find that legacy backup solutions generate excessive overhead on production systems, causing production service slowdowns or even outages.

Document Purpose

This whitepaper is the third in a series of four documents that outline Igneous' data management capabilities. A critical component of unstructured data management in an enterprise environment is the ability to protect data against loss or corruption. This document shows how Igneous' approach resolves many of the challenges that IT faces in data backup and restore operations.

For additional context around the concepts of data discovery, data protection, and data flow, please refer to the other papers in this series, which are available for download on the Igneous Resources page (https://www.igneous.io/resources).



Data Protection

In order to safeguard against infrastructure failure, logical corruption, user error, or malicious actions, any organization's data management approach must include robust data protection as a critical component of its overall strategy. Data is the lifeblood of any organization, and data loss can lead to business disruption, compliance failures, and revenue loss.

To safeguard against data loss, enterprises use one or more backup platforms to protect their critical data. Organizations with smaller data footprints may find that a single, legacy backup solution – whether based on tape or disk – can protect their entire data portfolio.

In larger enterprises, however, even a combined approach has become insufficient to the core objective of backing up a day's worth of new data in less than 24 hours. Organizations are now finding themselves generating new unstructured data faster than they can protect it.

Backup

While backup as an objective can take any of a number of forms, a core set of distinguishing attributes of "backup" – as opposed to "archive" (addressed later in this paper) – includes the following:

- Creates a versioned replica of the production dataset
- Requires a dedicated data platform, separate from the primary data source
- Maintained for a set retention period before being allowed to expire
- Not used for any active workloads

High-Performance Backup with Igneous DataProtect

Enabling robust protection of unstructured data from any NAS platform, Igneous DataProtect delivers native support for NFS and SMB file-access protocols, providing full vendor independence and a single consolidated backup solution for all NAS technologies. With a purpose-built, highly parallelized data-movement engine optimized for file protection, an Igneous data management platform backs up unstructured data at line speeds. At the same time, Igneous monitors NAS performance for latency changes, adjusting its workload accordingly to protect production applications.

Igneous' high-density storage appliance absorbs petabytes of primary data easily while its data-mover and indexing engines can scan and move data from multiple NAS endpoints simultaneously at line speed.

Enterprises who use Igneous have the flexibility to easily scale their backup capacity and service levels as their data needs grow. The modular architecture of an Igneous platform enables simple expansion of each layer as needed – more data storage capacity, more compute for the indexing engine, or both.

Legacy Backup Solutions

Despite significant changes in data growth and usage, and the ongoing evolution of storage and network technologies, many enterprises still use a core backup solution that is based on 30-year-old concepts. Other organizations use proprietary, vendor-based utilities that offer limited flexibility and data protection, or a hybrid of both these approaches.

NDMP-Based Backups

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 adequate data protection – typically one full backup per week, with incremental backups nightly. To minimize the length of time needed for backups to complete, the backup industry developed the Network Data Management Protocol (NDMP), which standardized how backup software interacted with NAS arrays and tape systems.

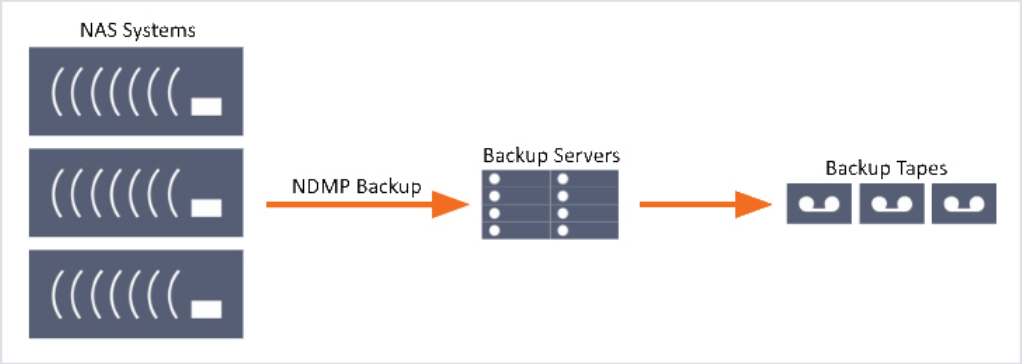
Limitations of NDMP Backups

Built on decades-old technology and practices, whether with tape or disk as the target, NDMP-based backup as a core data protection solution has reached the point at which its inherent constraints limit its usefulness in the modern enterprise.

