Igneous

WHITEPAPER

Igneous Unstructured Data Management – Data Discover

Part III: Finding, identifying, and analyzing unstructured data at petabyte scales

Abstract

This white paper is the third in a series that outlines Igneous' solutions portfolio for managing massive amounts of unstructured enterprise data.

Effective data management requires the protection of data against loss, movement of unstructured data between sites and storage tiers within the enterprise, and data archival as appropriate for possible later retrieval. For any of those services to be effective in petabyte-scale enterprise environments, data discovery – indexing, searching, and analyzing data – is a crucial element in unstructured data management, and the primary focus of this paper.

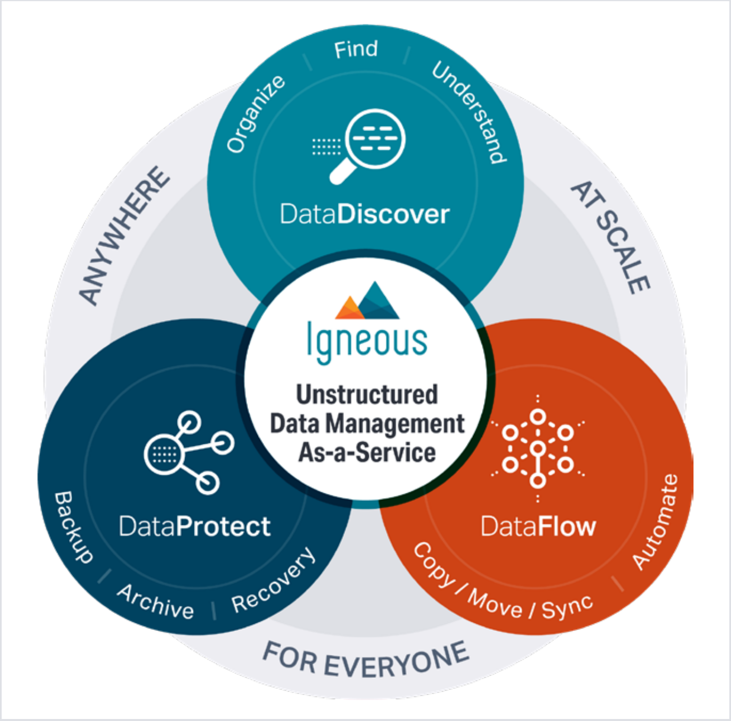
Igneous
2401 Fourth Ave., Suite 200
Seattle, WA 98121
(844) IGNEOUS
igneous.io

Table of contents

Introduction

Managing unstructured data requires more than just a NAS system to host it, or a backup solution to protect it. For organizations with massive-scale unstructured enterprise data, an effective management strategy requires a comprehensive suite of solutions that:

- Discovers, recognizes, and tracks data, throughout the enterprise and in the cloud
- Protects data against loss
- Migrates data as needed between storage systems, tiers, and sites
- Enables data owners and administrators to find, move, and use their data quickly and easily



None of these data management strategies can be effectively developed, configured, or executed without a comprehensive data discovery capability that enables enterprises to see and understand their entire unstructured data environment.

When an enterprise's unstructured data footprint exceeds hundreds of terabytes, the need for visibility becomes more critical to overall business operations at the same time that it becomes more challenging for IT to deliver.

Document Purpose

This document, the third in a series of papers that outline Igneous' data-management capabilities, provides an in-depth look at Igneous DataDiscover's deep-dive visibility into an organization's entire unstructured-data portfolio, offering index, search, and analytics capabilities that give data owners and administrators up-to-date visibility and understanding of their entire data environment .

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

Data Visibility

While “data management” as a term generally calls to mind the need for backup, replication, migration, and archive, these actions are hard to implement and optimize without a full understanding of the enterprise’s full data portfolio, including:

- How much data the enterprise owns
- Where data is stored
- How it is used
- Who uses it

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

The challenge that enterprises and organizations – both data owners and data administrators – must contend with is the scale of their data footprint. While data visibility is crucial to effective data management, and while enterprises have spent significant time, dollars, and resources in pursuit of data visibility, the tools available for viewing and analyzing data either don’t work at scale, don’t cover the full scope of the organization’s unstructured data portfolio, or don’t provide outputs that lead to actionable insights.

> *Data visibility is crucial to effective management, but despite enormous infrastructure and administrative investment, the current suite of tools is inadequate to the task.*

Data Discovery Objectives

At an enterprise level, data discovery requires the ability to:

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

When an organization’s data footprint scales to petabytes or more, full visibility becomes exponentially more critical for effective management. At the same time, however, it becomes exponentially more difficult, particularly in multi-site, multi-vendor environments where unstructured data can span multiple locations, platforms, and protocols.

At this scale of operations, legacy data discovery tools lose their effectiveness. Standard Linux admin utilities like ***grep*** and ***find*** provide only limited visibility even in small-scale environments. In large-scale enterprises, where a full end-to-end scan can take weeks or even months to complete, any insight these tools can deliver is out of data by the time their outputs are available.

