WHITEPAPER

Igneous Unstructured Data Management

Unstructured data backup and archive with Igneous DataProtect

Abstract

As a purpose-built management platform for enterprise unstructured data, Igneous DataProtect integrates with all NAS architectures and with every public-cloud platform to enable comprehensive backup and archive services at any scale. Igneous DataProtect overcomes the limitations of legacy data-protection solutions while expediting and simplifying the process of protecting and archiving unstructured data in even the largest environments.

This whitepaper is the second of a series that describes Igneous' solutions portfolio for simplifying the management of unstructured enterprise data at any scale.

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

Table of Contents

Introduction

In today's data-driven economy, nearly every industry – whether healthcare and biomedical research, earth sciences, electronic design, or media and entertainment – are finding that data is becoming increasingly critical to core operations. Statements such as "data is the new oil" or "data is the new gold" are commonly heard at trade shows and in the press.

The organizations that create, use, and profit from this data are faced with the challenges that accompany the opportunity. They're finding that double-digit annual growth rates, increasing operational complexity, and the evolving economics of cloud storage are driving operational complexity along with innovation.

The challenges of managing data during a period of compounding annual growth come from multiple angles. Limited data-center space means that data growth can't be accommodated simply by constantly adding new tier-one NAS capacity. As it ages out of active use, older data needs to move to secondary storage to free up primary-tier capacity. And at every stage in this cycle, all data needs to be regularly and reliably protected, sometimes daily.

New trends in data proliferation are breaking old data-protection models.

For years, even the most data-centric organizations were able to deliver comprehensive data protection using tape-based backup solutions. As new data – particularly new unstructured data (e.g. machine-generated files) – continued to scale up, the adoption of disk-based NDMP backups, followed by the use of disk-to-disk (D2D) replication, were able to continue providing effective data protection.

Now, however, the inherent limitations of legacy data-protection solutions are forcing data-centric organizations to reconsider their long-term viability. Enterprises that have passed the 100-200TB mark are finding that NDMP backup solutions have lost their ability to reliably protect unstructured data, and while D2D backups may provide adequate protection at that level, its overall cost makes D2D less attractive in the long term as well.

Document Purpose

This guide, the second in a series of white papers on unstructured data management, outlines the capabilities, simplicity, and overall value of Igneous DataProtect™ as a backup and archive solution, and as a critical component of Igneous' Unstructured Data Management portfolio. Designed and engineered specifically to overcome the limits of legacy data-protection solutions, and delivered as-a-Service for greater simplicity and reliability, DataProtect is "petabyte ready and terabyte friendly," providing effective backup and archive services at any scale.

For additional context around the concepts of unstructured data management, including data visibility and data flow, please refer to the other papers in this series, which are available for download here (https://www.igneous.io/resources).

Audience

This document is intended for any of the following stakeholders: IT executives, managers, and decision-makers, storage administrators, backup administrators; data scientists, and others concerned with or focused on managing unstructured data at scale.

Familiarity is assumed with the following:

- NAS services, concepts and architecture
- Unstructured data access protocols, including NFS, SMB, and S3/object
- Network concepts, including LAN, WAN, and Internet connectivity
- Public-cloud storage platforms and options from Amazon Web Services™ (AWS™), Google Cloud Storage™ (GCS™), and Microsoft Azure™ (Azure™)

Defining “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.

> *“Unstructured data” refers to machine-generated files and datasets, stored on enterprise NAS systems.*

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.

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 millions 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.

Protecting Critical Unstructured Data with Igneous

The value of any given dataset lies not just in the value it provides – e.g. revenue, research grants, patent creation, artificial intelligence/machine learning – for its parent organization, but in the cost of generating that data in the first place.

> *Loss of data can mean loss of revenue, impact to critical timelines, and legal or financial penalties.*

As a result, any loss of data means one of two things: the data must either be re-created (a process that costs the organization both money and time, and may not even be an option), or the data must be written off, leading to loss of revenue, business disruption, and/or legal risk.

Data-centric organizations whose unstructured data footprint is in the range of hundreds of terabytes (or beyond) often find that legacy backup solutions, such as NDMP and D2D are increasingly unable to deliver effective data protection. Complexity and cost may be too high relative to overall performance, or backup activities impact production workloads, or these solutions are unable to complete a full day’s backup in under 24 hours.

Igneous Protects Unstructured Data at Any Scale

To address the challenge of protecting massive amounts of unstructured data, Igneous engineered its Unstructured Data Management platform specifically to overcome the limitations of legacy backup solutions.

Designed for Enterprises of Every Size

Delivered as-a-Service, Igneous unstructured data management solutions enable data protection at scale by combining an unmatched scanning engine – capable of discovering and comparing hundreds of thousands of files per second – with a multithreaded data-movement engine that moves files simultaneously in parallel streams for maximum throughput.

Recognizing that larger datasets usually mean more complex environments, Igneous unstructured data management solutions – Igneous DataDiscover™ for data visibility, Igneous DataProtect for data backup, and Igneous DataFlow™ for programmatic data movement – were engineered to be fully compatible with any NAS platform, and can manage any type of NAS data, whether via NFS, SMB, or object storage protocols, on storage systems such as VAST Data™, WekaIO™ Matrix™, IBM™ Spectrum Scale, Quantum™ Stornext™, Panasas™ ActiveStor™, Hitachi™ HNAS™, and any other NAS platform.

> *Engineered for unstructured data at scale, DataProtect can discover and protect billions of files and terabytes of data per day.*

In addition to its universal support for any NAS architecture, Igneous services integrate at an API level with Pure™ FlashBlade™, Dell EMC™ Isilon™, NetApp™ Flexible Attached Storage™ (FAS™), and Qumulo™ File Fabric™ QF2™ – enabling the automation of export discovery for new systems, snapshot management for crash-consistent backups, path management for optimal throughput, and native multi-protocol support for NAS platforms, like Dell EMC Isilon and NetApp FAS, that support both NFS and SMB access to the same data.

The resulting solution offers true data protection at scale, moving up to 25,000 files per second (2.1 billion files per day). Even on the fastest local networks, DataProtect can move data at line speed – over 100TB per day on 10GbE networks.

Deployment Flexibility

Igneous' full suite of unstructured data management solutions offers a number of deployment options that can accommodate an organization's specific operational and environmental requirements, ranging from local-only hardware to a direct-to-cloud model driven entirely by virtual machines.

Customers who adopt DataProtect as their enterprise backup platform can streamline operations and lower expenses by taking advantage of Igneous' deployment flexibility, and retire their complicated legacy NDMP and/or D2D infrastructure.

Full Cloud Integration at Lower Cost

Organizations that leverage public-cloud storage – for added data protection, for long-term archive capacity, and for global data access – need a reliable, simple method of uploading their data.

DataProtect can move data to any tier of any public-cloud platform, using simple, straightforward policies that control replication and retention settings with a few clicks.

Additionally, the Igneous data-movement engine uses a proprietary process for reading and writing data that drastically reduces the number of operations to and from the public cloud. Since these individual transactions are all metered by the public-cloud service, DataProtect is able to provide cloud access for backup and archive services at a lower cost than other solutions.