NDMP runs in single-threaded mode; requires highest priority access to the data, and generates significant workload on the NAS array. Users may see performance issues during backup, which may affect their ability to read or write to the target storage.

Tape Backup

Tape backup requires many components to operationalize: backup servers, catalogues, tape drives, tape silos, backup management software, and backup tapes. All of these components add up to significant datacenter space consumption, ongoing maintenance costs, and operational overhead for the organization.



Although backup technology has evolved since the introduction of NDMP in the mid-1990s, many of the inherent limits in its original design specifications still persist, and continue to constrain backup throughput while consuming excessive resources on modern NAS systems.

Management Complexity

Even in smaller environments, a single backup job may require hundreds of tapes – one set of tapes for the actual backup job, and another set for the backup catalog. The sheer complexity of a single backup operation requires careful handling procedures to ensure the integrity of the backup data, since a single lost or damaged tape can ruin an entire backup set.

Cost

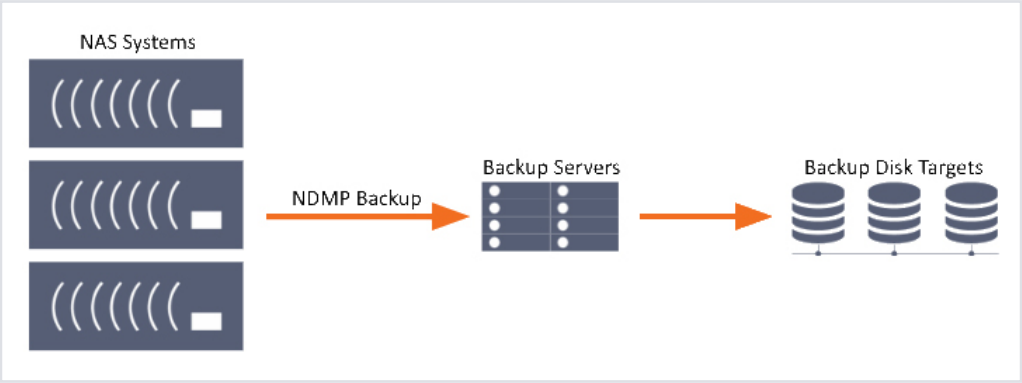
While it's often assumed that "tape is cheap," a one-week backup cycle at petabyte-plus scales with yearly retention can require hundreds of tapes. At an aggregate level, assuming one level-0 backup per month, a monthly change rate of 8% and a one-year retention schedule, a single petabyte of primary data requires over 24PB of tape. If a second backup copy is required for offsite storage – a common industry practice – then the number of tapes required per year is doubled.

In large environments, the scale and complexity of the tape-management process requires a tape silo, meaning more cost to IT – in hardware, in administration, and in data-center floor space. For enterprises that use a third-party service to rotate and store offsite backups, the cost goes up further, and may equal or even exceed the per-terabyte cost of primary-tier storage.