Data Discovery Tools

As critical as data visibility and analytics are for unstructured data management, there are a limited number of additional tools that can deliver the insights needed for data at scale.

Vendor-Specific Storage Analytics

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

Even within the context of their own platform, these tools may offer limited functionality, e.g. delivering only high-level information rather than in-depth file analysis, or providing only per-system visibility, in which analytics from individual systems can't be aggregated or centralized, even when all NAS systems come from the same vendor.

None of these utilities offer comprehensive platform analytics that span the entire unstructured data environment.

Open Source and Proprietary Utilities

As with some of the other components of a holistic data-management strategy (backup, archive, migration, and replication), there are dedicated open-source and third-party tools that attempt to meet this need.

Proprietary tools are generally licensed based on the organization's total managed-data footprint. At multi-petabyte scales, even with significant volume discounting, annual licensing costs alone can total millions of dollars and still account for only a portion of the total cost. As the company's data footprint continues to grow, high–performance compute and storage infrastructure, along with licensing, must be built out to keep pace.

Additionally, having originally been written for home directories and workgroup datasets, most proprietary tools were never intended to scale to the size that modern enterprises need. As the index grows beyond its design limit, overall performance will degrade to unacceptable levels under the weight of hundreds of data sources, billions of files' worth of metadata, and the sheer complexity of the compute and storage infrastructure required to host it.

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

Most vendor-specific, proprietary, and open-source tools offer only limited visibility and limited functionality that quickly breaks down at scale.

Whether proprietary or open-source, most existing data-discovery tools offer limited or no analytics or movement capabilities. Once the full set of unstructured files, folders, and file systems has been initially inventoried, any potential actions based on analytical insights – e.g. data protection, archive, migration, replication, etc. – require more tools and more investment.

Scalable Visibility with Igneous DataDiscover

Igneous DataDiscover has been built with scalable, robust components to deliver comprehensive, up-to-date unstructured data visibility and understanding, including:

- A crawler engine that is both high-performance and low-impact, which scans the target file system or export tree, enumerating and indexing file and directory objects and metadata at any scale
- An indexing engine that can store hundreds of billions of metadata rows efficiently, capable of answering deep queries quickly with fast response times
- A search engine to present a user interface for file-system insights and object queries, passing results back to data owners and administrators

Igneous DataDiscover can be deployed either as a virtual machine (for standalone use) or by adding it to an existing on-premises Igneous instance (if part of a larger Igneous UDM solution), and requires no customer configuration beyond the initial NAS discovery. Once all file-systems have been added to Igneous DataDiscover, the file-system crawler engine begins to scan and index files and directories immediately.

Engineered from scratch to be platform-agnostic and work with any NAS architecture, Igneous DataDiscover bypasses the limitations of siloed legacy data-protection and data-index platforms. Enterprises who use Igneous DataDiscover quickly find that its high-performance scan engine can successfully track changes across an organization's entire portfolio of unstructured data, even in multi-petabyte enterprises and unstructured-data footprints that include billions of active files.

> *Igneous DataDiscover's multi-threaded, hardware-agnostic scan engine, combined with its powerful, centralized index and search capabilities, delivers near-real-time data visibility even in multi-petabyte environments with billions of files.*

Igneous Discovery and Indexing

The combined services of the crawler and indexing engines are key components of the Igneous data-management platform, capable of consistently and reliably discovering new and changed files anywhere.

Igneous File-System Crawler

Rather than a single, brute-force tree walk from one end of the entire file system to the other, Igneous deploys dozens, even hundreds, of parallel crawlers across all managed file systems on all managed NAS platforms. Once deployed, these crawlers can discover and index billions of file-system objects and object changes per day.

Each crawler engine is multi-threaded and latency aware, allowing Igneous to automatically scale up or down discovery resources, managing threads autonomously in response to changes in available resources on the target system, prioritizing production workloads and system response to ensure data availability at all times.

