

Survey Report

The Komprise Unstructured Data for AI & Lakehouses Survey

Enterprises Are Reporting AI Returns Before Their Data Is Ready

October 2026

Table of Contents

01	Executive Summary	03
02	Key Statistics	04
03	Section I - The State of Enterprise AI	05
04	Section II - The State of Unstructured Data Readiness	07
05	Section III - Lakehouses & Unstructured Data	09
06	Section IV - Execution	11
07	Top 5 Takeaways	12
08	About Komprise	13

Executive Summary

Enterprise AI programs are running ahead of the data underneath them, at a time when AI's safety and validity are under fire more than ever before. **The Komprise Unstructured Data for AI & Lakehouses Survey** shows a wide gap between how mature enterprises believe their AI programs are and how much of their unstructured file data is prepared to feed those programs.

Two-thirds of respondents (68%) describe their organization as mature or maturing on AI, defined as setting standards, automating business processes, agentic workflows and applying governance models. **Yet only 7% have classified and cleaned the majority of their file data.** As well, 57% say they have half or less of their data ready and 13% cannot say how much is ready at all.

To generate measurable and meaningful ROI on AI, enterprise IT and AI leaders know they must understand and continually improve data quality, especially with their wildly diverse unstructured (file and object) data assets. The connection is evident, given that **only 21%–26% of enterprises have quantified ROI** on operations and revenue growth.

Concerning the largest gaps in data readiness for AI, respondents emphasized finding the right data and cleaning up redundant, obsolete and trivial (ROT) data across petabytes (49%) and extracting and enriching metadata to build a schema (42%).

The lakehouse readiness picture is equally nebulous. Unstructured file data lacks the schema that analytics platforms require, and **only 4% of organizations have built a schema for more than 70% of their file data and delivered it to lakehouse and analytics teams.**

Outside research affirms the complexity of getting enterprise data filtered, classified and governed for AI. [Gartner found](#) that 63% of organizations either lack or are unsure they have the right data management practices for AI.

Raising the stakes is evidence that enterprises are increasing AI-related spending: [BCG found](#) companies plan to invest about 1.7% of revenue in AI in 2026, roughly double the 0.8% of 2025.



About this survey

In September 2026, Komprise surveyed **500 technical decision makers** at organizations with **1,000 or more employees**, from IT managers and data engineers to directors of AI, CIOs and CTOs, to ascertain opinions on AI maturity and how data management practices are evolving.

Key Statistics



AI maturity

67% say they are maturing or mature in AI, with standards, strategy and governance forming and the most mature deploying agentic workflows connected across departments.



AI ROI

Customer Service is the highest area for ROI, with 63% saying they've achieved returns and 38% having quantified them. Quantified ROI for employee productivity and decision speed clocked 28% each, followed by revenue growth at 26% and operational savings at 21%.



Unstructured data readiness

Only 7% of organizations have 70% or more of their data cleaned, prepared and classified for AI, while the majority (57%) have 50% or less ready.



Largest gaps in readiness

The majority (nearly 50%) cite finding the right data for a project and cleaning up ROT, followed by metadata extraction and enrichment (42%).



Schema for lakehouses

Only 4% have more than 70% of their unstructured data ready, with schema, for analysis in a lakehouse and only 24% have more than half ready and available.



The most popular data lakehouse

Nearly half (49%) use Microsoft Fabric, followed by Google BigQuery (32%), Amazon Redshift (28%), Databricks (27%) and Snowflake (22%).



Data preparation tactics

Over 70% do not yet have the capability for advanced classification techniques that discover stale data and allow users to search and find the right data for specific AI use cases.



Data first strategy

Making unstructured data available for query in AI and lakehouses was deemed a priority (46%) over getting AI ready infrastructure in place (34%).

Section I: The State of Enterprise AI

In the labs of top AI companies such as Anthropic, OpenAI and China's Moonshot AI, there is a no-holds-barred race for technological superiority. Faster, more powerful models and new features are being launched at an unrelenting pace, while information workers are barely scratching the surface of models three versions behind.

Even in the halls of leading U.S. tech companies which have been the fastest adopters of AI, confusion and a bit of fear reigns. But there is no question that AI has entered the corporate mainstream, compared to just three years ago when ChatGPT was a mere play tool for answering random questions in seconds.

