Igneous

WHITEPAPER

Igneous Unstructured Data Management

Part I: Data Protection, Data Discovery, and Data Flow

Abstract

Effective management of unstructured data requires active maintenance and appropriate action at every stage of its lifecycle. Data owners and data stewards must ensure that critical files and folders are protected against loss, indexed for visibility and analytics, relocated or replicated for user access, and archived when no longer needed for active use.

This whitepaper is the first of a series that outlines Igneous' solutions portfolio – delivering high-performance, high-density data management – for managing unstructured enterprise data at multi-petabyte scales.

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

Table of Contents

Introduction

In the modern, data-driven economy, IT organizations are being squeezed from all sides, facing business pressure to: launch new applications; expand services on existing applications; integrate new cloud services; and meet the ongoing demands of runaway data growth. Although IT operations have become leaner and more efficient overall, IT departments continue to face flat budgets and headcount freezes.

While these industry headwinds combine to make it increasingly difficult for today's IT departments, the accelerating proliferation of data – unstructured data in particular – presents a particularly difficult challenge.

Document Purpose

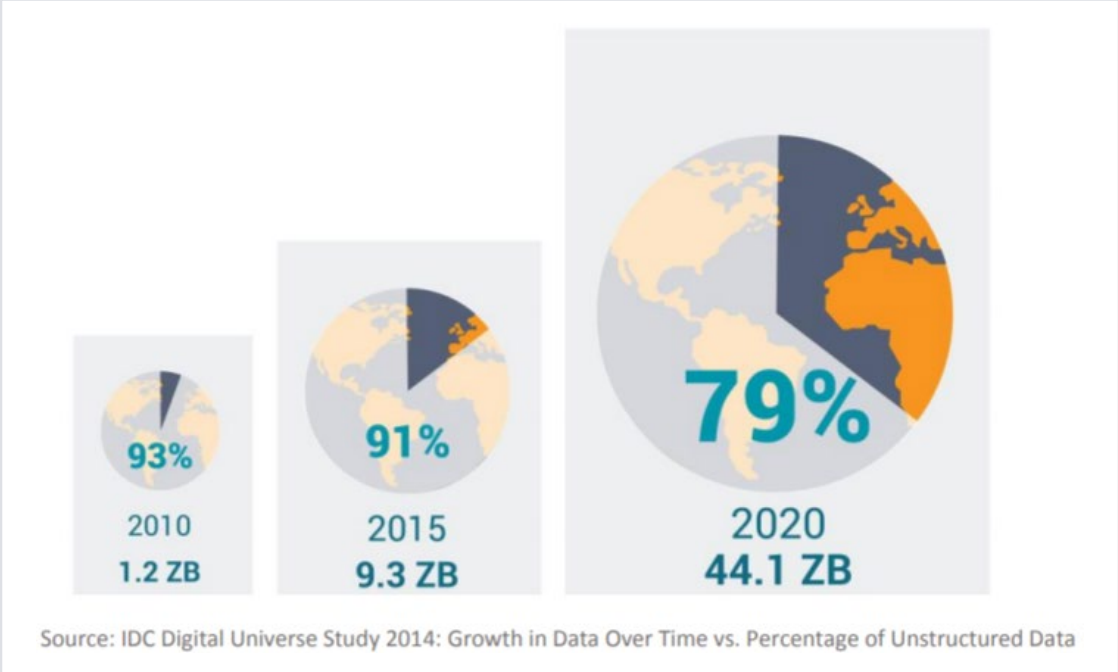
This document is the first in a series of papers that outline Igneous' data management solution offerings. While the other papers each focus on a core aspect of unstructured data management – data protection, data discovery, data flow – this document provides a more general overview of:

- Unstructured data as a concept
- The challenges of managing unstructured data at petabyte-plus scales
- Igneous' application and infrastructure architecture as a means to enable unstructured data management and simplify 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).

Business Challenge: Runaway Data Growth

According to IDC, there were 9.3 zettabytes (9.3 x 1021 bytes) of data stored worldwide in 2015, with up to 91% of that being unstructured data. IDC predicts that number will grow to 44.1 zettabytes by 2020.