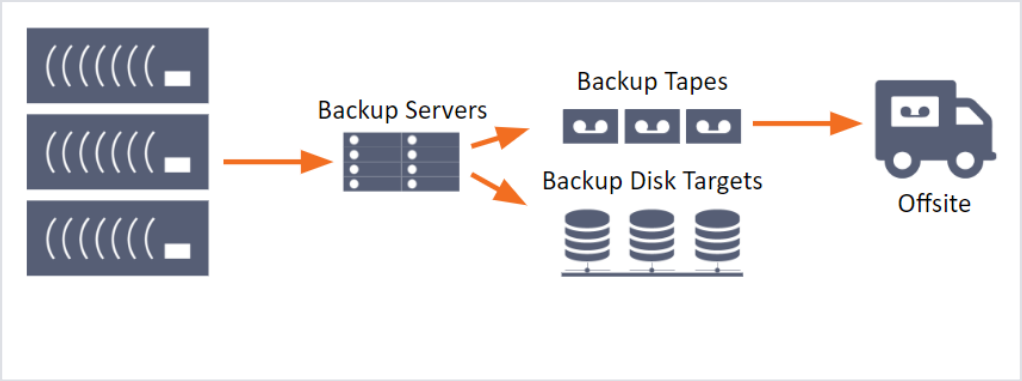
NDMP to Disk Backups

Once it became apparent that using tape as the primary backup medium was too complicated to meet most enterprises' daily backup throughput needs, and too costly relative to the value provided, backup vendors introduced disk-based storage that worked with NDMP backup solutions. This approach, still in widespread use today, pulls data from one or more NAS devices via NDMP and streams it straight to disk targets rather than tape.

This backup method is most common in enterprises consisting of mostly structured applications and some files.



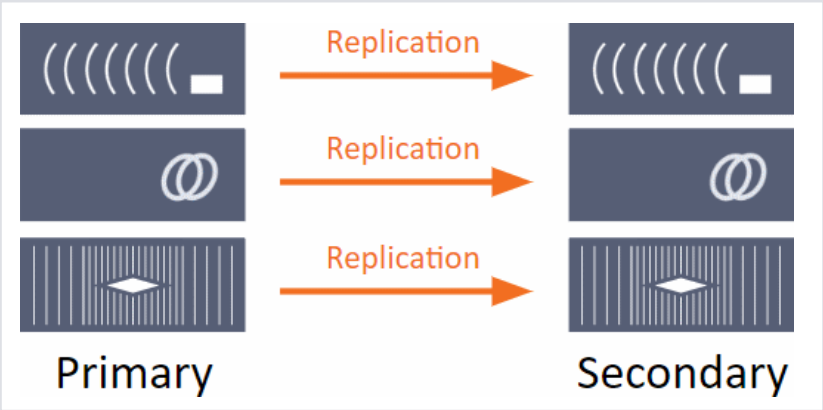
Overall backup throughput improved, simplifying operations and eliminating the need for expensive tape equipment, but the underlying limitations remain. NDMP still runs in single-threaded mode, continues to require highest-priority access to NAS system resources, and still risks affecting production services during backup windows.



Disk-to-Disk Backup

The limitations of NDMP-based backups in large environments have led IT to adopt disk-to-disk (D2D) replication, which was originally used for Disaster Recovery (DR), as a backup solution instead. Some modifications have been made to make the process compatible with backup requirements rather than DR objectives, but they also increase the complexity of D2D backup relative to its original use.

Each vendor sells a proprietary D2D utility, licensed separately from the cluster itself, and limited only to that vendor's architecture. Among the primary vendors, Qumulo QF2 includes replication tools within the core storage platform; NetApp offers SnapMirror™ or SnapVault™ for D2D replication, and Dell EMC Isilon uses SyncIQ™.



Disk-to-Disk Limitations

Using D2D as a backup platform requires a number of modifications from a standard DR configuration – the original purpose of D2D – which add significantly to the overall cost and complexity of the solution.

Disaster Recovery vs. Backup

A true DR platform replicates data to an identically configured system in a standby site, a standard requirement for IT shops looking to ensure that system performance will meet defined service-level objectives after a failover event. D2D works well as a DR solution since that's the specific use case it was intended to address.

As a backup solution, however, with additional requirements around capacity planning, storage performance, and management – particularly in enterprises with platforms from multiple NAS vendors – the overall cost and complexity of the resulting solution limit the effectiveness of D2D for many organizations.

Management Complexity

Enterprises with multiple NAS instances – both single- and multi-vendor environments – quickly find that D2D as a backup solution introduces a number of new challenges to their operating model, including vendor lock-in, capacity planning and management, and administrative overhead.

Vendor Lock-In

While most NAS vendors offer different storage tiers – e.g. low-capacity and high-performance, high-capacity and low-performance – to give IT some flexibility in hosting different data profiles, the use of D2D software still requires vendor lock-in: SyncIQ will not replicate from Dell EMC Isilon storage to NetApp, and Qumulo QF2 will not work with an Isilon target.