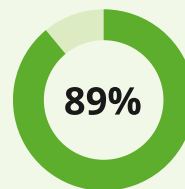
CXOs remain concerned and cautious, rightly so, about AI. Protecting internal data and preventing harmful hallucinations and false outcomes is the top priority. Yet our survey shows that despite these barriers, enterprise plans are progressing with an optimistic view of maturity: **only 11% characterize their organization as immature and without a strategy nor sanctioned tools for employees.**

In fact, **nearly 50% are “maturing,”** in terms of setting standards, automating processes and applying governance models. **One-fifth say their organization is “mature,”** defined as having AI deployed across departments with agentic workflows and unified security and governance.

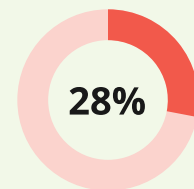
When you compare these findings with reported ROI from AI, there is a noticeable gap. Across employee productivity, operational savings, revenue growth, decision speed and customer service, only **an average 28% of organizations have quantified ROI from AI.** The highest reported returns are in customer service (38%), which runs on distinct, clean data sets such as chat and voice transcripts and tickets.

HERE IS THE CONFLICT

89% say they are ready for AI, yet less than a third are able to quantify returns.



say they are ready for AI



on average have quantified ROI from AI across the five areas surveyed

This is our take:

- First, the Komprise survey did not focus on ROI and without a detailed scoring methodology, participants are likely more apt to self-rate as high in maturity;
- Second, data readiness does not have a one-to-one relationship with AI maturity. According to [Gartner](#), data availability and quality are top challenges in AI implementation regardless of AI maturity levels.

How would you characterize your organization's maturity with AI?

Maturing

We are beginning to establish standards and use AI more strategically, automating key business processes and applying governance models across the organization.



Departmental

Teams and departments run their own versions of GenAI tools for general productivity, ad hoc purposes only with no centralized approach.



Mature

AI is deployed across multiple interconnected departments with AI agent workflows and a unified security and governance strategy in place.



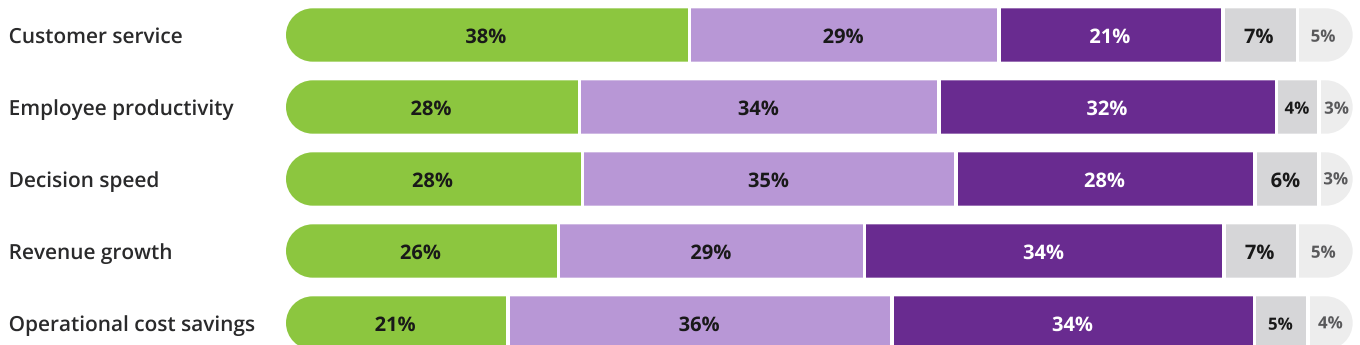
Immature

We don't have an AI strategy nor sanctioned AI tools for employees, but we know there is shadow AI occurring.



Where does your organization currently stand on ROI from AI investments using unstructured data?

● Achieved and quantified
 ● Achieved, not quantified
 ● Expected, not yet achieved
 ● Not expected
 ● Don't know



Figures may not total 100% due to rounding.



Top Takeaway

Data readiness does not resolve as organizations grow adoption of AI across departments and functions. In fact, the more AI workflows in progress, the greater the need to assess, scan, clean and prepare data continuously for evolving use cases. **Enterprise leaders focusing on AI can get ahead by investing in AI data management and readiness programs.**

Section II: The State of Unstructured Data Readiness

For enterprises, long-term AI value is not from employee productivity benefits such as aggregating and summarizing data, creating professional documents and writing content in seconds. The higher potential: using internal data such as research, contracts, market analysis, customer records and messages, sales data, field reports and more to generate intelligence that can move the business forward and grow market share and customer satisfaction. Yet since most (at least 80%) enterprise data is unstructured and lacking context, this outcome remains elusive for now.