> *DataProtect offers simple, flexible access to the public cloud at scale while lowering cloud access*

Protecting Data While Preserving Performance

Any Igneous activity – scanning and comparing files, moving data between systems, running backup tasks – includes a service that monitors latency times on the target system. Any detected increase in latency beyond expected levels will prompt the appropriate Igneous engine to scale back its resource usage in order to protect system performance. By prioritizing production workloads, Igneous enables 24x7 data protection, effectively eliminating the concept of the backup window.

Comprehensive Data Management

In addition to enabling easy, at-scale backup and recovery with DataProtect, Igneous also offers data index-and-search at scale through Igneous DataDiscover, as well as reliable, high-throughput data replication and movement to any endpoint (local NAS, remote NAS, any public-cloud location and tier) with Igneous DataFlow.

Scaling Data Protection to Match Data Volume

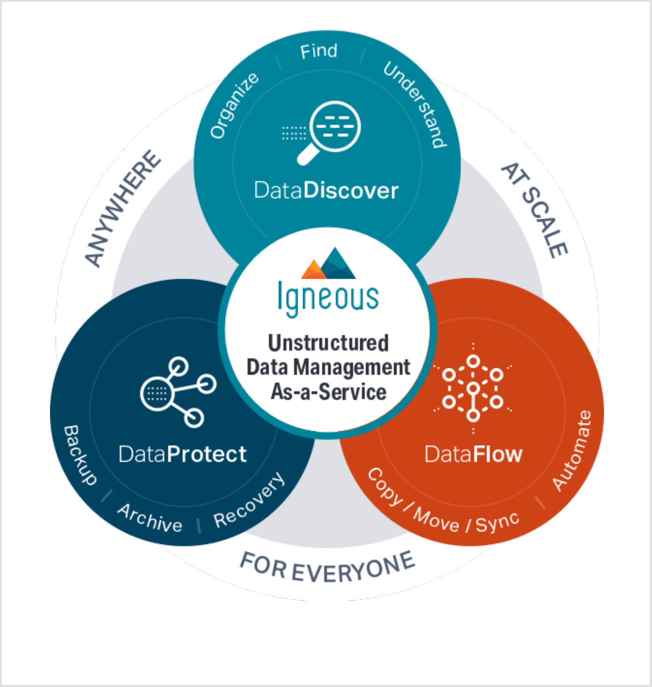
Delivered as-a-Service, with both physical and virtual deployment options available, Igneous' unstructured data management services offer the flexibility customers need for their specific environments.

A physical deployment starts with a 2U Application Service Router (ASR) component that provides scan, index, and data-movement services, and a 4U databox component that offers 426TB of storage capacity.

Customers with limited physical space for hardware can deploy a virtual-machine-based version that uses the customer's preferred public-cloud platform for all backup and archive storage capacity.

For increased scan and search performance, customers with physical deployments can add ASR components as needed to their local Igneous instance. Databox devices can be added to the local deployment for greater storage capacity and faster data throughput.

Virtual-only deployments can scale by adding more virtual machines to increase throughput or to provide protection for remote sites.



Challenges and Opportunities of Data Protection at Scale

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 (addressed later in this paper) – 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 it available for any active workloads

In the past, even large enterprises have been able to rely on NDMP and disk-to-disk (D2D) backup solutions. Today, however, those same enterprises' data portfolios have grown beyond what NDMP can effectively protect. D2D for daily backups offers larger-scale backup support, but can also be expensive and cause undue stress on production systems.

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

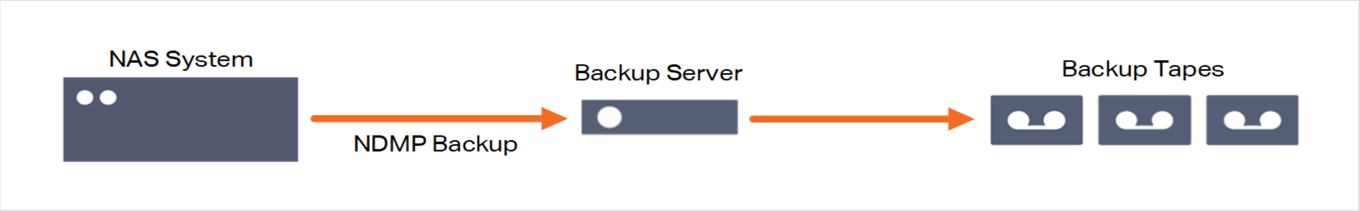
NDMP: An Old Solution for a New Problem

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 for years and enterprise operations have gotten more complex, with heterogeneous storage platforms, geographically dispersed business units, and the exponential proliferation of new data sources over the past two decades. Despite advances in NAS performance, and the decades-long proliferation of new data, tape backup technology has failed to evolve to accommodate these changes.

Limitations of NDMP-Based Backups

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.



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.

> *Designed for smaller datasets and simpler environments, NDMP's architecture limits its usefulness in large-scale enterprises.*

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: 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 the 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, since even one lost or damaged tape among those hundreds of tapes can ruin an entire backup set.

> *While tape backup is incorrectly thought of as inexpensive, infrastructure and operational factors together can drive the cost of tape backup to match or even exceed that of primary storage.*

Cost

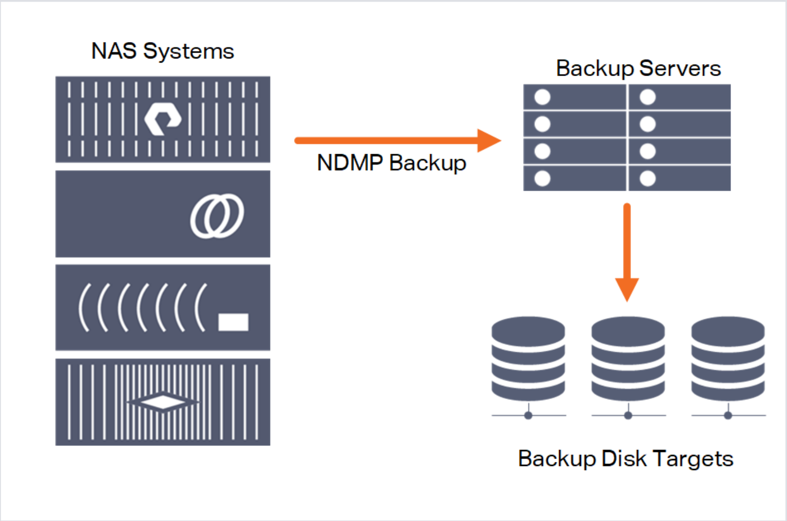
While it's often assumed that "tape is cheap", the number of tapes for a full week's backup can quickly add up. At an aggregate level, assuming one full backup per month, a monthly change rate of 8% and a one-year retention schedule, a single petabyte of primary data will consume over 24PB of tape over 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, meaning that IT needs to budget 48PB of tape capacity, per year, for every 1PB of primary data.

Offsite tape rotation services 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 salaries, and offsite service contracts can equal or even exceed the per-terabyte cost of primary-tier disk capacity.