A backup strategy based on D2D replication, particularly in multi-vendor NAS environments, requires IT to plan, purchase, configure and manage a completely separate set of processes for every replicated pair of NAS systems.

Resource Management

Whereas DR replication overwrites files on the target system with updated data, the target array in a D2D-for-backup pair must be configured to maintain version histories of the source file system with every change. This enables point-in-time recovery of files and datasets as needed, but it means that IT must ensure that the target array has sufficient storage capacity to meet the organization's defined recovery objectives.

Assuming a 12-month retention requirement, a growth rate of 4% per month and a change rate of 8% per month, every 1PB of primary data will require 4PB of capacity on the target array. For an IT storage administrator looking to budget for production capacity and data protection, every petabyte of primary data will require five petabytes of disk capacity.

In addition to planning and budgeting for the necessary storage space, IT must also:

- Monitor 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
- Monitor every configured replication job for success
- Identify and resolve replication failures as they occur

Even in enterprises that use NAS systems from a single vendor, D2D requires significant planning efforts and ongoing administrative overhead. In a multi-vendor environment, with multiple D2D relationships from each vendor, this model quickly becomes unsustainable.

Net Costs

To add even further to the overall cost of D2D-based data protection, most vendors license their D2D software separately from their storage capacity, and usually pegged to the amount of data being replicated.

As a backup solution, D2D does not scale; every replication relationship must be configured and managed individually, even in single-vendor environments. In multi-vendor environments, IT must:

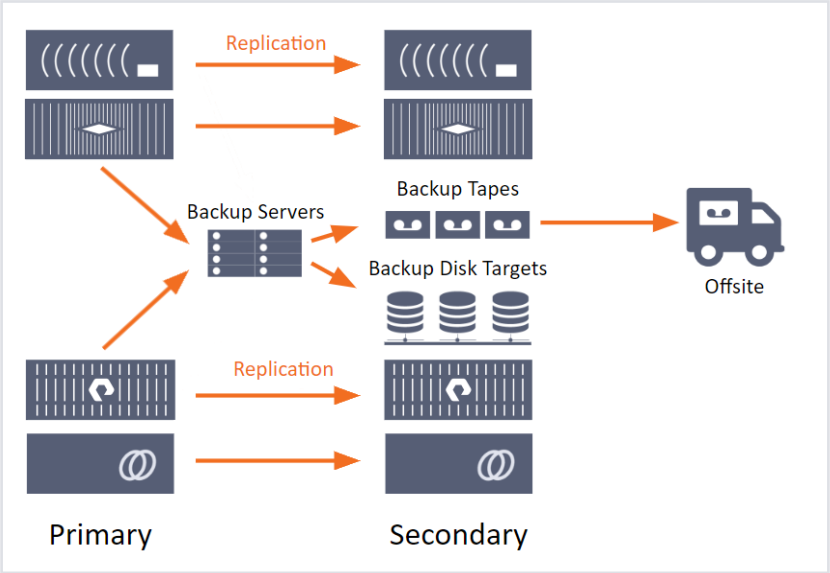
- Maintain sufficient storage capacity well ahead of need
- Ensure licensing compliance for all NAS vendors and replication pairs
- Maintain administrative expertise and support for every managed platform

When all these factors come into play – the cost of additional storage capacity, the D2D licensing and support fees, the required rack and floor space in the data center, and support staffing – IT may find that the cost-per-terabyte of simply hosting unstructured data may be quadrupled by the additional costs necessary to protect it.

Backups using NDMP and D2D

To overcome some of the NDMP and D2D data protection limits, and minimize the total cost of enterprise-wide data protection by using different strategies for different platform and data types, some enterprises use a hybrid configuration, with some unstructured data replicated using D2D, some data streamed via NDMP to disk for performance, and critical data backed up to tape for offsite storage.

For many organizations, 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.