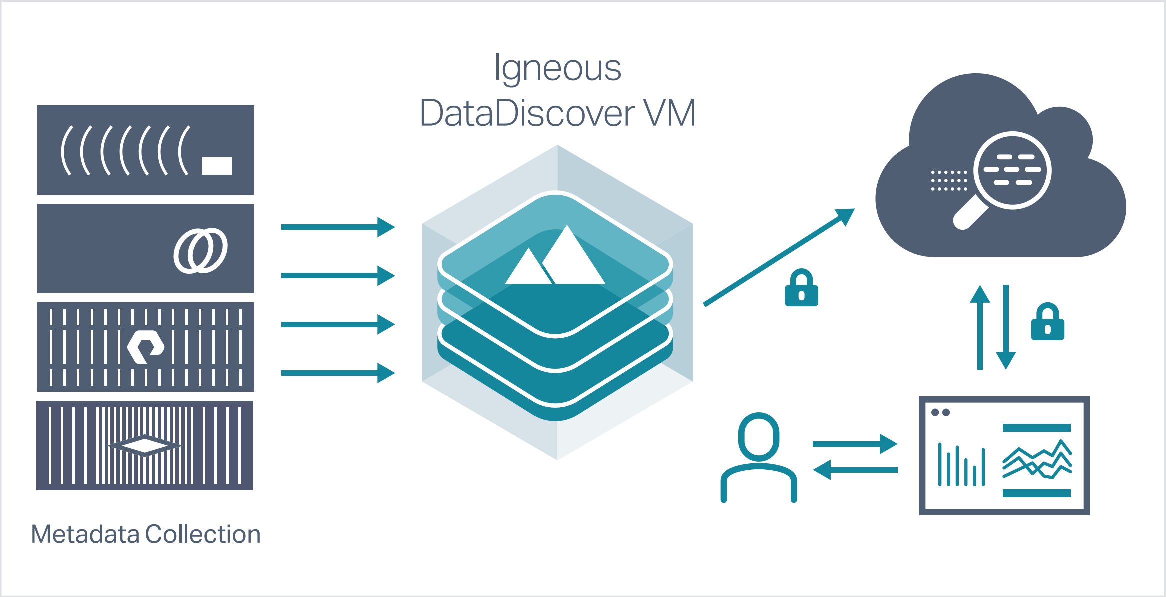
Igneous Indexing

Metadata collected by the crawler is fed into the index engine, which is also engineered to deliver consistently high performance as the overall unstructured data footprint scales, even to exabyte-plus scales and hundreds of billions of files.

As new files are created and existing files modified, the crawler service updates the metadata in the central index with each pass through the file system, providing near-real-time visibility to all unstructured data. Additionally, all Igneous-managed data movements to any and all destination tiers – whether datasets are backed up, migrated, replicated, or archived, locally or in the cloud – are also tracked via the same index service, meaning data can be found wherever it lives, even as it moves. Igneous DataDiscover maintains an active record of all managed, unstructured data at every stage of its life cycle.

Igneous Search

Using the metadata gathered during the discovery process, Igneous DataDiscover provides a single, searchable index that spans the full scope of managed unstructured data. With a cloud-native architecture that can scale to hundreds of billions of objects with no performance drop-off, Igneous' high-performance search engine provides rapid and direct access to the indexed metadata.



The information returned by the search engine lets organizations quickly identify specific datasets with key characteristics, even from a portfolio of billions of files. This enables IT admins to easily find patterns in their environment: excessively large datasets, unused datasets, duplicate datasets, or rapidly-growing datasets.

> *Igneous DataDiscover's near-real-time insights empower data owners and administrators to take control of even massive datasets, e.g. moving active data to high-performance storage, archiving stale data to the cloud, and optimizing enterprise storage capacity and value.*

With this information, enterprises can decide what further action to take on massive amounts of unstructured data, e.g. adjustment of backup settings, migration to a more suitable storage tier or site, archival to an onsite storage tier, or archival to cloud storage.

Igneous DataDiscover enables more effective stewardship of massive amounts of unstructured enterprise data, by offering deep and broad visibility into the full inventory of data, the progression of data lifecycle phases, and the necessary insights to manage this critical corporate asset.

As-a-Service Management

Monitoring, diagnostics, failure and event management, and software updates are all handled remotely by Igneous freeing up customers to focus on services – such as protection and management policies, index-and-search, data movement – and results, using a single, intuitive web portal. Igneous provides full system visibility and usage metrics, and alerts administrators to any issues detected during data-management operations.

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.

Conclusion

For comprehensive data management in petabyte-plus enterprise environments, organizations must be able to protect, archive, and migrate massive amounts of data on a daily basis, using methods that can be automated and audited across multiple storage platforms, between sites, to and from the cloud.

To meet the demand for these business-critical services, data administrators and owners need to be able to quickly discover data, track its movements throughout the enterprise, and analyze the aggregate sum of their data for patterns and actionable insights.

Igneous DataDiscover provides data visibility across an organization's entire unstructured data portfolio. Its highly-parallel file-system crawler engines offer near real-time insight into the full inventory of files – types, sizes, locations, and usage patterns – consolidating file-system metadata to a scalable central index that delivers high performance at any size. For end users, data owners, and administrators, the index engine presents a search tool to discover and manage data anywhere in the environment: quickly finding one file among billions, or one particular dataset among petabytes of unstructured files across hundreds of local and cloud-based endpoints.

With an Igneous platform, automated data management – backup, archive, replication, migration, and indexing – can be implemented across the entire portfolio of unstructured storage with a few clicks. Igneous' scale-out backup, index-and-search, and data-movement capabilities, seamlessly connected via API-level integration to all the major NAS platforms, enable protection of hundreds 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 us

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, API-driven data flow
- 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**.

Try Igneous for Free

Visit: www.igneous.io/datadiscover-free-trial