NDMP-to-Disk Backups

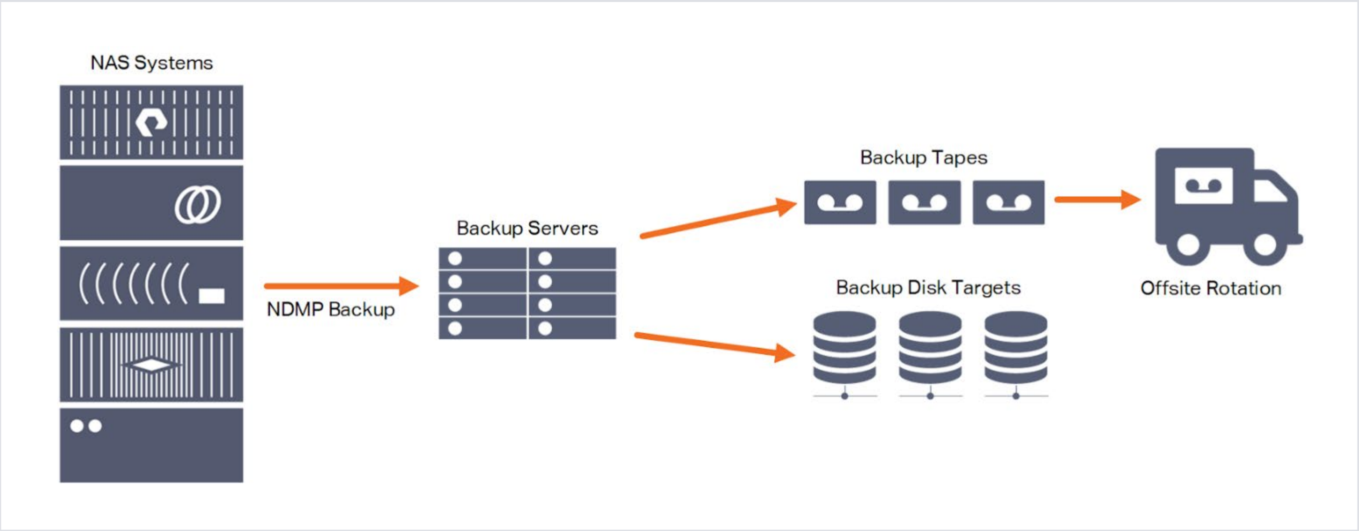
To address some of the cost and complexity factors of tape-based data protection, NDMP solution vendors have added support for disk-based backups, enabling some enterprises to replace their tape infrastructure. This change may have simplified backup operations, but it didn't address the core limitations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



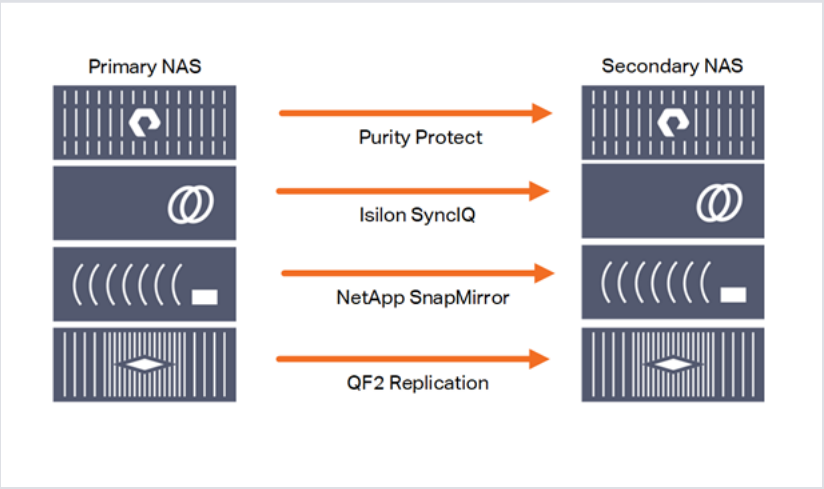
The only real benefit that this solution offers is a slight reduction in overall operational complexity, stemming from the reduction in total number of tapes handled per day.

Disk-to-Disk Backups

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™, and 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. Capacity and performance planning in a DR use case 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 doesn't apply 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.

Originally intended to simplify backup operations and improve performance, D2D complicates capacity management and deepens vendor lock-in.

Management Complexity

Additionally, enterprises with multiple primary NAS systems – both single- and multi-vendor environments – quickly find that D2D as a backup solution introduces factors and constraints that increase the complexity of their operating model, including vendor lock-in, capacity planning, performance management, and administrative overhead.

Vendor Lock-In

For all its inherent limitations, NDMP offers near-universal compatibility (Qumulo QF2 does not offer NDMP support). 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, 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.

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 snapshot and replication failures as they occur

In a large, multi-vendor environment, with multiple replication relationships from each vendor, D2D can quickly become operationally unsustainable.

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

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

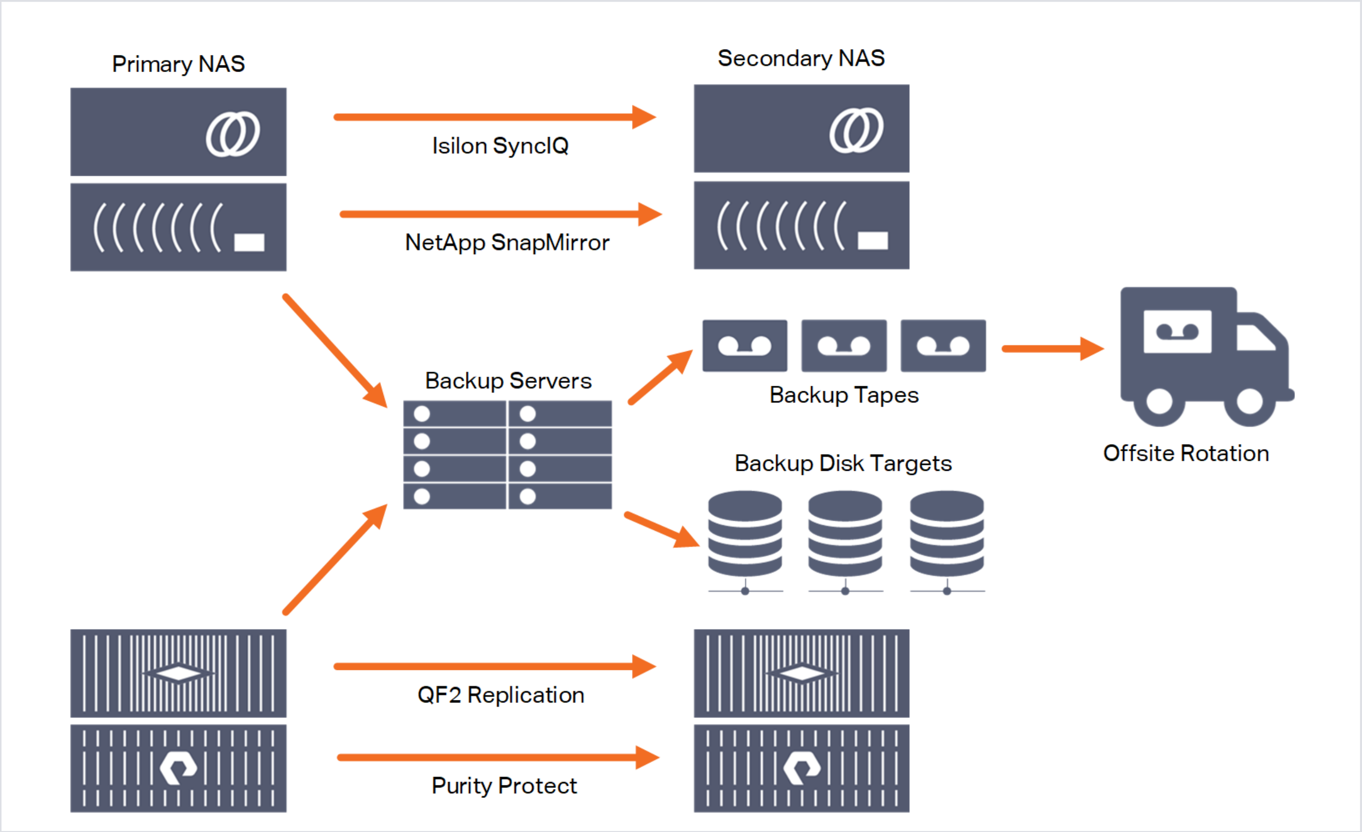
Additionally, three of the four major NAS vendors license their D2D software separately from their storage capacity, with the total licensing cost tied to the amount of data being replicated.