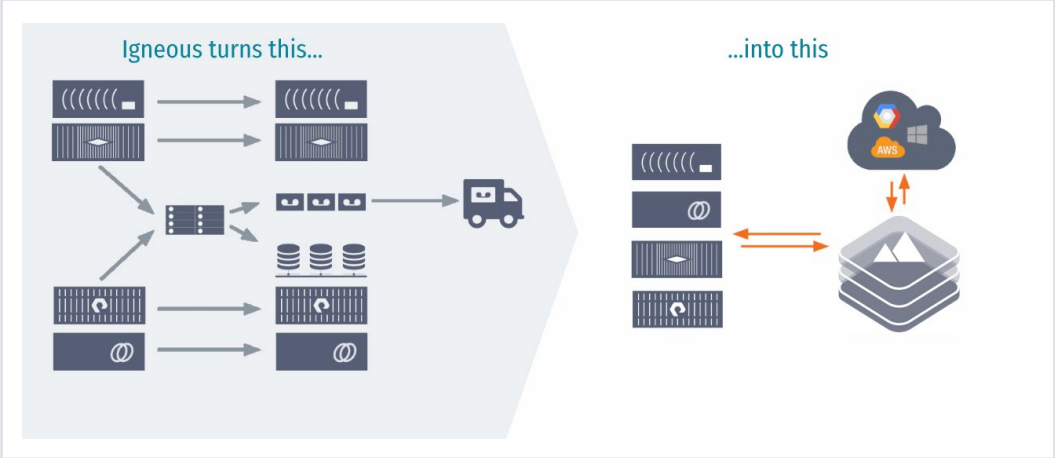
This multi-pronged approach to daily backup operations introduces significant attendant costs for every new terabyte of unstructured data capacity. Operationally, this strategy requires significant IT administrative bandwidth to monitor, maintain and support; it also limits opportunities to economize operations at scale.

In the modern datacenter environment, backup as a core function provides some of the worst return on investment of any IT operation, consuming datacenter 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.

Consolidated Scale-Out Backups with Igneous

Igneous DataProtect provides an easy-to-use, comprehensive data protection platform, enabling integrated enterprise file backup for unstructured file data that is both highly efficient and highly scalable. Policy-driven tiering and file movement to public cloud, and/or to Igneous deployments in other locations, are optional services that enterprises can leverage to ensure their data is always in the right place at the right time.

With a data-protection solution from Igneous, the challenge of managing a portfolio of disparate elements – backup servers, software, D2D replication, tape archives – is eliminated.



Policy-Driven Data Protection and Movement

Igneous DataProtect manages backups via individual policies, each of which is configured with unique, customizable settings for backup frequency, retention, and replication.

Once the appropriate policies have been defined, protecting petabytes of data can be as easy as assigning a policy to an export or file system. Igneous DataProtect will automatically import all NFS exports and SMB shares from each NAS array, and begin backup operations using the parameters associated with the assigned policy.

Replication and Archive

For each backup policy, administrators control whether data resides locally only, or if it should be replicated to a second location, e.g. another Igneous platform or cloud tier for long-term storage.

Faster Backups and Faster Restores

Unlike traditional backup methods, Igneous moves data in highly parallel streams, optimized for massive unstructured data libraries from any NAS storage platform. At the same time, the latency-aware data mover ensures that backups occur at line speed without impacting production users or workloads.