What shifts the dynamic is when IT and data teams can efficiently generate custom metadata denoting file contents and sensitivity. This extra data allows users to discover the exact data sets they need for AI with a keyword search, while excluding any data that is banned from ingestion, such as PII or IP.

Feeding just the right data to AI pipelines is not only critical for quality outcomes and ROI but also to control spending; AI compute and storage consume higher percentages of the IT budget every year.

93%

say their unstructured data is mostly not ready for AI

[Gartner forecasts](#) worldwide AI spending will reach roughly \$2.59 trillion in 2026, up about 47% year over year. Enterprise generative-AI spending more than tripled from an estimated \$11.5B in 2024 to \$37B in 2025, according to [Menlo Ventures](#).

Since data quality is a primary factor in AI success, we asked how much enterprise file data is classified, prepared and ready for AI. **Most (93%) reported that the lion's share of their data is not ready for AI, with nearly 60% saying that half or less is ready.**

This is not surprising. Until recent times, data engineers were primarily analyzing and cleaning data which was organically structured and organized in databases. The new world of unstructured data is amorphous and murky:

- Preparing unstructured data for AI is complex in a large enterprise which can easily store 10PB or more of data from wildly diverse systems (text, video, audio, image, sensor, industry-specific instruments) across on-premises, edge and cloud storage.
- IT organizations are struggling to determine how to preprocess, sanitize and add useful context required for this data to be leveraged for new value. The answer lies in the metadata.

SECTION II · UNSTRUCTURED DATA READINESS

Storage systems automatically generate metadata such as time stamps for creation, modification and access, file type, size and owner. While useful for data lifecycle management decisions, system metadata does little for AI and analytics.

Tools that can extract metadata tag files with information such as diagnosis or body part for a medical image, OSDU standard tags in oil and gas, quality scores in bioinformatics or copyright information in media files convert raw data into actionable information for AI. Furthermore, a global, searchable database of unstructured schema delivers a consistent, cohesive view of enterprise data across storage and clouds.

The ability to extract and enrich metadata to build unstructured data schema for AI is a top gap in AI data readiness, according to 42% of enterprises. An even greater pain point, encompassing schema needs, is finding the right data and cleaning up redundant, obsolete and trivial (ROT) data (49%).

AI data governance, a top concern in many surveys over the past two years, is the third pain point (41%), suggesting that tools and practices are evolving to better address sensitive data leakage to AI.



Top Takeaway

IT and data leaders can accelerate data readiness with metadata analysis, extraction and processes to manage metadata assets across the complete data lifecycle. This creates a foundation of trust, where AI users and engineers can easily find the most relevant, specific data for a project, without running up AI costs and risks.

What are the biggest gaps or pain points in data readiness for AI?

Finding the right data for my AI project and cleaning up low quality redundant, obsolete and trivial (ROT) data across petabytes



Metadata extraction/enrichment to get a schema for our unstructured data that can be exposed to lakehouses and AI



Governance and compliance including protecting sensitive and regulated data from exposure in AI and auditing outcomes and workflows



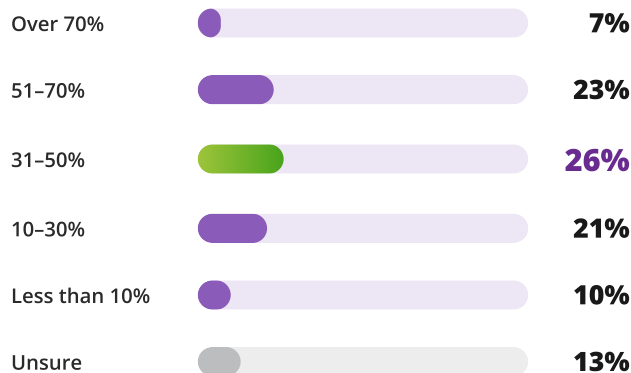
Visibility, aka knowing what data we have and where it is located



Presenting or copying scattered volumes and petabytes of file data to AI platforms



How much of your enterprise storage file data (NAS, archives, unstructured data stores, file shares, large image and videos, instrument data, etc.) have you cleaned, prepared, classified and made ready for AI?



Section III · Lakehouses & Unstructured Data

Data lakehouses have become the killer enterprise platform because they merge structured and semi-structured data and are becoming an efficient, central data pipeline for LLMs. They are poised to grow from \$10.33 billion in 2025 to \$12.58 billion in 2026, according to [Research & Markets](#).

