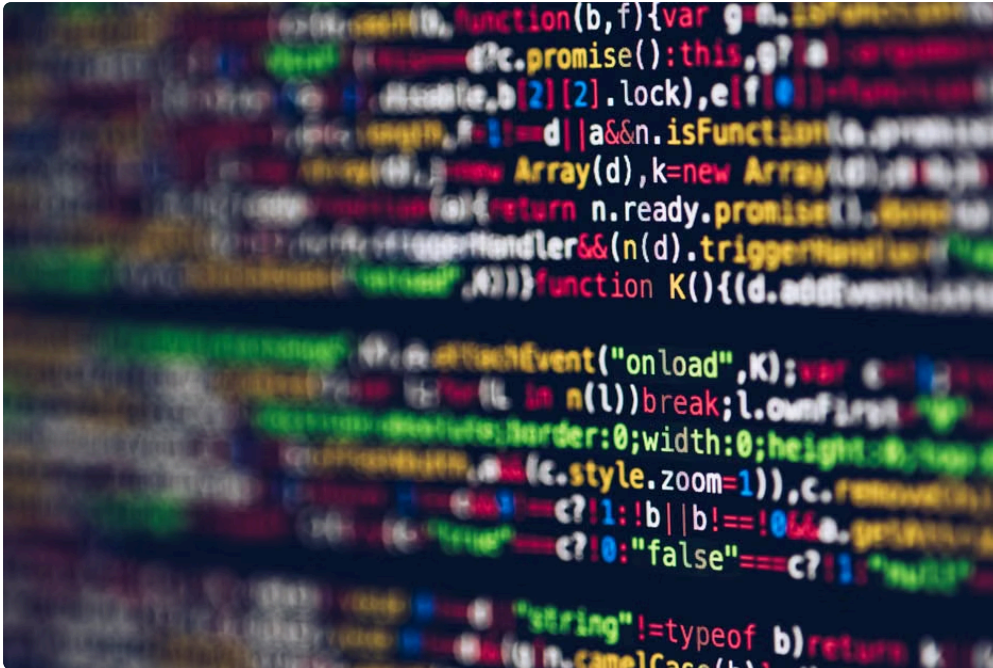


As organizations scale their data infrastructure, the ever-growing pools of unstructured data stored on NAS (Network Attached Storage) devices become both a treasure trove and a burden. Often vilified as "dark data," these files quietly accumulate—hidden from everyday analytics—yet they harbor critical insights and significant risk. Businesses must find ways to uncover, manage, and secure this metadata to fuel AI and analytics initiatives effectively, especially within modern data platforms like Databricks.



In this post, we'll explore how to query NAS file metadata in Databricks, leveraging tools like **Komprise Analysis** and **Komprise Intelligent Data Management**, and integrating with data lakehouse technologies such as Apache Iceberg. Along the way, we'll unpack key themes around *dark data*, unstructured data visibility, cost drivers, security exposures, and best practices.

What Is Dark Data and Why Does It Accumulate?

Dark data refers to data assets that organizations collect and store but do not actively use or analyze. Within enterprise file servers and NAS arrays, these often consist of old project files, documents, archives, unindexed images, and more. This "attic boxes" of digital information grows silently, driven by several factors:

- **Lack of clear ownership:** Files and shares created by departed employees or teams linger long after they're relevant—precisely the kind of mystery shares I keep track of.
- **Compliance inertia:** Steer-clear policies around deleting data "just in case" can cause exponential growth in data stored over time.
- **Backups and replication:** Multiple copies across storage tiers and backup systems create redundancy and bloat.
- **Decentralized storage strategies:** Teams provisioning their own NAS or shared drives without centralized governance add to sprawl.

According to Gartner, unstructured data grows at 50-60% annually on average, while only 20-30% is actively used, leaving a vast mountain of potential yet inert data. Dark data not only squanders storage costs but also exposes companies to compliance risks, as sensitive elements (PII, PHI) may be buried without controls.



Unstructured Data Visibility and Discovery in NAS Environments

Before querying NAS metadata in platforms like Databricks, the first step is to gain clear visibility into what data exists, where, and who owns it. This is where tools like **Komprise Analysis** come into play. Komprise provides metadata harvesting and indexing across file servers, NAS, and object stores, producing structured, queryable metadata without moving data immediately.

This discovery layer is critical because:

- It reveals file types, sizes, ages, access patterns, and ownership metadata.
- It identifies “mystery shares” with no clear owner or stale content candidates.
- It supports policy-driven analytics—what data is hot vs. cold, which files contain PII/PHI, etc.
- It informs data lifecycle management strategies by quantifying data growth trends and cost drivers.

By first indexing the NAS metadata with Komprise, enterprises can cleanse the attic boxes, enforce compliance, and optimize data placement—all without yet moving large volumes of data or disrupting user workflows.

Why Query NAS Metadata in Databricks?

Databricks, as a unified data analytics platform built on top of Apache Spark, provides powerful engines for big data processing, machine learning, and AI workflows. Increasingly, teams want to extend their capabilities to leverage the metadata baked in NAS environments, combining it with data lakehouse architectures powered by open standards like **Apache Iceberg**.

