

[White Paper]

Komprise Data Intelligence & Orchestration for AI

Solving the First-Mile Gap of
Unstructured Data Preparation

Table of Contents

Overview: Scaling GenAI to Vertical, Enterprise-Wide Programs3

The ROT and Dark Data Problem3

Common Approaches for AI Data Prep Aren’t Working4

The Four Requirements of Unstructured Data Preparation for AI5

 1. Global Data Discovery and Intelligence 5

 2. Iterative, Custom Data Tagging and Classification 6

 3. Policy Driven Orchestration..... 7

 4. Efficiently Ingest, Execute and Govern..... 8

Komprise AI Data Management8

Conclusion10

About Komprise10

Overview: Scaling GenAI to Vertical, Enterprise-Wide Programs

AI is moving from ad hoc, small use cases and general productivity to enterprise-wide line of business strategies. Vertical AI use cases such as claims processing in insurance, clinical trial analysis in biotech, improved patient diagnosis in healthcare and legal case analysis depend entirely on the unstructured data layer: documents, emails, instrument and research data, images, video, sensor telemetry, scanned records. If the AI can't reach the right data reliably, these burgeoning use cases will fail.

The struggle is evident, with 90% of vertical, function-specific use cases stuck in pilot mode, according to [research from McKinsey](#)¹. The consulting firm calls this the "Gen AI Paradox": nearly 8 in 10 companies report using AI, yet just as many report no significant bottom-line impact.

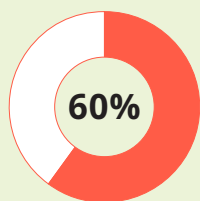
The ROT and Dark Data Problem

So what is keeping enterprise AI projects in pilot mode? There's not a single answer to this question, given the complexities of setting up the right infrastructure, acquiring sufficient budget, training or hiring requisite talent and getting alignment on the most promising use cases for AI. Yet these are perennial IT challenges with new technology.

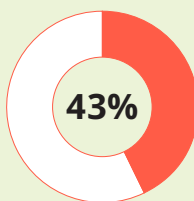
The elephant in the room: petabytes of disorganized, unclassified and siloed data, obscure data context and intelligence that is needed to fill AI pipelines.

Less than 1% of unstructured data is in an AI-ready format, [according to IBM](#)². This is because it's too large and difficult to move and too noisy. IT doesn't know what's in it nor how much is redundant, obsolete and trivial (ROT). When an AI model retrieves from a corpus filled with this low-quality data — common in enterprise NAS environments, cloud object stores and backup drives — it produces hallucinations and unreliable outputs.

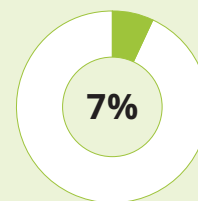
Enterprise Data is Not AI-Ready



Through 2026, organizations will abandon 60% of AI projects that lack AI-ready data foundations. -- [Gartner](#)



Infrastructure, skills, and data readiness (43%) are the biggest obstacles to AI plans in global enterprises. -- [Drexel University's LeBow College of Business and Precisely](#)



Only 7% of organizations say their data is completely ready for AI adoption. -- [Cloudera and Harvard Business Review](#)