Source: IDC Digital Universe Study 2014: Growth in Data Over Time vs. Percentage of Unstructured Data

Unstructured Data

As a term, "unstructured data" generally refers to data that is stored as files on enterprise storage, as opposed to "structured data" comprising databases and virtual machine images, or file data hosted locally on application servers.

Depending on the industry, unstructured data may take any of a number of different forms. Electron microscopy images, electronic design schematics, media files, seismic data, market data, and text files are common examples, though there are a number of other types as well.

Although end users can generate significant amounts of data, most of these files are primarily created by automated processes and applications. For some enterprises, both sources can collectively add up large amounts of new, unstructured data on a daily basis. In extreme cases, organizations and their IT teams need to plan for terabytes of new data per day.

Business Challenge: Management Sprawl

The unstructured data within a single large organization can easily total billions of files. In the modern enterprise, these files can be anywhere and everywhere – hosted on hundreds of local NAS systems, spread across dozens of sites, as well as in public cloud-based data repositories – collectively consuming petabytes of storage capacity.

To accommodate the demands of different data types, and the workloads that generate and consume unstructured data, a single IT shop typically uses multiple NAS platforms from multiple vendors – Dell EMC Isilon™, Pure FlashBlade™, Qumulo QF2™, and NetApp are among the most popular – each of which is optimized for specific cost, capacity, and performance profiles, and which may collectively host hundreds, even thousands, of individual exports. While there are different levels of administrative complexity and overhead associated with different NAS platform types, an enterprise that uses a heterogeneous mix of NAS platforms needs to maintain administrative bandwidth and and engineering expertise for each of them.

In recent years, as datacenter space has become scarce (and costly), with limited expansion opportunities, end users and IT enterprises have turned increasingly to public-cloud storage options for both quick capacity expansion and long-term data retention.

While this resolves the short-term problem of where to host massive amounts of unstructured data, it opens up new challenges:

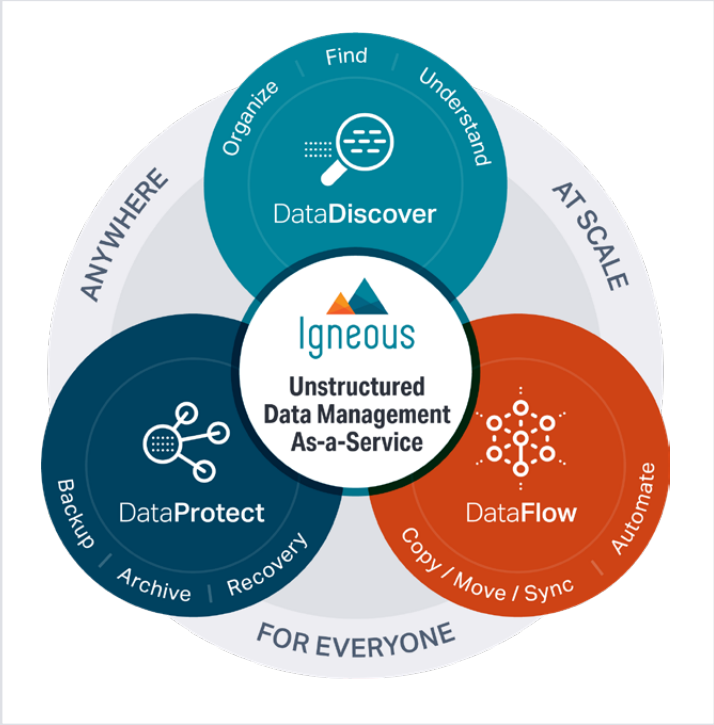
- Determining which datasets can be appropriately migrated to cloud storage
- Migrating massive datasets to the cloud regularly
- Rapidly and reliably locating data anywhere in the enterprise, including on cloud storage

When budgets, headcount, and datacenter resources are heavily constrained, IT needs to ensure that the unstructured data in their environment can be appropriately hosted, protected, and tracked at every stage of its lifecycle, from the moment it is generated until the day it is retired.