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.

D2D provides better overall throughput, but at a significantly higher cost-per-terabyte of protection than NDMP

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.



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

For those 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.

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.

Data-center space restrictions and NDMP/D2D costs are driving the adoption of cloud-based backup services, which offer flexibility and virtually unlimited capacity.

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 file or object they upload to cloud storage, and an egress charge for every file or object that they download. “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.

Cloud-based backup and restore services incur fees for every object uploaded or downloaded, as well as for actual space consumed.

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 one 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.

Data Protection at any Scale with Igneous

Explicitly engineered to overcome the challenges of legacy backup and archive solutions at any scale, Igneous DataProtect simplifies backup architecture and operations with an as-a-Service delivery and support model that is both highly efficient and highly scalable. Policy-driven tiering and file movement – whether to public cloud, to other NAS endpoints, or to Igneous systems in other locations – are additional, optional services that enterprises can leverage to ensure their data is always in the right place at the right time.

Simple-at-Scale Backup with Igneous DataProtect

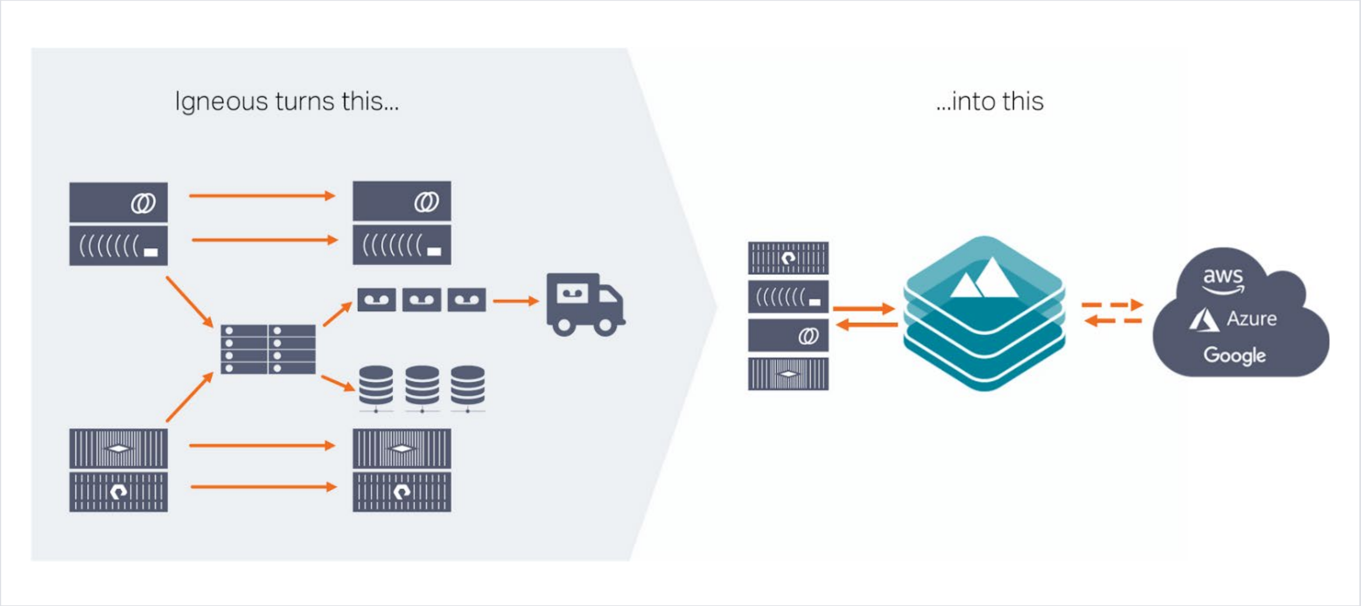
Igneous DataProtect delivers native support for NFS, SMB, and object protocols, as well as API-level integration with all major NAS platforms, providing full vendor independence and a single consolidated backup solution for all NAS technologies.

DataProtect's purpose-built, highly parallelized data-movement engine is optimized for file movement and works seamlessly with any NAS platform to back 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.

DataProtect delivers better interoperability than NDMP and better performance than D2D.

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 capacity for data storage, more compute for the indexing engine, or both. Virtual-only instances can scale by adding new virtual machines to the deployment.

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.



Faster Backups, Faster Restores, More Options

Unlike traditional, single-threaded backup methods, Igneous moves data in concurrent streams, automatically optimized for massive unstructured data libraries from any NAS storage platform.

Optimized Data Movement

When a backup job launches against a newly added dataset (e.g. directory, export, or file system), a multi-threaded crawler engine scans the entire directory tree, enumerating files by location, type, and size. For incremental backup runs, the crawler engine compares each file's metadata with the index from the previous backup, scanning and comparing the records from hundreds of thousands of files per second, to identify new and changed files.

DataProtect moves data in parallel streams, automatically optimized for data type, NAS type, and the available network bandwidth.

The data-movement engine then begins optimizing files for transfer. To normalize the data for maximum throughput, small files are bundled into larger blocks, while larger files are broken up into multiple smaller blocks. Once normalized into standard sizes, these blocks of data are then moved in parallel streams over the network to the backup target.

The number of separate data-movement streams varies with each job, and is determined by a combination of factors, including the throughput capabilities and real-time system latency on the source NAS, available network bandwidth during configured backup hours, and the number of active backup jobs.

A Shortcut to Cloud Integration

Especially at scale, the number of data-protection solutions that include native support for cloud-based backup storage is limited. Enterprises who have wanted to expand their options to include public-cloud storage have in some cases hesitated or delayed implementation due to the difficulty of incorporating data movement to and from cloud endpoints into their normal operations.

Any Cloud, Any Tier

For DataProtect, which was engineered for parallel data movement at scale across any network link, and which includes full support for all major public-cloud providers, access to any tier of any cloud platform is easily enabled and implemented using simple data-management policies to control both replication and retention from a single management portal.