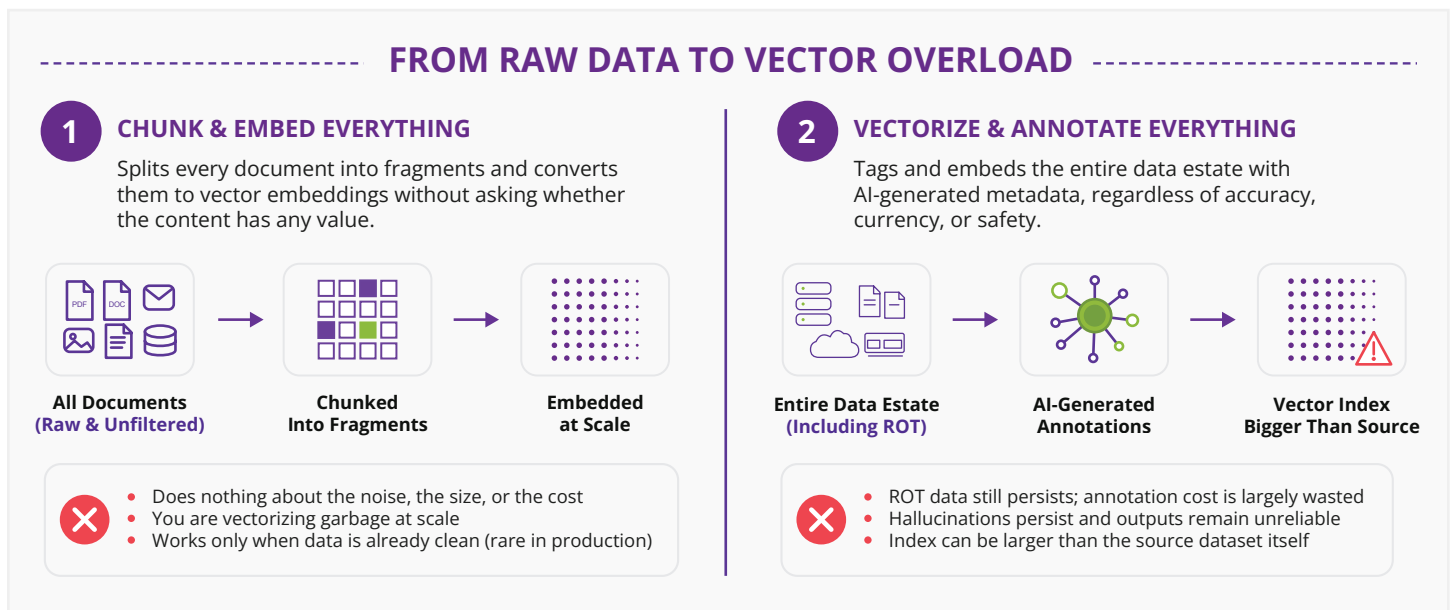
¹ <https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage>

² <https://www.ibm.com/new/announcements/ibm-and-unstructured-io-partner-to-accelerate-ai-ready-data-in-watsonx-data>

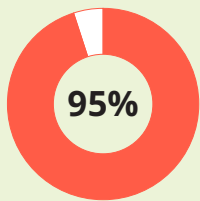
Common Approaches for AI Data Prep Aren't Working

Two tactics dominate early AI data preparation efforts, and both break at enterprise scale.

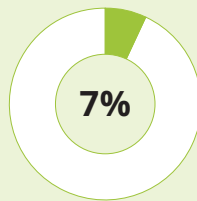
1. **Chunk and embed** splits every document into fragments and converts them to vector embeddings without asking whether the content has any value. It does nothing about the noise, the size, or the cost. You are vectorizing garbage at scale. It works only when the underlying data is already clean, which is typical of a small pilot and rare across a production estate.
2. **Vectorize and annotate everything** tags and embeds the entire data estate with AI-generated metadata, regardless of whether that metadata is accurate, current, or safe to expose. It fails the same way. The ROT data is still there, the annotation cost is largely wasted, hallucinations persist, and the resulting index can be larger than the source dataset itself.



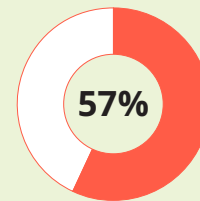
Enterprise AI Adoption is Growing but ROI is Still Lacking



Roughly 95% of generative AI pilots delivered no measurable profit. -- [MIT Project NANDA](#), [The GenAI Divide: State of AI in Business](#)



Only 7% of companies have fully scaled AI across their organization. -- [McKinsey](#)



57% of AI projects aren't delivering their objectives and nearly all (94%) struggle to manage unstructured data -- [Nasuni](#)

The Four Requirements of Unstructured Data Preparation for AI

Enterprise AI needs unstructured data intelligence and preparation to support continuous, AI-ready data. The right method must scale to petabytes, span data estates both on premises and cloud, and cut AI token and compute costs by reducing the estate to just the right data for each use case.