Igneous Manages Unstructured Data

To address these challenges, and to position customers for future success, Igneous delivers the first comprehensive data management solution for unstructured data. Purpose-built to handle billions of files, hundreds of file systems, and petabytes of data, Igneous provides data management solutions for unstructured file data at scale, all delivered as-a-Service. These services include:

- Data protection with **Igneous DataProtect**: backup, archive, and restore critical unstructured data
- Data movement with **Igneous DataFlow**: automated or user-managed data copy, movement, and synchronization operations
- Data visibility with **Igneous DataDiscover**: organize and find data, manage, and optimize data and capacity usage



Igneous offers highly-scalable, policy-based data movement to enable backup, migration, and archive operations at line speeds. Its index and search engines let IT locate, retrieve, and analyze petabytes of unstructured data quickly and easily. Offering full integration with primary NAS systems at the application programming interface (API) level, Igneous is unique in enabling comprehensive backup operations while protecting production NAS performance.

With on-premises infrastructure designed for data-centric computing, and transparent integration with cloud services, Igneous eliminates the challenges of managing disparate (and complex) unstructured data backup solutions, data migration and replication operations, data tiering and archive services, and data analytics.

Igneous offers simplicity at scale for enterprises whose unstructured data environments range from hundreds of terabytes to hundreds of petabytes, and tens or even hundreds of billions of files. Engineered for performance at any scale, to run anywhere, and with an as-a-Service management model, Igneous delivers powerful data management services while reducing the administrative and financial burdens that unstructured data creates.

Effective Unstructured Data Management

For effective lifecycle management, unstructured data requires more than just providing a NAS array and cloud capacity to store it. Most data passes through a lifecycle as it ages, transitioning from "hot" to "cool" to "cold," with each phase having its own requirements for user access and performance. As data progresses through these phases, an effective management strategy will recognize the change in access frequency and in data age, then migrate the data to the appropriate storage tier in response.

Backup and Archive with Igneous DataProtect

Most unstructured data, even at enormous scales, is considered to be business-critical, meaning it must be treated as a crucial corporate asset, protected against loss, inventoried and indexed for discoverability, and hosted on storage appropriate to its business value.

Without an effective data protection solution built on regular backup operations, companies are at risk of data loss resulting from hardware failure, user error, or malicious attack. Data loss can lead to business disruption, loss of revenue, and financial penalties. A business that irretrievably loses critical data may be forced to cease operations altogether.

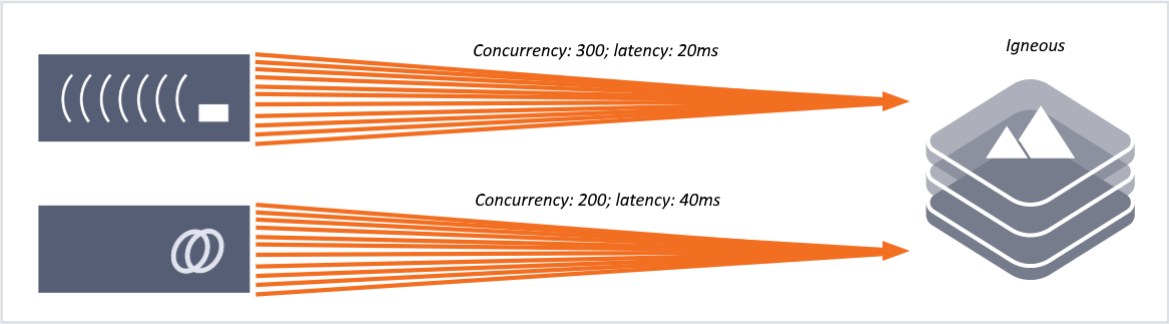
With native support for NFS and SMB file-access protocols, Igneous offers full vendor independence for a single, consolidated backup solution that protects unstructured data on any NAS platform. Igneous' high-density appliance absorbs petabytes of primary data easily. Its modular architecture enables capacity expansion on an as-needed basis – more storage for the data layer, more compute for the indexing engine – independently.

Unstructured Data Protection at Scale

Igneous provides an easy-to-use, comprehensive data protection platform, enabling integrated enterprise backup that is both highly efficient and highly scalable for unstructured file data. 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.