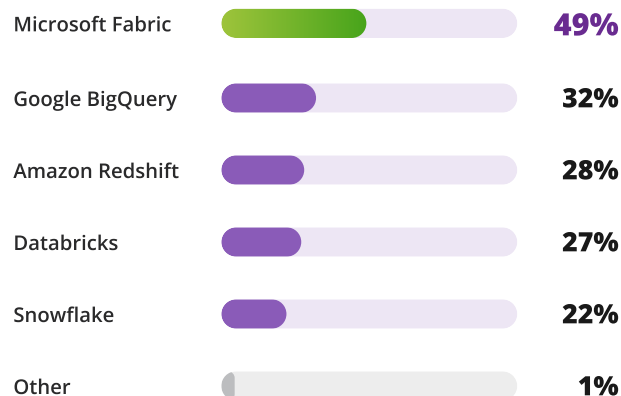
Enterprises have several mature data lakehouse platforms to choose from: the most popular in the Komprise survey is Microsoft Fabric, used by nearly half of respondents, followed by Google BigQuery (32%), Amazon Redshift (28%), Databricks (27%) and Snowflake (22%). On average, organizations are using roughly 1.6 lakehouses, which adds to the unstructured data preparation challenge explained next.

Powerful lakehouse tools are only as valuable as the data they ingest and enterprises have a long way to go when it comes to unstructured data.

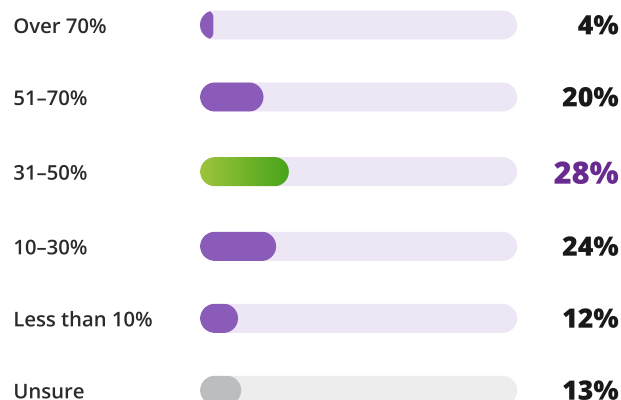
- Only 4% have most (70% or more) of their file and object data classified with schema;
- Less than a quarter (24%) have more than 50% of this data available with schema.

4% of unstructured data is ready for lakehouses

Which lakehouses do you use today?



Unstructured file data lacks the schema needed for data lakehouses. What percentage of your enterprise file data have you developed a schema for and made available to data lakehouse/analytics teams?



SECTION III · LAKEHOUSES & UNSTRUCTURED DATA

This reality is a barrier for the true promise of enterprise lakehouses as a natural repository for all data types and the means to deliver a single source of truth for traditional analytics and AI. Automated, scalable metadata extraction and enrichment with tagging across all storage and lakehouse silos helps close this gap.

A related IT challenge is delivering unstructured data access without the time and expense of copying and moving terabytes to petabytes of data to a lakehouse. On average, 33% of stored enterprise file data is classified as ROT (Redundant, Obsolete, or Trivial), according to the [Veritas Global Databerg Report](#).

If you add “dark data” or hidden, unorganized data, the range of useless data in an enterprise can exceed 80%, according to various estimates. It’s imperative to filter and segment this data before moving any of it to a lakehouse.



Key Takeaways

Lakehouses now anchor the enterprise data pipeline, and most enterprises run more than one. Yet only 4% have developed schema for most of their file data and just 24% have passed the halfway mark. Until unstructured data carries the metadata that lakehouses require, these platforms deliver a single source of truth for structured data only, eroding their usefulness for AI.

How important would it be to leave data in place and still query it from analytics, lakehouse and AI tools?

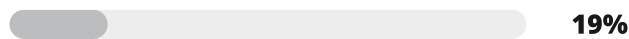
Very important to build dashboards and/or AI pipelines with structured and unstructured data



It's high on the list but behind getting our AI-ready GPU compute and storage in place



Unsure right now



Not important



By leveraging metadata, however, IT can make unstructured data visible in lakehouses without first moving it. Importing Apache Iceberg tables of metadata based on a user query can accomplish this from any lakehouse.

This zero-move capability is increasingly popular due to the size of unstructured data; lakehouse users can view and manipulate tables to decide which full data sets are required for ingestion rather than moving volumes of raw data first.

- **Most (46%) said it would be very important to query data from lakehouses or AI without moving it** to build dashboards with structured and unstructured data.
- Just over a third (34%) reported that getting AI-ready compute and storage in place was more important.