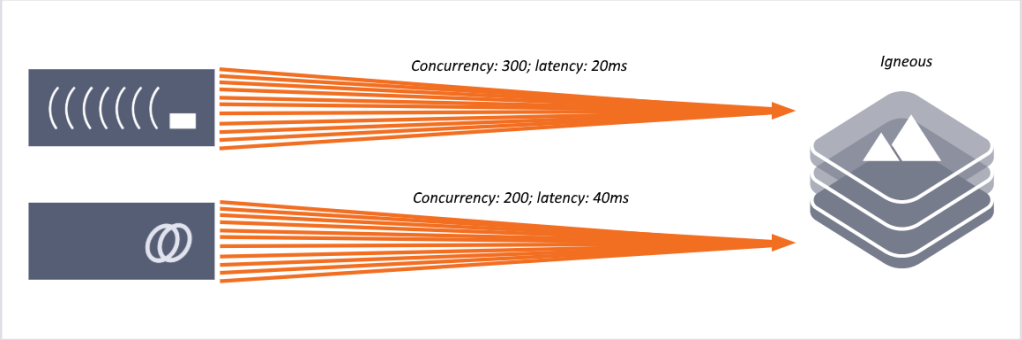
Optimized Data Movement

Once a file system or export has been added to the Igneous inventory, a multi-threaded crawler engine scans the entire directory tree, enumerating files by location, type, and size, and indexing their metadata. Workloads are broken up into separate thread pools appropriate to the data profile: large files are handled differently from small files, new files (which need to be scanned and indexed) differently from existing files (which need to be scanned, indexed, and compared with prior versions for potential changes).

After a job has started, and the index-scan-compare process has identified files for backup, additional thread pools begin optimizing the data for transfer. To minimize TCP overhead and network latency, small files are bundled into larger blocks for optimal throughput. Larger files are broken up into parallel copy streams to reduce the backup window even further.

Workload Protection

All data movement is highly parallel and latency-aware, ensuring fast data movement without disrupting business-critical applications that depend on the primary storage tier. If the data mover engine monitoring the primary NAS system identifies storage performance changes during the file-system crawl or backup operations, it automatically scales back the workload as appropriate to protect production applications and users.



After the initial full backup operation, subsequent jobs are incremental only, called "virtual fulls," which capture new and changed files and ensure even faster time to completion.

Streamlined Restore Process

The "virtual full" approach eliminates the time-consuming process of recovering a file incrementally, as from tape, or having only a single file version available for restore, as with D2D replication.

Industry-Leading Data Protection

As a comprehensive protection solution for unstructured data at massive scales, Igneous DataProtect offers significant advantages in the areas of backup administration, backup performance, and administrative simplicity, which set it apart in the industry.

Vendor Integration

Igneous is the only backup solution that includes full API integration with Dell EMC Isilon, NetApp, Pure Storage FlashBlade, and Qumulo QF2 storage. Using platform-specific programming commands, Igneous DataProtect simplifies backups by:

- Automating discovery of NAS exports
- Automating permissions provisioning for the entire export structure
- Monitoring overall system latency to ensure backups do not impact business-critical applications
- Managing snapshots for read-consistent backups
- Indexing and backing up both SMB and NFS file permissions for all storage types.

For all other NAS platforms, Igneous DataProtect still offers parallel, latency-aware backup capabilities and share/export discovery.

Unique Per-Vendor Platform Awareness

Igneous offers the industry's only multi-protocol support for Dell EMC Isilon, capturing both NFS and SMB permissions for the same data set, simplifying the backup process while also reducing backup time and the amount of storage capacity required.

For enterprises using Pure FlashBlade in their unstructured-data portfolio, Igneous is the only vendor to provide native support for backing up object storage as well as file data.

Organizations using Qumulo File Fabric will find that Igneous is the first data protection platform to offer native API integration. This enables Qumulo QF2 users to leverage backup as a data protection solution for the first time, with customizable retention policies and multi-version file protection rather than D2D-based replication.

Industry-Leading Backup Performance

Even the newer non-NDMP-based backup solutions, while better engineered to use file-access protocols that lower the burden on NAS resources, still require file-by-file comparisons that slow down the overall process of incremental backups, limit its usefulness in very large environments, and do not monitor the effect of backup jobs on production system performance.

With its highly parallel crawler threads that quickly identify new and changed file-system data, and a highly parallel data movement engine that packages small files in bulk and breaks large files up into multiple simultaneous move operations, Igneous DataProtect delivers industry-leading throughput that copies data at line speed while continuing to monitor and protect production system availability.

Archive Files with Igneous DataProtect

As data grows, particularly at petabyte-plus scales, IT must choose between continuing to expand its Tier-1 NAS footprint indefinitely, or migrating data off the primary tier at the same rate at which new data is created.

An effective archive strategy involves identifying data whose relevance to daily business operations has diminished to the point that it would be more cost effective to move it to an archive tier, designed for dense storage at a lower performance point, for long-term retention.