Faster Backups and Faster Restores

Igneous DataProtect can connect to and protect data from any NAS platform, pulling data from NAS sources in parallel streams and enabling greater overall throughput rates. At the same time, its latency-aware data movement engine monitors overall storage performance during backup operations, throttling back when necessary to protect business-critical applications that depend on primary NAS performance.



Configuring and enabling backup protection for unstructured data is a simple process. Igneous DataProtect leverages easy-to-define, easy-to-apply policies to determine each target's backup frequency, data retention, replication target and archive settings. A policy can be applied at the individual export/share level, or an entire NAS system can be protected using a single policy.

Different data profiles require different handling for optimal performance and throughput. Backup jobs that target small files are handled differently from backup jobs consisting primarily of large files. Igneous DataProtect automatically adjusts its scan-and-move processes in response to the data it discovers during normal backup operations.

Igneous DataProtect runs an initial full (i.e. "Level 0") backup operation on a new target. All subsequent backup jobs are incremental, capturing only new and changed objects for faster backup completion. When a file or data set needs to be restored from backup, Igneous presents a "virtual full" look at the file system corresponding to the requested recovery point.

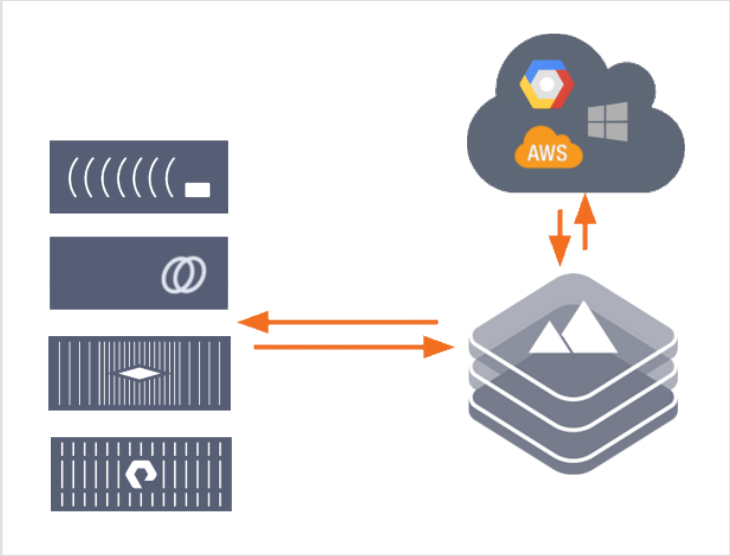
Compared with most NDMP-based solutions, in which a restore operation begins with a search through the catalog to identify the necessary tapes, then continues through a multi-step process of restoring from multiple incremental backups, Igneous' approach is simpler, faster, uses less overall capacity, and shortens recovery turnaround.

Simplified Archive with Igneous

For most organizations whose unstructured data footprint reaches petabyte-plus scales, new files are generated constantly, often numbering terabytes' worth of new files per day. In order to maintain sufficient capacity on their primary NAS systems, IT must move older data off the primary tier to an archive repository, either on-premises or in the cloud.

At multi-petabyte scales, identifying data for archive presents a formidable challenge. Moving data between platforms at scale is another challenge, particularly between heterogeneous storage platforms, between sites, and to public-cloud endpoints.

In addition to offering high-throughput data movement and dense storage at petabyte scales, Igneous DataProtect enables the consolidation of archive data from every NAS source in the enterprise. Organizations that deploy a mix of NAS architectures can use an Igneous data management solution to centralize data from all these systems to a single archive repository, or to a tiered archive solution that leverages both on-premises and cloud storage.



Based on settings defined in the assigned policy, Igneous DataProtect can be configured to use an onsite Igneous storage appliance as the primary archive tier, with secondary replication to cloud storage; or stream archive data immediately to a public cloud archive repository.

The enhanced integration that Igneous 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

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. 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.

Indexing and Finding Files

Backup and archive alone, however, are only two services in what needs to be a larger management approach for unstructured enterprise data. To ensure effective stewardship of company resources, an automated, scalable management strategy for unstructured data becomes essential to daily operations.

