

In today's data-driven world, organizations grapple with ever-growing volumes of unstructured data scattered across file servers, NAS devices, object storage, and cloud buckets. This "dark data" — data that remains largely unused and unmanaged — accumulates quietly, incurring costs and increasing risk. For businesses leveraging **Snowflake** for analytics, a common challenge is gaining visibility into file metadata without the costly and complex process of moving large datasets around.

In this article, we'll explore what dark data is, why it accumulates, and how companies can achieve unstructured data visibility through metadata querying in Snowflake — leveraging technologies like *Apache Iceberg* tables and intelligent data management platforms such as **Komprise**. Along the way, we'll reference industry insights from **Gartner** and community resources like **DataStackHub** to provide [RAG pipelines](#) a well-rounded perspective.

What is Dark Data and Why Does It Accumulate?

Dark data refers to information that is collected, processed, and stored but not actively used for business decision-making, analytics, or improving operations. According to Gartner, dark data can represent up to 90% of an organization's data, and its unmanaged presence leads to escalating costs and security risks.

Why Does Dark Data Build Up?

- **Legacy storage sprawl:** Organizations retain old file shares and network-attached storage (NAS) systems for years due to compliance or operational inertia.
- **Lack of ownership:** Files often get orphaned as employees leave or projects close, creating "mystery shares" nobody claims ownership of.
- **Backup and replication:** Copying large datasets repeatedly for backup and disaster recovery inflates data volumes exponentially.
- **Compliance hold:** Regulatory mandates may require retaining files containing sensitive information (PII, PHI), even if the data is no longer actively useful.
- **Unstructured data growth:** File formats ranging from PDFs and images to videos and emails proliferate rapidly, often without classification or tagging.

This accumulation results in what I like to call "attic boxes" of data — files tucked away, forgotten, yet still incurring rent every month.

Understanding Unstructured Data Visibility and Discovery

Visibility into unstructured data is foundational for good governance, cost control, and reducing risk exposure. Simply knowing what files exist, their age, size, type, and access patterns is the first step toward intelligent data management.

Many organizations make the mistake of focusing immediately on tools or cloud solutions without first answering the critical question: *Who owns this data?* Once ownership and data context are understood, companies can apply technologies like **Komprise Intelligent Data Management** to analyze usage patterns and metadata at scale.



Metadata Tables and Apache Iceberg in Snowflake

Snowflake has emerged as a dominant analytics platform, widely used for structured data warehouses. But recently, Snowflake's support for external tables and integration with open table formats like **Apache Iceberg** has opened paths to query unstructured data metadata without ingesting or moving entire files.

Apache Iceberg is an open table format built for huge analytic datasets. It helps organize and maintain metadata about data files in object stores or file systems, allowing query engines like Snowflake to read table metadata efficiently.

By creating **metadata tables** over Apache Iceberg-managed directories, you can:

- Query file-level metadata (size, modification time, format) via SQL in Snowflake.
- Filter files according to attributes without scanning or moving the data itself.
- Integrate metadata information with other Snowflake tables for richer analytics.

This approach lets organizations gain visibility into unstructured data “attics” without expensive, risky, and time-consuming migrations.

Cost Drivers Related to Dark Data

Hidden in dark data are some of the most significant cost drivers for IT budgets. These include:

Cost Category Details Impact Storage Paying for capacity to hold infrequently accessed files on expensive primary or secondary storage. Wasted spend on unused capacity. Backup and Replication Duplicating dark data sets for data protection. Storage overhead grows exponentially. Disaster Recovery (DR) Copying entire file shares offsite — often including unused/obsolete files. Higher DR costs and recovery complexity. Security and Compliance Unmanaged sensitive data (PII, PHI) risks breaches, fines, and operational damage. Potential regulatory fines and brand damage.

How Komprise Enables Metadata-Driven Intelligent Data Management

Platforms like **Komprise** have pioneered solutions that combine data analytics, metadata extraction, and automated lifecycle management — all without moving files upfront. Their **Komprise Analysis** tool performs deep dives into file metadata across heterogeneous sources (file shares, NAS, cloud buckets), helping organizations:

1. Discover and classify dark data at scale
2. Prioritize data for archiving, tiering, or deletion based on access patterns
3. Create metadata layers compatible with analytic environments like Snowflake
4. Apply policies ensuring secure, cost-effective data usage

With this approach, your Snowflake environment can query external Iceberg metadata tables populated by Komprise, combining analytics with operational data management — all while preserving original data locations.

Integrating Snowflake Metadata Queries with Industry Insights

Industry observers such as **Gartner** emphasize the growing importance of data observability and metadata-driven governance. Gartner's strategic recommendations align well with a metadata-first approach: before attempting data migration or AI/ML project integration, ensure you know what data exists, who owns it, and its value.

Resources like **DataStackHub** showcase best practices for data lakehouse architectures, including how Apache Iceberg and Snowflake enable metadata abstraction layers crucial for unified analytics.

Common Challenges to Anticipate

- **Metadata freshness:** Ensuring that Iceberg tables reflect current file system state requires disciplined maintenance and integration with data management tools.
- **Security integration:** Metadata querying mechanisms must incorporate access controls to avoid exposing sensitive file details inadvertently.
- **Ownership clarity:** Clear accountability and data stewardship remain organizational challenges beyond tooling.

Summary: Your Metadata Roadmap in Snowflake

Querying file metadata in Snowflake without moving files demands a mix of data governance mindset, metadata-centric tools, and leveraging open formats like Apache Iceberg. By taking ownership of dark data, building visibility with **Komprise Analysis**, and integrating metadata tables into Snowflake, organizations can reduce costs, lower risk, and prepare for trusted AI and analytics — all without disruptive migrations.

Next time you hear the buzzwords pushing “just move everything to the cloud,” remember: the real question is always “Who owns and understands your attic boxes?” Addressing that first unlocks tangible insights and savings.



Further Reading & Resources

- [Komprise Analysis Overview](#)
- [Komprise Intelligent Data Management](#)
- [Gartner Data and Analytics Summit Resources](#)
- [DataStackHub: Apache Iceberg with Snowflake](#)