Archive versus Backup

Whereas "backup" as a term refers to a recurring operational cycle in which live data is copied to an alternate platform as protection from data loss, an "archive" operation moves an entire dataset from its original location – generally on higher-performance storage – to a platform optimized for dense storage and infrequent access.

Archive Objectives

Data moved to archive storage typically retains some value to the organization – whether for legal reasons, financial reasons, to preserve intellectual property, or for historical analytics – but needs only to be available for potential access, not for regular use. The retention time for archive data may vary depending on its usefulness: often 1-3 years for most data types, 3-7 years for some financial and legal data, or even indefinitely.

Regardless of the specific retention periods, tagging policies, or data types, an archive solution in a petabyte-scale enterprise must minimally include the following:

- Identifying data for archive based on defined criteria
- Data movement from primary to archive storage
- Auditable data movement trails that leave a record of what data was moved, to what location, based on what parameters
- Discoverable ("searchable") updates to enterprise index and search engines to ensure that data can always be located quickly

While there are a number of tools and utilities on the market capable of addressing the above requirements, they typically don't scale to the multi-petabyte level effectively. At petabyte-plus scales, there are very few consolidated archive solutions that can satisfy all of the above conditions. Organizations with large unstructured data portfolios may need to deploy a series of utilities to create a holistic, effective archive solution.

Challenges for Archiving Data

While each organization has its own criteria for determining which data is fit for archive, e.g., files older than a certain age, not accessed for a given time frame, or associated with defunct teams / projects – the task of identifying that data amid petabytes of active files presents a daunting challenge for file and storage administrators.

Vendor software is system-specific and does not work with other NAS systems; 3rd party software breaks down at scale: a 70PB UD footprint takes 3 months to scan. Data at that point is worthless.

Moving Data to Archive Storage

Once files and directories have been identified for archival, the next challenge is the process of migrating data sets from the primary NAS to the archive storage tier.

Platform-Specific Archive

NAS vendors offer tiered storage options, in which a single NAS array includes both high-performance disk and high-density, low-speed storage capacity. The data lifecycle model in use cases such as this typically involves data being generated and hosted on the high-performance tier initially, then migrating automatically to the archive-level storage as it ages out of active use.

Just as with other vendor-specific solutions, native storage utilities only work on the hosting NAS architecture. An organization looking to archive data from a mix of Dell EMC Isilon, NetApp, Pure and/or Qumulo must manage each platform's archive capacity and operations separately.

Licensing costs again factor into consideration here, since automated data tiering requires IT to purchase the additional functionality separately, per-platform and per-instance, from the archive-tier capacity.

Cloud Storage

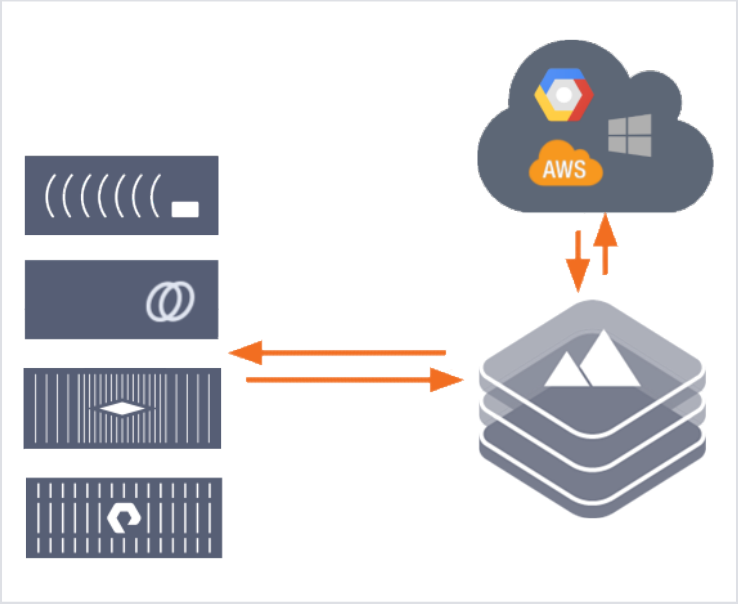
As enterprise data centers run out of available space for new equipment, and as storage capacity and licensing costs continue to accrue in response to ever-increasing unstructured data footprints, many organizations are turning to cloud service providers for additional compute and storage capacity, such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

