

In the rush to harness artificial intelligence and machine learning capabilities, many organizations face a sneaky and costly problem: AI data ingestion processes pulling in large [how to find duplicate files](#) volumes of **junk files**, or what is commonly known as *dark data*. This dark data is not just a storage nuisance but a significant risk factor for security, compliance, and cloud cost overruns. If you've been wondering how to ensure your AI projects only work with **curated datasets** and avoid the overhead of irrelevant or potentially harmful data, you're not alone.

What is Dark Data and Why Does It Accumulate?

Simply put, **dark data** is the information collected by organizations but left unused and unanalyzed. This often consists of old archives, redundant files, outdated backups, or miscellaneous "junk files" — think of them as *attic boxes* stuffed with stuff nobody remembers owning. According to Gartner, dark data can make up to 60-70% of an organization's unstructured data.

Why does this stuff accumulate? Here are the main reasons:

- **Lack of Ownership:** When nobody explicitly claims responsibility for certain data, it just lingers.
- **Over-retention Policies:** Legal or compliance mandates sometimes force keeping everything "just in case."
- **Legacy Systems and Shadow IT:** Multiple departments and users create copies, set up unknown shares, or dump files without centralized management.
- **Unstructured Data Growth:** Files like documents, videos, emails, images grow exponentially and often without clear tagging or metadata.

This pile of dark data becomes a black hole for storage, backup, and replication. Worse, when AI workflows indiscriminately ingest data, these junk files bog down training and inference systems, increasing costs and reducing model quality.

Unstructured Data Visibility and Discovery — The First Step

Before you can exclude dark data from AI ingestion pipelines, you need a clear picture of what's there. It's like trying to curate a museum exhibit without knowing which boxes in the attic are fragile artifacts and which are just old holiday decorations.

This is where tools like **Komprise Analysis** come into play. Komprise offers deep insights into your unstructured data spread across file servers, NAS, and cloud storage. It scans and categorizes data by age, type, owner, and usage patterns — essential clues to identify data that can be safely excluded from AI ingestion.

Visibility tools help answer crucial questions:

- What percentage of data is truly active and relevant to AI model needs?
- Which shares, folders, or buckets are filled with large volumes of stale files or "mystery shares"?
- Where do sensitive files containing **PII** or **PHI** reside, and are they securely protected?

Understanding these details empowers data owners and security teams to act decisively before AI pipelines blindly pull in everything.

Cost Drivers Tied to AI Ingestion of Dark Data

Many organizations are surprised when their cloud storage and processing bills balloon after enabling AI workflows. The root cause often traces back to ingesting unnecessary, uncurated data:



Cost Driver Impact of Dark Data Why It Matters for AI Ingestion Storage Excess storage consumed by irrelevant files inflates cloud bills and onsite storage maintenance. More data ingested means longer processing times and higher retention costs for AI datasets. **Backup and Replication** Unnecessary backup cycles and replication overhead multiply costs and slow disaster recovery. Ingesting dark data can increase backup windows and replicate irrelevant historical files. **Disaster Recovery (DR)** Storing dark data extends DR recovery times and impacts RTO/RPO objectives. Critical when AI platforms must quickly restore high-quality curated training data. **Security and Compliance Exposure** Dark data often harbors sensitive information unaccounted for in governance policies. Ingestion of PII/PHI into uncontrolled AI systems risks breaches, fines, and reputational damage.

Ignoring these cost and compliance aspects can turn your AI journey into a money pit and regulatory risk.

How to Ensure AI Data Ingestion Uses Curated Datasets — Best Practices

Eliminating junk files and dark data from AI ingestion requires a combination of strategy, tools, and governance:

1. **Identify Data Owners:** Before any tool or platform, ask “Who owns this data?” This prevents your AI from digging deep into “mystery shares” and orphaned folders. Komprise emphasizes data ownership in its Intelligent Data Management approach, helping build accountability.
2. **Deploy Visibility and Classification Tools:** Use analytical platforms like **Komprise Analysis** to map data landscapes. This process identifies data types, sensitivity, and usage patterns.
3. **Establish Data Policies:** With visibility in hand, work with legal, security, and business units to decide which data sets can be ingested for AI and which must be excluded as dark data.
4. **Leverage Intelligent Data Management:** Use tools like **Komprise Intelligent Data Management** that automate data movement, archiving, and deletion based on lifecycle policies. This reduces dark data buildup over time.
5. **Integrate Security and Compliance Controls:** Ensure AI ingestion pipelines respect data governance rules, including masking or outright excluding PII/PHI. Collaboration with platforms like DataStackHub can help monitor compliance during data workflows.

6. **Continuously Monitor and Reassess:** Dark data is a moving target. Regularly revisit your datasets and update AI ingestion filters to adapt to evolving business and compliance needs.

Why Komprise, Gartner, and DataStackHub Are Critical Allies

Komprise delivers scalable unstructured data management solutions designed for defining and maintaining curated datasets. Its dual tools — *Komprise Analysis* for visibility and *Intelligent Data Management* for action — help companies proactively reduce dark data, thus safeguarding AI pipelines.

Gartner

DataStackHub offers data workflow automation with integrated compliance checks, creating an environment where curated datasets can flow safely into AI and analytics platforms without dragging unnecessary risk along.

Final Thoughts: Don't Let Your AI Ingest the Attic's Junk Boxes

AI and machine learning promise powerful insights — but these insights depend on the quality of your data. Letting AI ingestion processes blindly consume all data can backfire by increasing costs, exposing sensitive information, and muddying model outcomes. Addressing this starts with enhanced visibility and control over unstructured data to exclude dark data and curate datasets tailored to AI requirements.



Approaching this challenge with a data-centric mindset—“Who owns it?” before “What tool can I use?”—and leveraging best-in-class tools like Komprise combined with solid governance frameworks supported by research from Gartner and automation via DataStackHub will set your AI projects up for success. After all, you wouldn't put every old attic box into an exhibition — your AI datasets deserve the same thoughtful curation.

Ready to reduce your AI ingestion junk files and transform dark data into AI assets? Start with Komprise Analysis for visibility, then apply Intelligent Data Management to keep your curated datasets healthy, secure, and cost-effective.