Querying NAS metadata inside Databricks offers multiple advantages:

- **Unified analytics:** Combine structured data from data warehouses and unstructured file metadata for richer operational insights.
- **AI readiness:** Prepare datasets for machine learning models that include metadata features such as last access patterns, file sizes, or ownership tags.
- **Cost management:** Use data-driven insights to recommend data tiering or archiving strategies based on usage and sensitivity.
- **Security and compliance:** Identify and report on files containing sensitive information to meet audit and governance requirements.

Leveraging Iceberg Metadata for NAS Files

Apache Iceberg is an open table format for huge analytic datasets that enables ACID transactions, schema evolution, and partitioning, making it ideal for data lakes and lakehouses. By surfacing NAS file metadata as Iceberg table structures, Databricks users can query this metadata using SQL with high performance and reliability.

For example, Komprise's Intelligent Data Management solution can export indexed metadata into formats compatible with data lakehouses and tools like Iceberg. This metadata includes:

- File path and name
- File size
- Creation, access, and modification timestamps
- File type and classification (e.g., media, documents)
- Ownership and permissions
- Sensitivity tags (PII, PHI)

Once this metadata is available as an Iceberg table in Databricks, analysts can run queries such as:

```
SELECT file_type, COUNT(*) AS file_count, SUM(file_size) as total_size_gb FROM nas_file_metadata WHERE last_accessed < DATE_SUB(CURRENT_DATE(), 365) GROUP BY file_type ORDER BY total_size_gb DESC;
```

This query provides immediate insights into old, infrequently accessed files by type—helping prioritize data lifecycle decisions.

Cost Drivers in NAS Data Management

Unstructured data storage is not just a space problem—it's a cost problem that spans multiple facets:

Cost Driver	Description	Impact
Storage	Primary costs of storing data on NAS, including SSD, HDD tiers	Direct proportional to stored volume and tier
Backup and Replication	Additional storage and bandwidth for copies to meet RTO/RPO goals	Often 2-3x primary data volumes—major cost multiplier
Disaster Recovery (DR)	Off-site replication for resiliency	Extra infrastructure and operational costs
Security and Compliance	Costs around encryption, access controls, governance audits, and potential breach fines	High risk and regulatory penalties if unmanaged, especially for PII/PHI

Gartner's research underscores the risk of exponentially rising costs if unstructured data sprawl isn't addressed early. By querying and analyzing NAS metadata within Databricks, organizations can visualize cost drivers and identify cost-saving measures such as archiving cold data or deleting redundant files.

Security and Compliance Exposure in Dark Data

Dark data often contains hidden sensitive information—Personally Identifiable Information (PII), Protected Health Information (PHI), or corporate intellectual property—that if left unmanaged increases exposure to breaches, audits, and hefty fines under HIPAA, GDPR, and similar regulations.

Without metadata visibility, security teams operate blind to:

- Where sensitive files are stored
- Who accessed or modified files
- Legacy data beyond retention policies

Komprise Intelligent Data Management integrates with Databricks to tag and surface sensitive metadata at scale, enabling proactive governance within the lakehouse architecture. This unification minimizes breach risk and supports compliance auditing by making dark data visible and accountable.

The Role of DataStackHub and the Data Ecosystem

DataStackHub is a valuable resource for discovering tools [orphaned data cleanup](#) and best practices around data infrastructure, including NAS data analytics and lakehouse integration. Their curated ecosystem highlights how companies combine metadata management, governance, and analytics platforms like Databricks with data movers such as Komprise to orchestrate end-to-end data intelligence.

For data engineering teams steering AI projects, referencing DataStackHub helps in selecting interoperable tools that respect governance while driving value from dark data assets.

Bringing It All Together: A Practical Workflow

1. **Inventory and Index NAS Data:** Deploy Komprise Analysis to scan NAS shares, identify mystery shares, and extract file metadata.
2. **Classify and Tag Data:** Use Komprise Intelligent Data Management to apply sensitivity tags, ownership, and lifecycle policies.
3. **Export Metadata to Data Lakehouse:** Transform indexed metadata into Apache Iceberg tables residing in your lakehouse storage.
4. **Query in Databricks:** Utilize Databricks SQL or Spark to query the Iceberg tables containing NAS metadata for analytics, AI, or compliance reporting.
5. **Create Automated Workflows:** Integrate insights with data lifecycle management tools to archive, tier, or remediate based on metadata-driven triggers.

This workflow transforms dark, unstructured NAS data into a strategic, governed asset powering analytics and innovation.

Conclusion

Dark data lurking in NAS environments can no longer be ignored. Its mounting storage costs, hidden risks, and untapped insights demand structured approaches to metadata visibility and governance. By leveraging Komprise Analysis and Intelligent Data Management to index and classify NAS metadata, exporting this metadata into Iceberg table formats for Databricks, organizations gain critical unstructured data visibility within modern lakehouse architectures.

Your attic boxes of data need clear ownership, smart policies, and analytics-driven oversight—not just vague “cloud migration” buzzwords. Combining metadata management with Databricks’ powerful query engine provides an actionable path forward to transform NAS dark data from a compliance liability into a competitive advantage.

For organizations ready to move beyond static file servers and mystery shares, embracing this metadata-driven lakehouse approach with trusted tools from Komprise and insights from Gartner and DataStackHub offers both governance and innovation in one package.

Have you started querying your NAS metadata in Databricks yet? What challenges or successes are you seeing? Share your experience in the comments below or reach out to discuss how to take the next step toward smarter data management.