Four requirements, all supported in the **Komprise Intelligent Data Management** platform, make that possible:

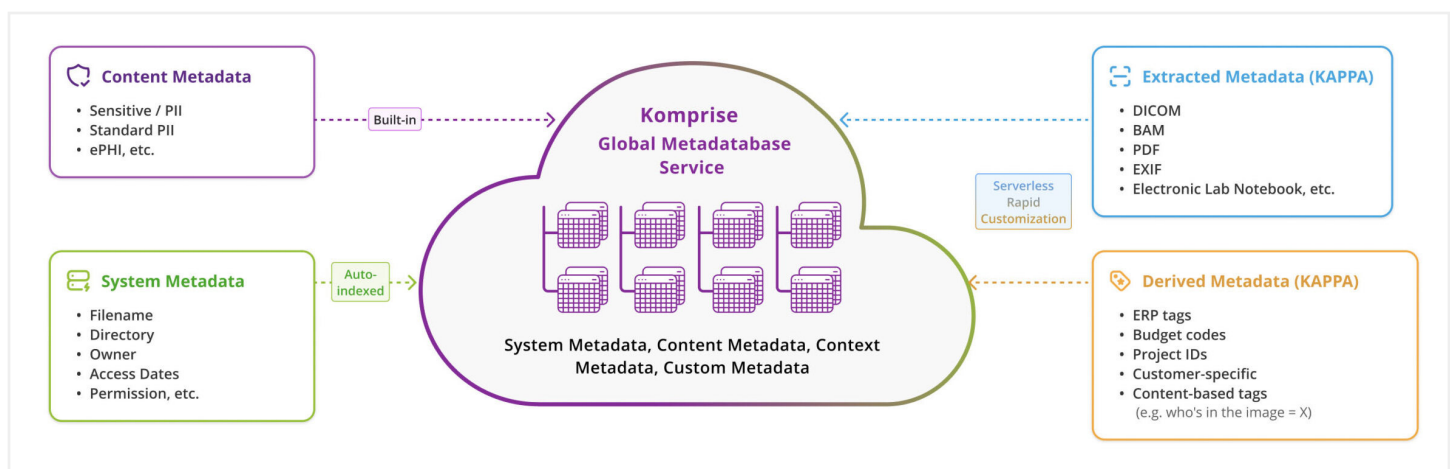


1 Global Data Discovery and Intelligence

First, organizations need to get one version of the truth: discover every data source across the estate, weed out duplicate and low-quality data, and build a schema that is query ready from the metadata. That schema must work across vendors and standards to deliver unified context no matter where the data lives. An AI team should be able to locate specific research files across a hybrid estate spanning multiple NAS vendors and cloud providers without running a separate query per environment.

What's needed is in-place indexing into a single searchable layer, and the [Komprise Global Metadatabase](#) provides it:

- Many vendors claim global discovery, but that pertains only to what their products store, when most enterprises store data across multiple storage vendors and clouds.
- Others force data to be loaded before indexing. Komprise indexes file and object across all storage in place, delivering unified visibility and consistent policy enforcement from a single platform.
- Stateless Komprise Observers bring the indexing presence next to the data for optimal performance and data security, since the data remains in place.
- The [Komprise elastic, distributed architecture](#) has scanned hundreds of petabytes without impacting production performance and while bringing compute to the data.



Unstructured data lacks context and structure and without that, it's not searchable and usable for AI. Once data is indexed and visible, it's time to enrich basic file system with metadata for narrow classification needs and as new data is created.