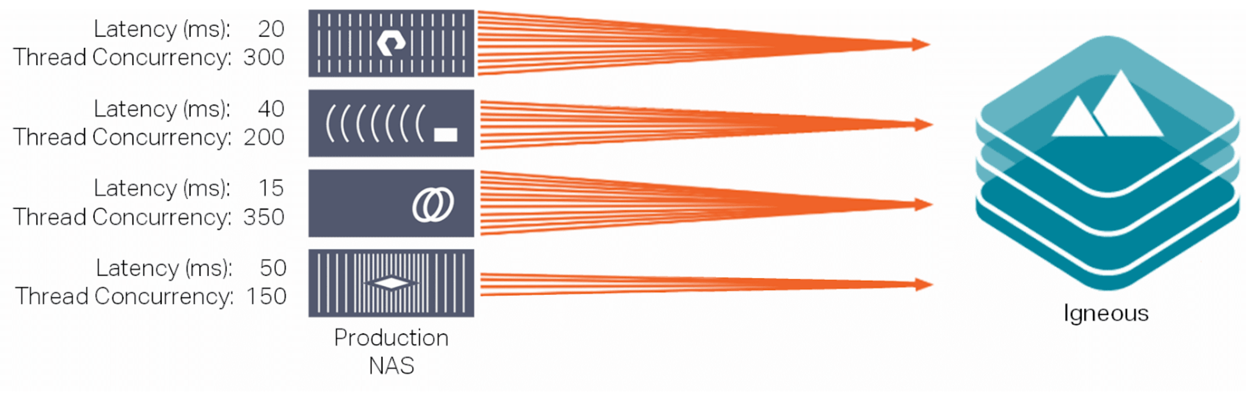
Efficient Data Movement

For cloud-based data movement, DataProtect aggregates files into roughly equal-sized "blobs" of data, each containing a bundle of hundreds, even thousands (depending on average file size), of individual files, before writing data to any public-cloud target. With every individual transaction being metered and charged by the cloud provider, this feature alone can lower the cost of cloud ingress and egress operations by over 90%, making cloud-based backup and restore operations much more cost-effective for Igneous customers.

> *Igneous' unique data-movement architecture and universal cloud compatibility maximize customer flexibility and speed, while minimizing cloud costs.*

Workload Protection

The file-scanning and data-movement engines that power DataProtect are also latency-aware, ensuring fast data transfers without disrupting business-critical services, applications, and workflows that depend on the primary storage tier. If the Data Mover engine identifies storage performance changes on the primary NAS system during the file-system crawl or backup operations, it automatically scales back the workload as appropriate to protect production applications and users.

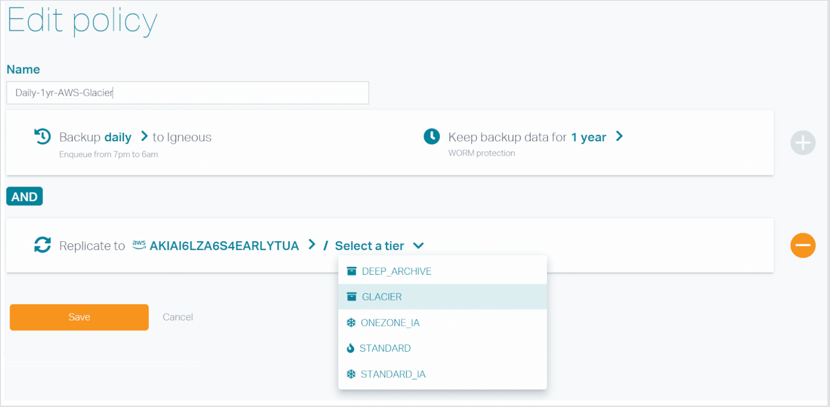


DataProtect can run 24x7without impacting system availability, effectively eliminating the backup window from operational consideration.

Since DataProtect is engineered to automatically move data in concurrent streams, and configured to optimize throughput while adjusting system resource consumption in response to overall NAS latency, multiple backup jobs can run at any given time. This means that backup runs can occur at all hours – enabling 24x7 data protection, providing protection at scale, and eliminating the entire concept of the backup window.

Policy-Driven Data Protection and Movement

DataProtect manages all backup, archive, and replication settings through flexible, intuitive policies that let customers configure specific parameters for backup frequency, backup target (e.g. onsite Igneous or directly to cloud storage), along with optional replication to a public-cloud endpoint and tier. Once the necessary policies have been defined, protecting unstructured data at any scale – whether it's in a 100TB environment or a multi-petabyte enterprise – is as easy as assigning the appropriate policy to an export or file system.



Edit policy

Name: Daily-1yr-AWS-Glacier

Backup **daily** > to Igneous
Enqueue from 7pm to 6am

Keep backup data for **1 year** >
WORM protection

AND

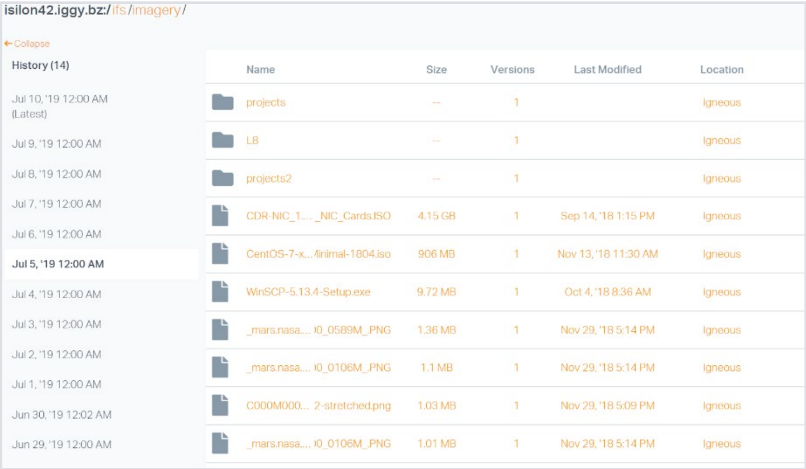
Replicate to aws AKIAI6LZA6S4EARLYTUA > / Select a tier

Since different datasets will often have unique protection requirements, a backup policy can also be assigned to individual directories where required.

Policy Flexibility

The same policy can be assigned to any target, on any file system, using any protocol. DataProtect will automatically detect the correct access settings for the target and begin backup operations using the parameters associated with the assigned policy.

Streamlined Restore Process



isilon42.iggy.bz:/ifs/imagery/

History (14)

Jul 10, '19 12:00 AM (Latest)
Jul 9, '19 12:00 AM
Jul 8, '19 12:00 AM
Jul 7, '19 12:00 AM
Jul 6, '19 12:00 AM
Jul 5, '19 12:00 AM
Jul 4, '19 12:00 AM
Jul 3, '19 12:00 AM
Jul 2, '19 12:00 AM
Jul 1, '19 12:00 AM
Jun 30, '19 12:02 AM
Jun 29, '19 12:00 AM

Name	Size	Versions	Last Modified	Location
projects	--	1		Igneous
L8	--	1		Igneous
projects2	--	1		Igneous
CDR-NIC_1... _NIC_Cards.ISO	4.15 GB	1	Sep 14, '18 1:15 PM	Igneous
CentOS-7-x...Minimal-1804.iso	906 MB	1	Nov 13, '18 11:30 AM	Igneous
WinSCP-5.13.4-Setup.exe	9.72 MB	1	Oct 4, '18 8:36 AM	Igneous
_mars.nasa.... 0_0589M_PNG	1.36 MB	1	Nov 29, '18 5:14 PM	Igneous
_mars.nasa.... 0_0106M_PNG	1.1 MB	1	Nov 29, '18 5:14 PM	Igneous
C000M000... 2-stretched.png	1.03 MB	1	Nov 29, '18 5:09 PM	Igneous
_mars.nasa.... 0_0106M_PNG	1.01 MB	1	Nov 29, '18 5:14 PM	Igneous

To simplify the process of data recovery, DataProtect presents a "virtual full" view of the dataset for every date on which a backup run was completed. Restoring data – whether the full dataset or a single file – is a simple matter of choosing the appropriate recovery-point date and selecting a destination directory.