For cloud storage to be a useful archive solution, however, IT needs a means for uploading data that is simple, automatic, and searchable after the fact for archive data that needs to be retrieved. Platform-specific solutions that integrate with cloud endpoints offer some of this functionality on a limited basis, but do not offer consolidated data archival from heterogenous NAS platforms, offer only partial search and retrieval capabilities, and may not be optimized for high-latency, low-bandwidth, wide-area network (WAN) transfers.

Simplified Archive with Igneous DataProtect

Designed to move data rapidly and efficiently for numerous use cases, Igneous' data movement engine moves data between platforms in multiple parallel streams that are sensitive to the impact of data movement on production workloads.

In addition to offering high-throughput data movement and dense storage at petabyte scales, Igneous DataProtect is storage-agnostic, centralizing archive data from every NAS instance in the enterprise. Organizations that deploy a mix of NAS architectures can use Igneous DataProtect to consolidate data from all these systems.



From there, data can be archived further, using policy-driven workflows that seamlessly move older files to any of the major cloud storage providers for longer-term retention. The enhanced integration that Igneous DataProtect delivers enables customers to automatically replicate or archive unstructured data directly to any of the following:

- Microsoft Azure's Hot, Cool, and Archive storage tiers
- Google Cloud Platform's Regional, Nearline, and Coldline storage tiers
- Amazon Web Services' S3 Standard, S3 Infrequent Access (IA), and Glacier storage tiers

Igneous also automatically updates its index and search engines as data is moved, meaning that files and directories can always be quickly and reliably found regardless of location. IT, data owners, and data users can always find their data sets quickly and simply, no matter where it is.

Conclusion

For nearly any IT enterprise, unstructured data as a source of administrative overhead and capacity consumption is a daily operational challenge. In order to support business-critical applications that depend on it, IT needs to ensure that petabytes of files are active and online at all times.

Most or all of that data must be backed up regularly and reliably. In many cases, that data also must be searchable: data owners and data users may need to quickly locate one file among hundreds of billions – either for recovery or for data analytics purposes.

Legacy backup models, involving NDMP-based backup to tape or disk-to-disk replication jobs, are insufficient to the full scope of the unstructured-data problem that IT needs to solve. Silo-based data storage limits IT agility, and primary-tier storage for second-tier data is not economically feasible over the long term.

Most legacy and platform-specific management utilities for replicating, archiving, moving, and indexing data are too limited in scope or functionality to be of use at an enterprise level, particularly multi-petabyte enterprise that uses heterogeneous NAS storage, spans multiple geographical sites, and leverages one or more public-cloud storage endpoints.

For nearly any medium-sized or large IT enterprise, unstructured data as a source of administrative overhead and capacity consumption is a daily operational challenge. In order to support business-critical applications that depend on it, IT needs to ensure that petabytes of files are active and online at all times. Most or all of that data must be backed up regularly and reliably.

Igneous DataProtect offers scale-out backup, indexing, and storage features, seamlessly connected via API-level integration to all the major NAS platforms, and protects of billions of files in very short timeframes. Its cloud-native architecture means resiliency at massive scales without sacrificing performance, and its as-a-Service implementation and support model means that IT can enable across-the-enterprise data management without sacrificing its own limited support bandwidth.

Contact Igneous

Igneous offers a modern, simple-at-scale architecture to:

- Effectively manage and scale growing unstructured data farms
- Eliminate backup windows and accelerate data restore operations
- Reduce the primary storage footprint by archiving data
- Expand access to data and services through platform-agnostic data movement
- Make all unstructured data easy to locate, track, and access
- Achieve cloud-level economics for secondary data
- Reduce management overhead so IT can be focus on strategic initiatives and operations

To learn more, please contact Igneous at info@igneous.io or **844-IGNEOUS**.

2401 Fourth Ave, Suite 200, Seattle, WA 98121, USA / 1-844-IGNEOUS / www.igneous.io