Section IV: Execution

Preparing and moving unstructured data to lakehouses and AI requires specialized tools and methods that are custom made for the unique requirements of large, diverse file data. Understanding data to aid search and discovery means first indexing and classifying it: scanning across silos for PII, IP and other sensitive content that typically cannot be exposed to AI tools and then tagging files for mitigation.

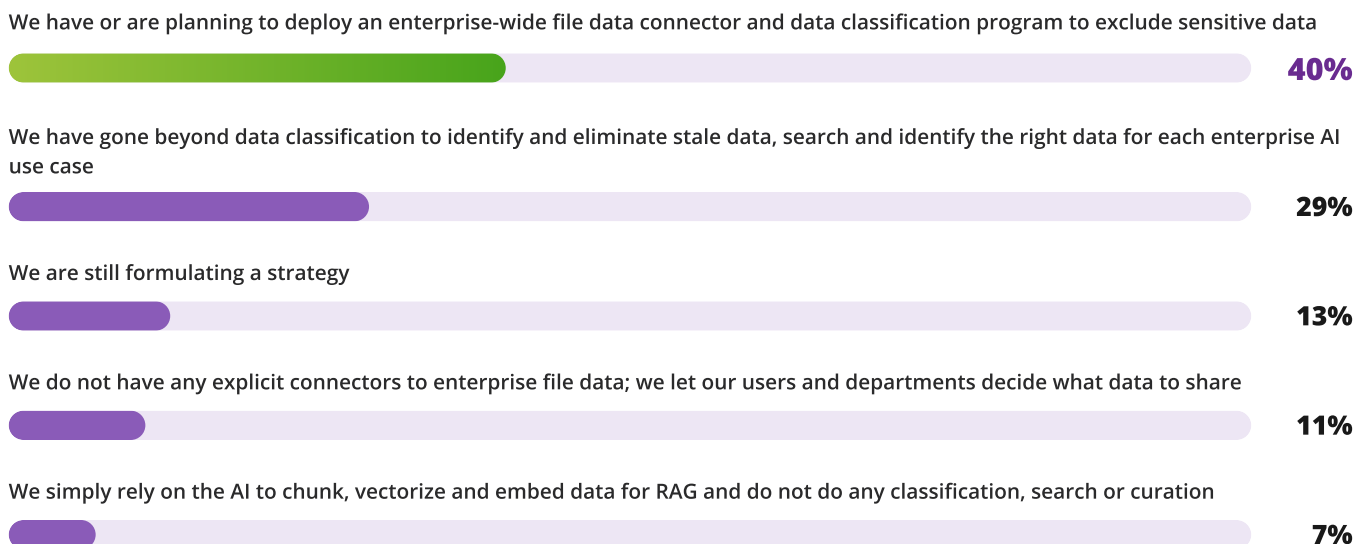
This part of data preparation is the most mature, with **the largest bucket of organizations (40%) reporting that they have or are planning to classify and exclude sensitive data from AI.**

With AI governance practices maturing, teams are moving beyond risk mitigation to the critical area of data quality: **29% are now filtering data assets to remove stale, irrelevant data and pinpoint just what's needed for their use cases.** This too, requires automated ways to crack open files and tag data with identifying information at scale.

Progress is promising, although nearly a quarter are not preprocessing data at all:

- 13% are still developing a data preparation strategy and
- 11% are letting individuals decide what to send to AI.

How are you connecting to dark enterprise file data, classifying the data, managing its security and ingesting it to AI?



Top 5 Takeaways

1

AI returns map to data readiness

Low AI ROI correlates with low unstructured data readiness. To generate measurable and meaningful ROI on AI, improving data quality must be job number-one among IT, data and AI leaders.



2

Data readiness is an ongoing tactic

Data readiness does not resolve as organizations grow adoption of AI across departments and functions. By investing in AI data management, enterprises can get ahead. Top pain points are metadata enrichment/extraction and clearing up redundant, obsolete and trivial (ROT) data to discover the right data for each project.



3

Sophisticated metadata management builds trust and lowers costs

Metadata analysis and extraction across the complete data lifecycle creates a foundation of trust, where AI users and engineers can easily find the most relevant, specific data for a project, without running up AI costs and risks.



4

Lakehouses are behind on unstructured data

A scant 4% of enterprises have developed schema for most file data. Until unstructured data carries the metadata that lakehouses require, these platforms deliver a single source of truth for structured data only, eroding their usefulness for AI.



5

Automation is still immature for unstructured data preparation

Filtering data for security risk is moving forward but less than one third filter for quality and 31% are not preprocessing unstructured data.



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.

Learn more at www.komprise.com



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.