In many cases, particularly in enterprises with petabytes of unstructured data, a comprehensive approach needs to support lateral movement of data – between production NAS systems and across multiple sites – in addition to a hierarchical migration from primary to archive to cloud layers.

Enterprises may also find themselves needing deeper visibility into the files they own or maintain, in order to ascertain which files may be necessary for legal discovery purposes, for intellectual property defense, for financial history, or for any of a number of other reasons.

In all of these organizations, data needs to be actively managed: protected, replicated, moved, or archived. None of these actions can be effectively planned, configured, or executed without data discovery – the ability to see the full portfolio of unstructured data that the company owns, complete with actionable information.

In large-scale enterprises with multi-petabyte unstructured data portfolios, this requirement can present an insurmountable challenge. When billions of files are distributed over hundreds of NAS systems in dozens of sites, a centralized indexing engine capable of reaching all these nodes and discovering all these files would require hundreds of CPU cores, hundreds of terabytes of high-performance storage, and one or more full-time IT administrators.

For all that investment, however, a full scan of the entire enterprise from start to finish would require weeks of discovery. Unless the organization operates with a very low change rate across the enterprise, any information collected using this approach would be obsolete by the time it was accessible.

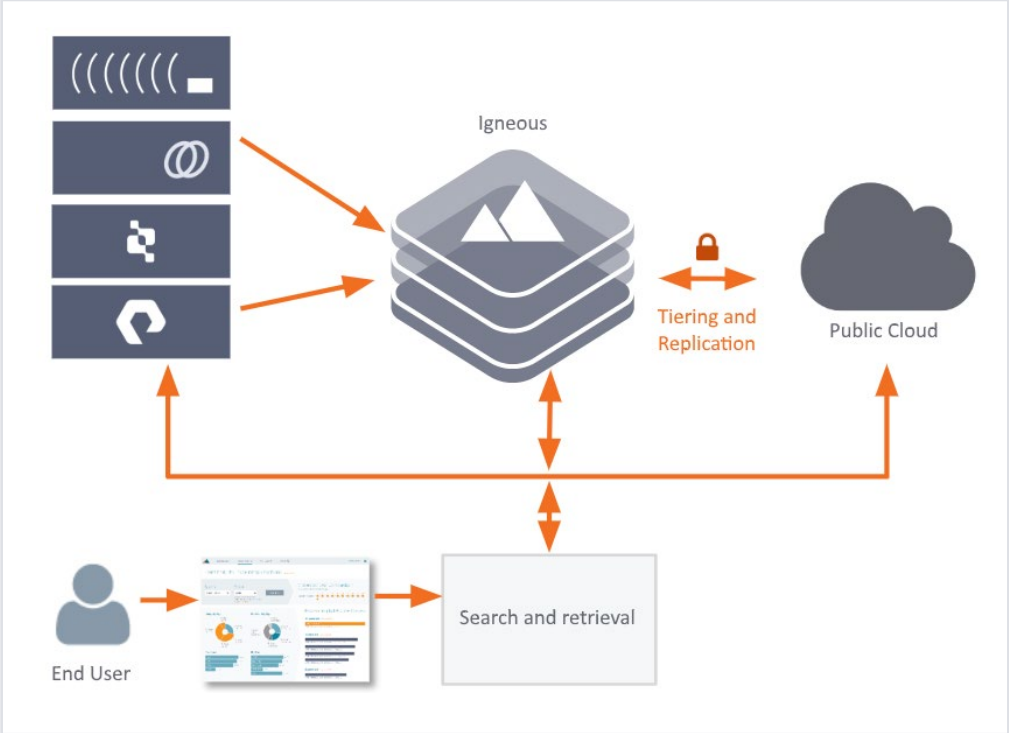
Igneous DataDiscover

An integral component of any Igneous data management platform, Igneous DataDiscover offers high-performance, platform-independent data discovery to provide deep and broad visibility into the entirety of any enterprise's unstructured data, even at petabyte-plus scales. Leveraging metadata collected during file-system scans and routine data-management operations, Igneous DataDiscover offers insights into capacity planning and management, usage metrics, and file lifecycle patterns to answer key questions about unstructured data at scale. Engineered for massive unstructured-data environments, Igneous DataDiscover scales to any size without sacrificing performance.