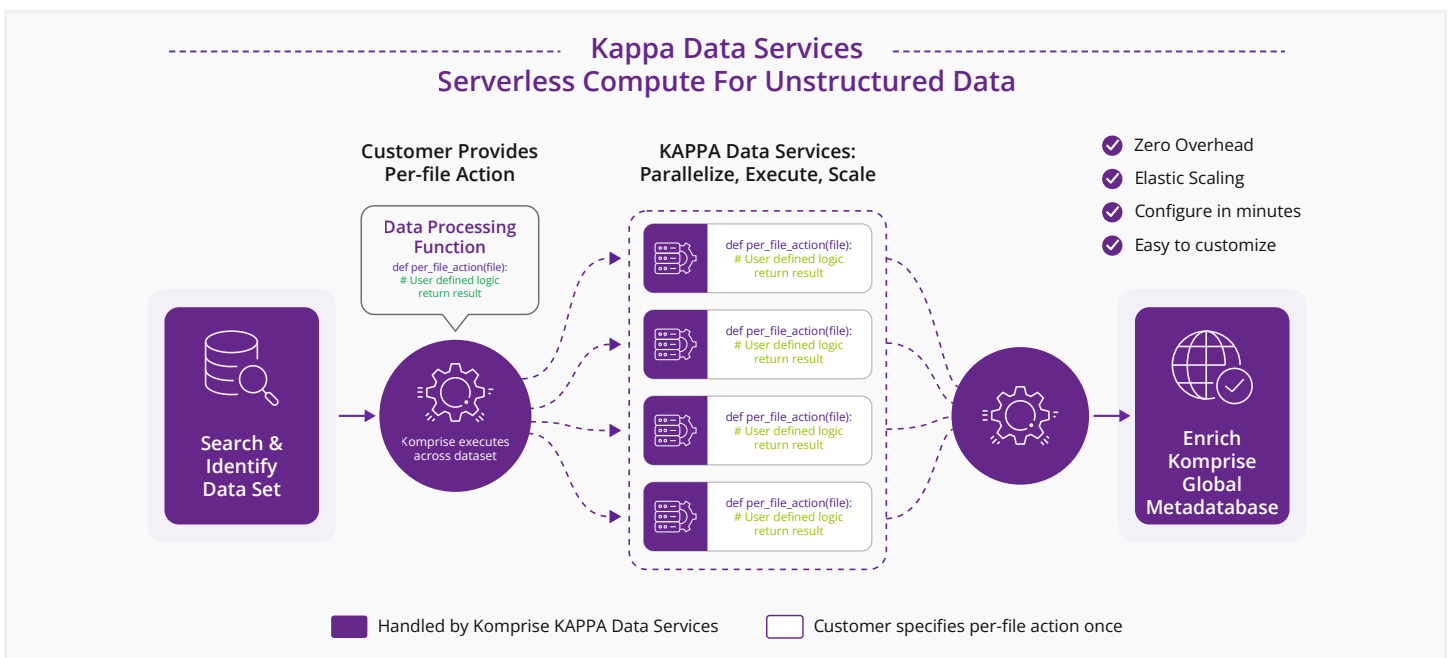
Analyzing system metadata such as time of creation and last access, owner and file names is the first step in purging too old or redundant data. Then organizations can extract industry and custom metadata for precise classification and filtering.

The iteration matters: tier off data older than six months, scan the smaller remaining set for PII, confine what is sensitive, then extract domain metadata to isolate exactly what a model needs.

Using Komprise, Duquesne University achieved a 99% faster process for image search and tagging to support an important fundraising campaign.

Komprise delivers continuous, iterative metadata enrichment and classification across storage and cloud vendors to support RAG pipelines.

- [Komprise AI Preparation and Process Automation \(KAPPA\)](#) is a service that delivers project and industry-specific metadata extraction such as DICOM attributes in medical images, OSDU metadata for energy industry, or electronic lab notebook (ELN) project codes.
- KAPPA handles the full processing stack automatically: provisioning compute via Komprise Observers that reside next to the data and scaling to iterate the function across billions of files.
- Komprise integrates with any third-party data classification and filtering tool.
- Built-in sensitive data classification in [Smart Data Workflows](#) identifies PII, PHI, and regulated content across the organization.
- The Global Metadatabase reduces a petabyte-scale estate to precise datasets for specific use cases and requirements. Wherever data moves, the metadata persists.
- Komprise delivers the Global Metadatabase as a managed service: no setup required.