Data can be quickly and simply recovered, without the time-consuming juggling of NDMP catalog and backup-set tapes, or the highly complex process of managing D2D recovery snapshots.

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, DataProtect simplifies backups by:

- Automating discovery of NFS exports and SMB shares on all NAS systems
- Automating permissions provisioning for the entire export structure where appropriate
- Managing network paths for optimal throughput
- Managing snapshots for read-consistent backups
- Monitoring overall system latency to ensure backups do not impact business-critical applications
- Indexing and backing up both SMB and NFS file permissions for NetApp FAS and Dell EMC Isilon systems

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

> *DataProtect is the only backup solution specifically engineered for API-level integration with all four major NAS platforms.*

Unique Per-Vendor Platform Awareness

Igneous offers the industry's only multi-protocol support for Dell EMC Isilon and NetApp FAS, protecting both NFS and SMB permissions for the same dataset, 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, in addition to supporting backup of both NFS and SMB datasets, Igneous is the only vendor to provide native support for backing up object storage as well as file data.

Igneous is also the first data-protection solution to offer native API integration with Qumulo QF2, and is also the only non-D2D-based data-protection option for Qumulo customers.

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, continue to scan and move unstructured data one file at a time. This approach to data comparison and movement slows down the overall process of both full and incremental backups, limiting their usefulness in very large environments. And, since these other solutions do not monitor the effect of backup jobs on production system performance, they may still be constrained to run only during off-peak hours.

> *Many of the newer, post-NDMP backup solutions still include architectural constraints that limit their ability to protect unstructured data at scale.*

Only DataProtect utilizes highly parallel crawler threads that quickly identify new and changed file-system data, and a data-movement engine that packages small files in bulk while breaking large files up into multiple simultaneous move operations.

DataProtect even outperforms other new backup platforms, delivering industry-leading throughput that copies data at line speed while protecting production system availability.

Archiving Unstructured Data

Even with effective backup and recovery protection in large-scale environments, data-centric organizations have an additional challenge to contend with.

As their unstructured data footprint grows at double-digit rates, they must either continue to expand their Tier-1 NAS capacity indefinitely, or develop a strategy for continuously migrating older data off their primary NAS storage to make room for the new data being 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.

Without an effective strategy to offload aged data, enterprises with double-digit data-growth will need to expand their Tier-1 NAS capacity at unsustainable rates.

Archive vs. Backup

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

Archive Objectives

Even as it ages, data 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: typically 1-3 years for most data types, 3-7 years for some financial and legal data, or indefinitely in rare instances.

Regardless of the specific retention periods, tagging policies, or data types, an archive solution 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 don't scale effectively. In very large environments, there are few consolidated archive solutions that can satisfy all of the above conditions.

Archiving data at scale requires the ability to search through hundreds (or thousands) of discrete datasets to find any that have aged out of active use.

Challenges for Archiving Data

While each organization will have its own criteria for determining how data qualifies for archive – e.g., files older than a certain age, datasets that have not been accessed in a predetermined number of months (or years), or data associated with defunct teams / projects – the task of identifying that data, particularly in large enterprises with hundreds of terabytes or more of active files, presents a daunting challenge for file and storage administrators.

While NAS vendors offer software to automate this process, these solutions are platform-specific and do not work with other NAS systems. Multiplatform third-party software breaks down at scale: a 70PB dataset can take months to complete a single end-to-end scan, by which point the early scan results are already stale and of limited value in identifying data for archive.

Moving Data to Archive Storage

Identifying datasets for archive, amid petabytes of active data, is only part of the overall problem confronting large-scale enterprises. Once aged files and directories have been identified, the next challenge is the process of migrating massive datasets to archive storage. There are several available options, each of which comes with its own constraints and limitations.

> *Each NAS vendor offers their own tiering solution – requiring more hardware and software costs, adding operational complexity and increasing vendor lock-in.*

Platform-Specific Archive Options

NAS vendors offer tiered storage solutions that usually include some combination of costly flash capacity and high-density, low-speed storage. 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 archive-level storage as it ages out of active use.

Just as with other vendor-specific solutions, these solutions are each restricted to their own NAS systems. A heterogeneous enterprise that needs to archive data from a mix of Dell EMC Isilon, NetApp FAS, Pure FlashBlade and/or Qumulo QF2 must manage each platform's archive capacity and operations separately.

Additionally, since all these data-tiering options require the software purchase separate from storage capacity, licensing costs again factor into consideration. The net result is often higher cost, increased complexity in heterogeneous environments, and an even deeper state of vendor lock-in.

Many organizations, particularly in very large environments, find themselves either unwilling or unable to continue to expand their onsite storage capacity using that model.

Archiving Data to Cloud Storage

The same factors that are causing enterprises to adopt a cloud-based backup strategy – data center space constraints, infrastructure and licensing costs, a need for increased operational agility – are also driving them to archive some or all of their aged datasets to cloud storage.

Unlike backup use cases, with regular incremental upload transactions and restore requests that mean additional transactions to pull files back down from the cloud – at a metered cost per transaction – archive workflows require only a single upload of the dataset. Intended only for very occasional use, archived data can sit in cold storage for months, or even years, before any data needs to be retrieved. This makes archive use cases a very attractive fit for cloud storage.

> *Public-cloud storage offers deep, cost-effective archive capacity for space-constrained enterprises.*

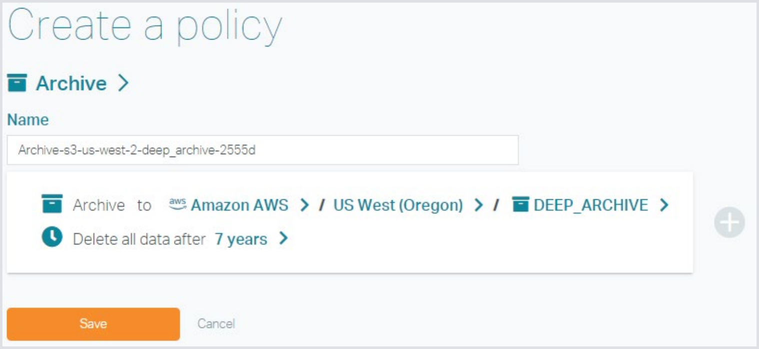
For any of these cloud storage platforms to be a useful hosting endpoint, however, IT needs a means for uploading data that is simple, automatic, and searchable after-the-fact for archive data that needs to be retrieved.

Platform-specific solutions that integrate with cloud endpoints deliver some of this functionality, but only on a limited basis. They do not offer consolidated data archival from heterogenous NAS platforms, and provide only partial search and retrieval capabilities. They may also not be optimized for high-latency, low-bandwidth, wide-area network (WAN) transfers.

Simple, Cost-Effective Archive with Igneous DataProtect

Unlike the vendor-specific solutions, Igneous DataProtect provides a single, unified platform that is universally compatible with all NAS systems.

In addition to providing regular, high-speed backup for unstructured data at scale, DataProtect also offers policy-driven archive services. Data can be archived in a single operation to an optional onsite Igneous storage tier, archived locally and replicated to the customer's preferred public-cloud endpoint, or archived immediately and directly to public-cloud storage.