Index and Search

Igneous uses multi-threaded crawlers that can consistently and reliably discover and process billions of file changes on a daily basis, across multiple NAS and cloud platforms simultaneously. Information collected during the scanning process is compiled into a centralized, scalable, high-performance index, even in environments with hundreds of billions of files, locally, remotely, and in the cloud.

In addition to finding new and changed files during scanning operations, Igneous DataDiscover also tracks data movements to any and all managed destination tiers as files are backed up, migrated, replicated, or archived. Once compiled, the indexed metadata can be aggregated and transformed to searches, queries, and analytics; letting data owners and data consumers actively manage data for optimal access and utilization.

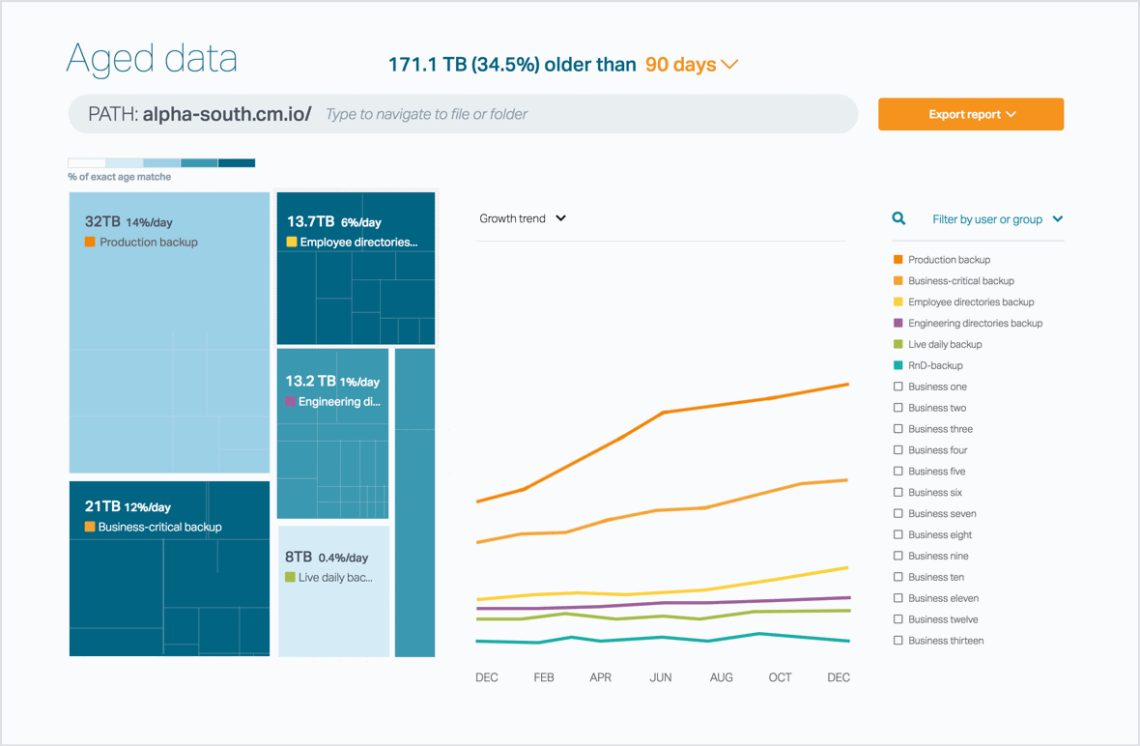


Data Management and Visibility

Enterprises that leverage Igneous DataDiscover can quickly organize and build a comprehensive understanding of their entire unstructured data portfolio, getting regular and reliable answers to critical questions about their environment:

- How much file and object data is in the environment?
- Where are all these datasets located?
- How old are these datasets and how are files changing?

With its always-current metadata index, Igneous DataDiscover can turn these questions into actions, offering a simple user interface to provide deep visibility into enterprise unstructured data.