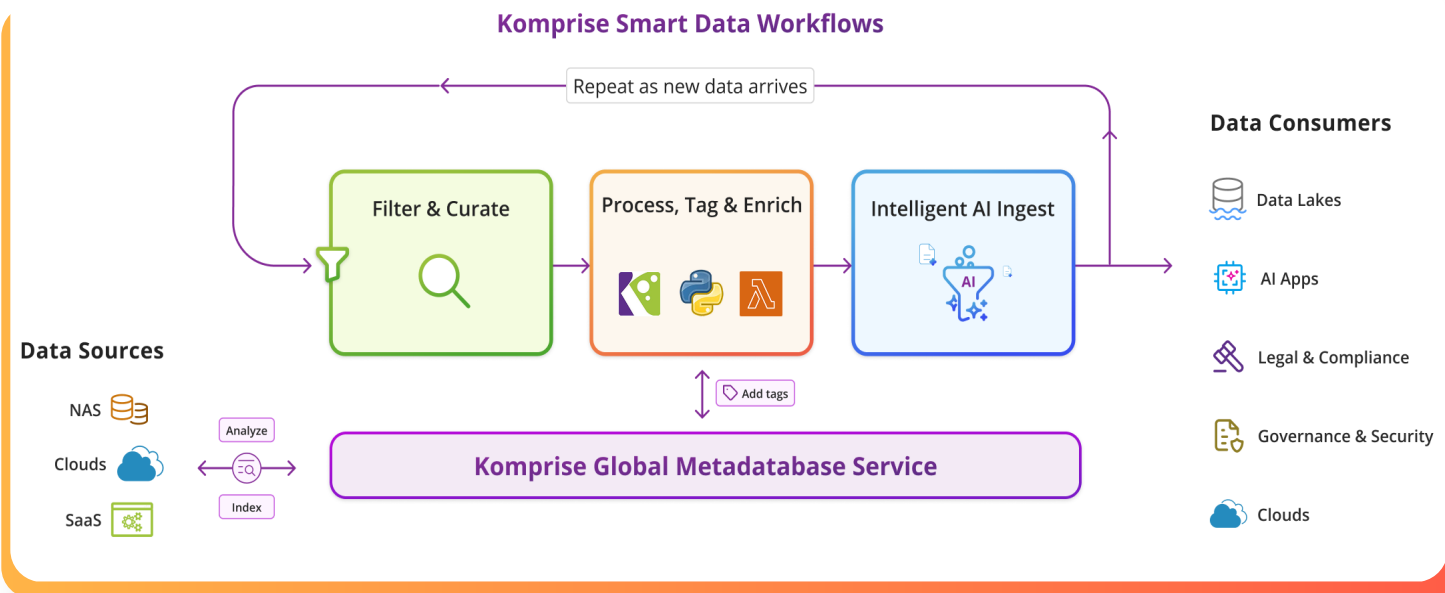


3 Policy Driven Orchestration

Discovery, classification and ingestion to AI are not one-time events. As new data lands and use cases shift, enrichment and delivery must run continuously by policy. Multiple workflows may be running in parallel and need to be orchestrated. The third requirement is automated orchestration that connects discovery to enrichment to delivery with built-in governance. It is imperative that the Global Metadatabase persists the latest updates from one workflow to another so the same data is never reprocessed, your data intelligence is always current, and you keep costs down.

Komprise Smart Data Workflows orchestrate the execution of KAPPA functions and other actions at scale, with no scripting required.

- Each workflow runs on a schedule, continuously finds new or updated files that match ingestion criteria, applies enrichment tags, runs governance checks, and delivers fresh datasets to AI platforms without manual intervention.
- Enriched tags are written back to the Global Metadatabase as searchable file characteristics, avoiding AI services from reprocessing the same data as datasets grow.
- Context compounds with every pass, producing progressively richer inputs for retrieval and training.
- IT can monitor hundreds of workflows across different environments from a single interface, providing centralized control as the number of use cases scales.



4

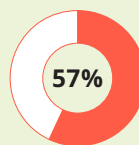
Efficiently Ingest, Execute and Govern

The final requirement is to move only what you must. Unstructured data is expensive and cumbersome to move; a growing best practice is to leave it in place when only metadata is needed. Increasingly, data engineering teams working in platforms such as Databricks and Snowflake want to bring unstructured data into their projects without excess overhead.

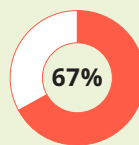
Along with cost management, AI data governance must be integrated functionality for any AI data workflow, no matter the source and target.

AI Data: Risky, Ungoverned

Classification is the top challenge in preparing unstructured data for AI. -- [Komprise 2026 State of Unstructured Data Management](#).



of leaders view data reliability as a key barrier to moving AI projects from pilots to production. -- [Informatica](#).

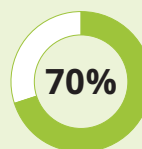


of executives believe their company has already suffered a data leak or breach due to unapproved AI tools. -- [Writer, Enterprise AI adoption 2026 report](#).

Komprise AI Data Management

Komprise brings rapid, highly-efficient capabilities to bring the right unstructured data to AI and lakehouses with 70-80% lower overhead and costs.

- [Komprise Transparent File Tables](#) brings metadata to the lakehouse, without moving the data, through Apache Iceberg tables and with access controls intact. Lakehouses see unstructured data as if it were structured, described in rows and columns.
- [Komprise Deep Analytics](#) curates and eliminates noise before moving data to lakehouses or AI through flexible custom queries while Smart Data Workflows allows for sensitive data detection and mitigation.
- [Komprise Intelligent AI Ingest](#) eliminates **70%+ of unstructured data noise** before it reaches AI, reducing GPU compute and token costs and improving model accuracy.
- Automated workflows delete leftover copies later to reclaim cloud storage and **reduce costs by as much as 90%**.