Create a policy

Archive >

Name

Archive-s3-us-west-2-deep_archive-2555d

Archive to aws Amazon AWS > / US West (Oregon) > / DEEP_ARCHIVE >

Delete all data after 7 years >

Save Cancel

The enhanced integration that DataProtect delivers enables customers to automatically replicate or archive unstructured data directly to any of the following:

- Microsoft Azure – Hot, Cool, and Archive storage tiers
- Google Cloud Platform – Regional, Nearline, and Coldline storage tiers
- Amazon Web Services – S3 Standard, S3 Infrequent Access, 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.

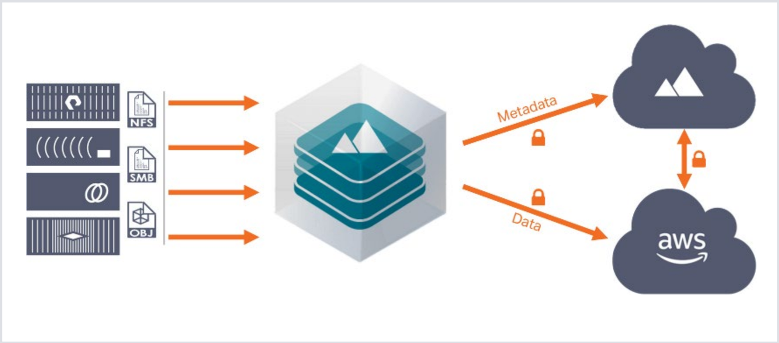
Deployment Scenarios and Options

Where NDMP requires a heavy investment in proprietary hardware – including servers, tape drives or disk shelves – and D2D comes with vendor-specific requirements for duplicate NAS hardware, Igneous offers customers a choice of deployment options that meet their specific environmental and operational needs, whether for backup, archive, or both.

To meet these requirements – which can include a combination of factors such as security, cost, recovery-time objectives, local environment restrictions, and performance – Igneous DataProtect is available in a number of different deployment scenarios, ranging from local-only hardware and storage to an exclusively cloud-based deployment that uses virtual machines for onsite data scanning and movement to the cloud.

Virtual-Only

The simplest iteration of Igneous DataProtect is deployed in the customer's environment using only a virtual machine. Rather than use onsite disk capacity for either backup or archive data, this deployment option makes use entirely of cloud storage.



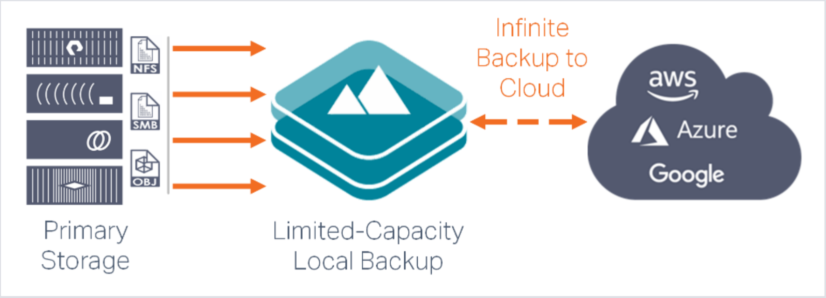
Data is backed up directly to the customer's own public-cloud endpoint, while the accompanying metadata discovered during each backup run is uploaded to the customer's own Igneous-specific cloud target, for use in managing data restore operations or as part of Igneous DataDiscover.

This solution is intended for customers whose available rack space is either limited or nonexistent.

Limited Local with Full Cloud

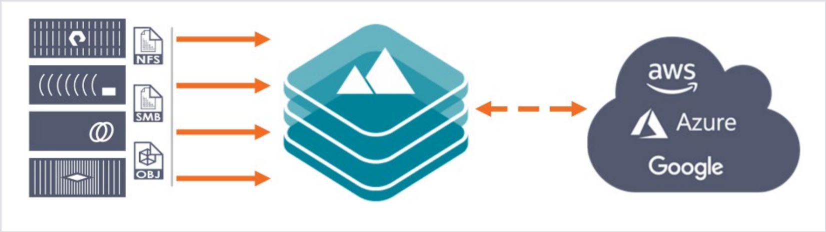
For customers who need quick access to only a partial subset of their backup data, and who can move the rest to cloud for longer-term retention, DataProtect can be deployed using a small-footprint hardware instance.

Backup policies are configured to replicate all data to the customer's cloud endpoint, keeping only the most recent backup images on local Igneous storage. Data beyond a certain age (e.g. more than 30 days old) is automatically deleted from the local Igneous storage.



Full Local with Full Cloud

The most common deployment option for DataProtect is to use an onsite hardware instance that backs up the full complement of their unstructured data, then replicates it to the customer's own preferred cloud platform.

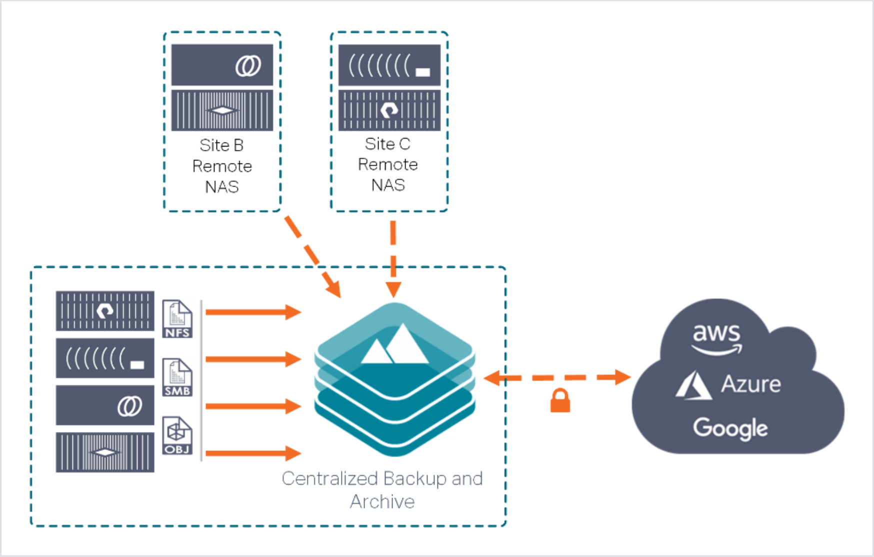


This type of deployment provides a full duplicate of customer data in the cloud, and offers similar data protection to the legacy approach of NDMP-to-tape with regular offsite rotation of one backup set.

Centralized Backup with Full Cloud

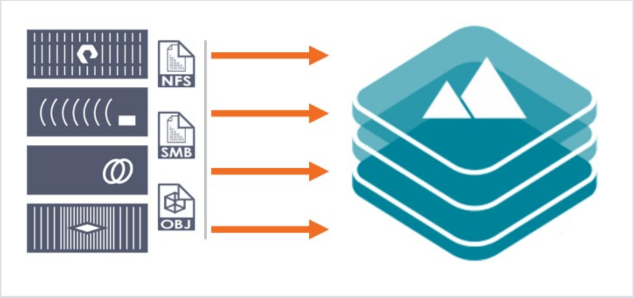
With DataProtect's efficient data-movement engine that makes use of all available network bandwidth while continuing to prioritize and protect production workloads, some enterprises are able to use a centralized physical Igneous instance as the backup platform in a multi-site environment, pulling data from remote NAS systems and sites to a consolidated central backup repository.