With Igneous DataDiscover, unstructured data management becomes a shared effort, with all stakeholders – data owners, data stewards, and data users – each taking an active role in ensuring that unstructured data is managed like the critical business asset that it is.

Moving and Replicating with Igneous DataFlow

The final aspect of a comprehensive unstructured data management strategy requires enterprises to replicate or move data – between platforms, tiers, and locations as needed – to ensure that data owners, data stewards, and data consumers have direct and optimized access to their own data.

This service, called Igneous DataFlow, empowers data users to move, copy, and/or synchronize active, machine-generated datasets, using self-service APIs and IT-enabled processes, to ensure critical data is available where they need it, and when they need it.

Organizations with massive unstructured data footprints face a series of challenges in implementing Igneous DataFlow as a regular function of their overall operations: identifying the appropriate data for each step in the mobility process; taking action to replicate or migrate the identified data; and tracking data changes from source to target throughout the entire process.

Consolidated Data Movement

An Igneous solution uses standard protocols (NFS, SMB, and S3) to pull data from any type of NAS system, push data to any NAS system, and specifically engineered to provide optimal transfer speeds across LAN connections and WAN links, regardless of bandwidth speed, latency, or competing WAN traffic.

Configured, enabled, and managed by IT, these tools – move data between local NAS systems, between remote NAS systems, between NAS and cloud, or between clouds – can be given directly to end users. IT can enable consumers to move, copy, or synchronize the data they need, with access access controls to provide security and services managed to optimize costs.

With Igneous DataFlow, users can move data automatically from a NetApp storage instance in one location to a Pure FlashBlade array in another, or from a Dell EMC Isilon cluster to multiple Qumulo QF2 clusters and an Amazon AWS endpoint, with no additional agents or software, and without any specific WAN tuning.

Users, applications, and partners can access and consume data locally, even data generated on the other side of the world, using Igneous DataFlow as a key enabler of their scale-out data-management strategy.

As-a-Service Management

Monitoring, diagnostics, failure and event management, and software updates are all handled remotely by Igneous. Customers have only to 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.

With an Igneous 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 – can be 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 non-disruptive 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.

Conclusion

Any enterprise organization with a significant unstructured data footprint can find it challenging to manage effectively. New data needs storage capacity to land on. Active datasets need to be accessible on-demand for work groups, teams and projects to use. Nearly all data must be regularly and reliably protected. Old data must be identified and archived.

For data to be managed, it must be visible. An additional challenge in unstructured data shops is to maintain an active index of the organization's datasets – including location and path, size, usage patterns, and age – to ensure appropriate action is taken at the appropriate time for every phase in the lifecycle of the data.

At petabyte-plus scales, what was difficult in smaller environments can quickly become impossible using legacy enterprise data-management solutions. From maintaining adequate Tier-1 NAS capacity for new datasets, to regular backup and archive services, and data flow operations, old strategies quickly break down under the strain of massive data loads. This is especially true when it comes to finding, indexing, and searching for data, since most enterprise file-system crawlers are unable to keep pace with the rate of change in most unstructured data enterprises.

To address these challenges, and provide a comprehensive solution for discovering, protecting, and managing unstructured data at massive scales, Igneous engineered an all-new platform that recognizes the difficulties that petabytes of unstructured data can create.

- With Igneous DataProtect, enterprises have a high-performance backup and archive solution that works with any NAS platform and cloud provider; optimized for high-speed data transfers in parallel streams without impacting production file services.
- Igneous DataDiscover enables deep and broad visibility into the entire unstructured data portfolio, with near real-time indexing and high-performance data search and query capabilities.
- For enterprises that need to enable end-user management of large-scale dataset migration and synchronization, Igneous DataFlow lets IT offer unprecedented consumer control, delivering business agility without compromising data security.

With the unstructured data management features that an Igneous solution provides, and with its as-a-Service implementation and support model, Igneous has created a new, modern architecture that simplifies operations, speeds up services, and lowers total-cost-of-ownership.

Large-scale enterprises that are grappling with the challenges unstructured data create will find that, an Igneous solution simplifies the management of these challenges, while delivering the services that unstructured data requires.

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-independent 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