Komprise Intelligent AI Ingest eliminates 70%+ of unstructured data noise



2x-10x
the speed of standard transfer tools



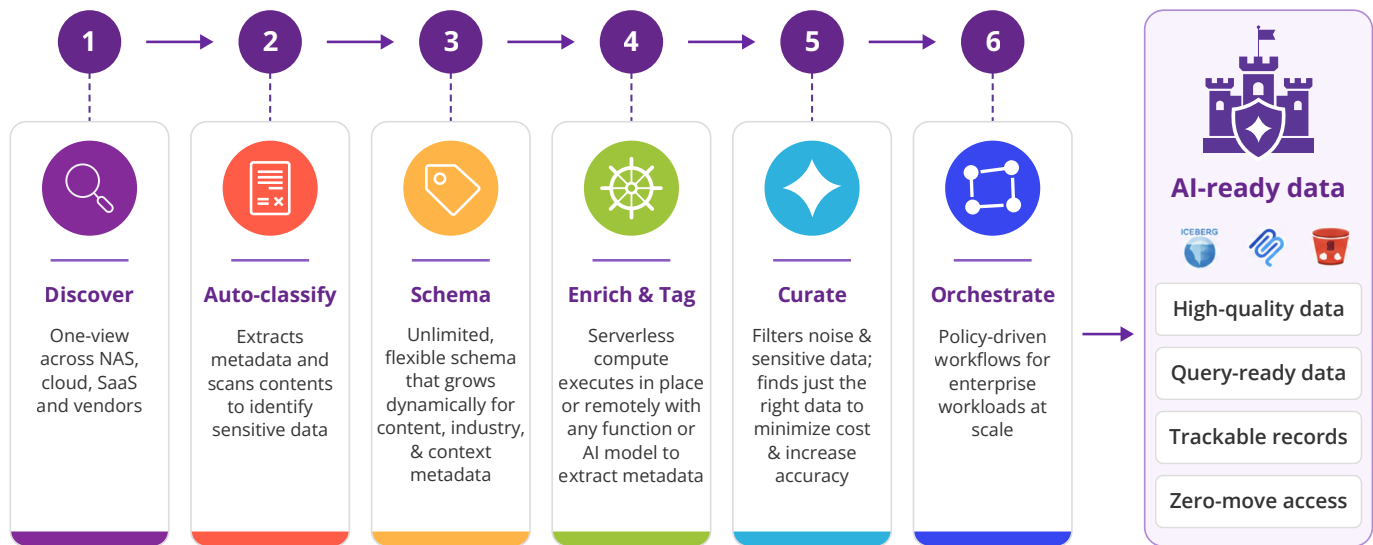
Reduce AI costs by as much as 90% with Komprise



Most of our data is unstructured, untagged and nobody knows what's in there or how frequently it's being accessed. The storage vendors haven't provided good tooling to look inside that data set. Komprise gives us visibility into this data, it can tag that data, and then have it ready for AI."

Senior IT Architect, large U.S. academic health system.

Komprise Delivers AI-Ready Data



Conclusion

Enterprise AI stalls on the unstructured data layer, not on the model. The path from pilot to production runs through data preparation built on four requirements: global discovery, iterative annotation and classification, policy driven orchestration, and governed ingest.

Komprise Intelligent Data Management delivers all four across every silo, on premises and in the cloud. The result is a smaller, cleaner, governed dataset that lowers token and GPU cost, reduces hallucinations, and provides a data foundation your line of business teams can trust. Data dumping does not scale nor deliver the ROI that the C-suite expects. Disciplined data preparation does.

About Komprise

Komprise unlocks unstructured data for AI while optimizing data placement to cut storage costs. With Komprise [Intelligent Data Management](#), enterprise IT gets global data intelligence, policy-driven data orchestration, fast data mobilization, and governance while data teams get AI-ready data that is easy to query. Enterprises use Komprise to discover, classify, curate, tier and migrate file and object data across any storage, cloud or AI environment.



Komprise, Inc.
307 Orchard City Dr #360
Campbell, California 95008

For more information:
Call: 1-888-995-0290
Email: info@komprise.com
Visit: komprise.com

For media requests email
marketing@komprise.com
© Komprise, Inc. All rights reserved.