As with other local-to-cloud hybrid deployments, backup and archive data can be replicated in its entirety to the customer's cloud endpoint, offering an additional layer of data protection against site-level failures.



Local Only

For enterprises whose security or operational requirements preclude the use of cloud storage as an option, DataProtect is also available as a local deployment, in which both backup and archive storage are hosted onsite.

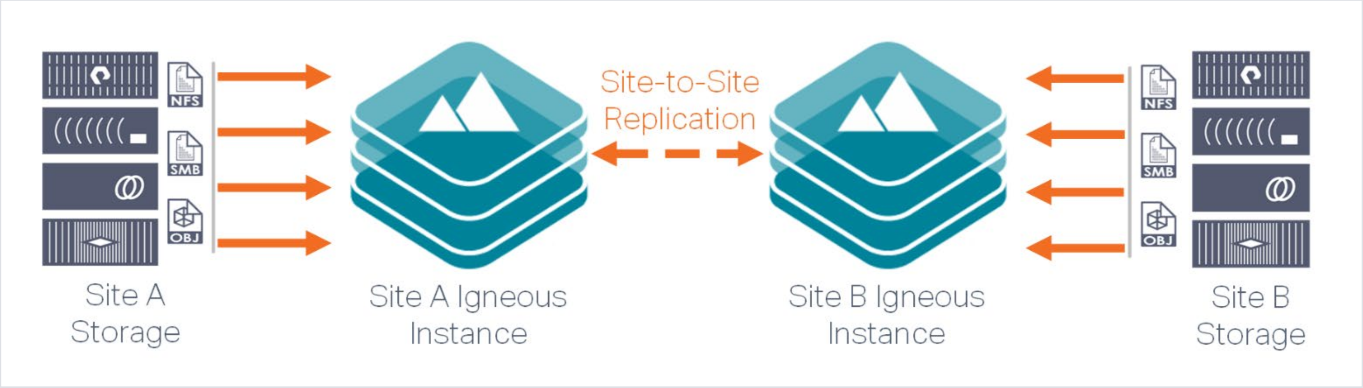


Igneous' easy scalability means that, even with a local-only deployment, the Igneous instance's local storage capacity can be expanded to whatever capacity is needed to host the full portfolio of customer data.

This type of deployment offers data protection and archive capability, but without the use of a cloud replication target to provide added data protection, customers who use this approach are vulnerable to loss of data resulting from site-level failure.

Site-to-Site Replication

Even without the option for cloud replication, DataProtect can still be configured to provide an added layer of data protection, and safeguard against site-level failures, by using a multi-site Igneous deployment with site-to-site (S2S) replication between Igneous instances.



This deployment type is for multi-site customers whose operational or environmental conditions preclude the use of cloud storage as an added data-protection layer, but who still need multi-level data redundancy and protection against sitewide failure events.

As with the local-only deployment type listed above, the capacity of each site's local Igneous instance can be expanded as needed to ensure full data protection across the entire enterprise.

Delivered and Managed as-a-Service

All Igneous solutions – including physical-only, virtual-only, and hybrid instances – are delivered and supported using the same as-a-Service model. Monitoring, diagnostics, failure and event management, and software updates are all handled remotely by Igneous. Customers need only add their NAS systems to the Igneous inventory, then create the appropriate data-management policies. Igneous provides full system visibility and usage metrics, and alerts administrators to any issues detected during data-management operations.

Igneous DataProtect lets customers focus on data, not hardware.

With an Igneous unstructured data-management solution, the control plane is managed remotely by Igneous. The data plane – including protection and management policies, index-and-search, all Igneous-related services in all sites – is managed through a single, intuitive web portal.

Cloud-Native Compute Model

Built using resilient, container-based microservices, Igneous delivers scalability and resiliency across all components: data movement, data index, and data storage components are purpose-built for optimal availability and performance at scale. This approach enables nondisruptive software releases, while ensuring that system performance remains unchanged, even while the size of the managed environment continues to scale.

Additionally the Igneous cloud-native model enables system updates – feature releases, bug fixes, and security patches – to be remotely and transparently added to a production Igneous deployment on a weekly basis, with no impact to production performance or system uptime.

Success Assurance

Igneous' as-a-Service delivery model includes automatic enrollment of all Igneous customers in the Success Assurance program, which includes a white-glove onboarding experience and SLA monitoring of all data protection tasks through the Backup Assurance program. Customer success is also ensured through Igneous' proactive capacity management and monitoring services and custom reporting capabilities.

Seamless, Effortless Scalability

A physical Igneous UDM instance is based on a modular architecture, consisting of one or more 2U Application Service Router (ASR) devices, and one or more databox devices, each of which offers up to 426TB of raw disk capacity.

As enterprises continue to scale their unstructured data footprint, they can also scale their Igneous deployment to match, by adding more ASR devices for increased scan and search performance, or adding more databox devices for greater backup throughput and/or storage capacity: with compression, upwards of 4PB of data can be hosted in a single 42U rack, and customers can continue to scale from there as needed.

> *DataProtect scales simply by adding components: ASR and databox instances for physical environments, and more virtual machines for software-only deployments.*

Since each added ASR component brings more compute capacity to the deployment, and each new databox contributes disk capacity and network connectivity, performance and throughput scale linearly as new components are added.

With Igneous' as-a-Service deployment and support model, adding components to a physical deployment is a seamless process, with new ASR and/or databox capacity automatically joined to the existing instance.

Virtual-only customers can add more virtual machines to their deployment as needed, whether to expand data-management services to additional sites or to increase overall backup throughput.

Conclusion

Regardless of overall size, every enterprise has critical data that needs to be protected. Also regardless of size, nearly every enterprise has some subset of data that is no longer in use but not ready to be deleted either. Smaller and medium-sized enterprises have been able to use standard backup solutions – either NDMP or D2D – to provide effective data protection and archive services.

As unstructured data scales up in size, however, the ability of these legacy backup and archive solutions to deliver effective data protection and data tiering diminishes. Adding more NDMP servers and silos, or creating new D2D replication pairs, increases operational cost and complexity while failing to scale service levels to meet ongoing demand.

Additionally, as onsite data centers run out of space, or as enterprises phase out their local operations in favor of co-located data centers and cloud services, these same organizations find themselves unable to continue with their legacy data-protection and tiering strategies, and often facing overwhelming challenges associated with moving petabytes of data into the public cloud.

> *Igneous simplifies alignment with industry trends moving away from legacy backup platforms, away from local data-center operations, and toward cloud-based platforms and solutions, even at scale.*

For these enterprises, Igneous DataProtect offers scale-out backup and archive, seamlessly connected via API-level integration to all the major NAS platforms, and native integration with all public-cloud platforms through simple, intuitive data-management policies.

DataProtect's highly efficient file-system crawler and data-movement engines can move data at line speeds, protecting billions of files with almost no administrative overhead. Igneous' 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.

With Igneous' flexible deployment models, scalable architecture, and efficient, high-performance data-management solutions, even the largest customers can simply, reliably, and effectively manage their data, at any scale.

Contact Igneous

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

- Effectively manage and scale growing unstructured data environments
- 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 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

Igneous