komprise

The Komprise 2024 State of Unstructured Data Management

Building AI on a Budget

August 2024

Table of Contents

Top of Mind in 2024

Creating AI Infrastructure & AI Governance and Security

2

Executive Summary

Across the globe, AI mania is reaching epic proportions. Investors spent $27.1 billion on AI startups in the United States alone from April to June of this year, accounting for nearly half of all startup funding in the country during that period, according to PitchBook[1]. Yet when it comes to implementing these technologies, it's still early days in the enterprise, according to the fourth annual **Komprise 2024 State of Unstructured Data Management.**

Most (70%) organizations are still experimenting with these new technologies as "**preparing for AI**" remains a top data storage and data management priority for IT leaders. Yet cost optimization is an even higher priority this year and they are trying to fit AI into existing IT budgets. Only 30% say they will increase their IT budgets to support AI projects.

At this stage in the AI game, IT has not yet enforced much restriction on data, users or tools. Enterprises are instead focused on building out the proper infrastructure and technology stack for AI by upgrading data storage and data management technologies. Another key trend is the need to improve security and governance across GenAI and unstructured data management and more broadly by filling security skills gaps in the IT workforce, our research shows.

As the startup community and established technology vendors alike rush to launch new products and capabilities for AI, enterprise IT teams are still laying the technical foundation and determining the necessary economic strategies to get started. This year's survey shows that, when it comes to AI, it all starts with the data and how it is managed, mobilized, protected and prepared for new uses.

The 2024 Komprise State of Unstructured Data Management Report examines the challenges and opportunities with unstructured data in the enterprise. This report summarizes responses of 300 global enterprise IT leaders (director and above) at enterprises with more than 1,000 employees in the United States. The survey was conducted by a third party in June 2024.

[1] "Investors Pour $27.1 Billion Into A.I. Start-Ups, Defying a Downturn,"
The New York Times, July 3, 2024 https://www.nytimes.com/2024/07/03/technology/ai-startups-funding.html

Key Statistics

AI Plans
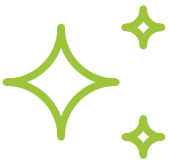
- Only 13% restrict what corporate data can be used in AI while 31% have no restrictions for users, apps or data in AI.

- Nearly half (44%) are creating AI-ready data infrastructure and 32% are building their own learning models.

- Only 30% are increasing the IT budget for AI projects.

- The leading challenge in prepping data for AI is managing governance/security concerns (45%), followed by data classification and tagging (41%).

- The leading tactic to address AI data concerns is to upgrade data storage/data management technologies (53%).

Data Storage Priorities

- The top priorities for data storage in the next year include cost optimization (54%), preparing for AI (51%) and investing in data management and data mobility (41%).

- Supporting AI projects (30%) was lower on the priority list, as IT organizations have competing needs to modernize disaster recovery and backups (32%) and migrate data to the cloud (38%).

- When it comes to unstructured data, IT leaders care the most about data access for the workforce followed closely by ease of data mobility.

Top Unstructured Data Challenges

- Moving data without disruption to users/apps is once again the top technical unstructured data management challenge (54%), followed by using AI to classify and segment data (48%).

- Prepping for AI is the top business challenge for unstructured data management (57%).

Future Needs

- AI data governance/security is the top future capability (47%) for unstructured data management.

- Nearly 60% need more staff with skills related to security, compliance and sensitive data.

Data Estate & Spending

Almost half of enterprises in the survey are storing more than 5PB of unstructured data and nearly 30% have more than 10PB. Storing and backing up billions of files across hybrid cloud environments is expensive, exceeding 30% of the IT budget for more than half (55%) of respondents this year. While IT budgets are growing this year at 8%, according to Gartner[2], funds are spread thin across many areas of need—including preparing for AI and hiring more IT staff (see sections V and VII respectively). Finally, as in years past, 90% of respondents plan to spend more on data storage this year than the prior year.

59% of respondents plan to spend more on data storage this year than the prior year.
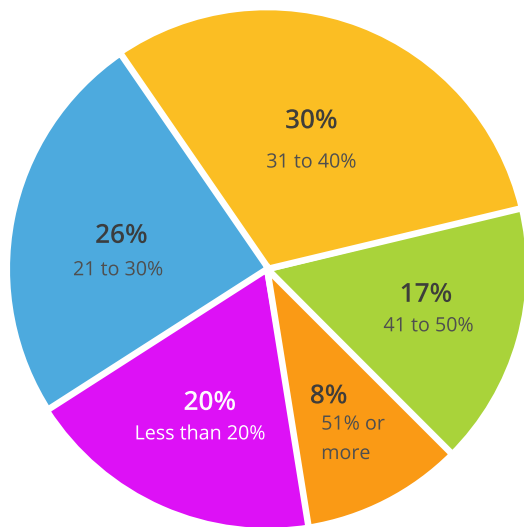
How much data are you currently managing?
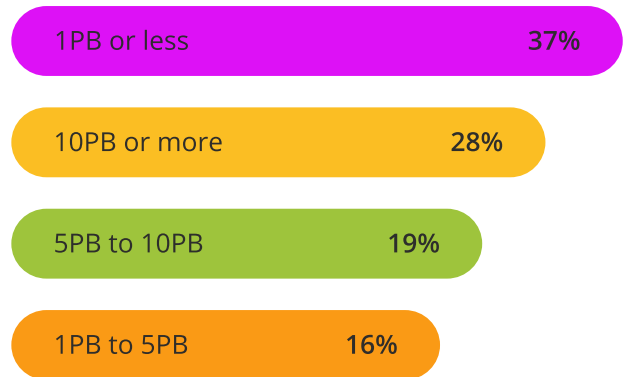
1PB or less	37%
10PB or more	28%
5PB to 10PB	19%
1PB to 5PB	16%

Storage and IT Budget

What percentage of your IT budget is spent on data storage and data protection (storage, backup, cloud, DR)?

- **30%** 31 to 40%
- **17%** 41 to 50%
- **8%** 51% or more
- **20%** Less than 20%
- **26%** 21 to 30%

Spending projections for 2024

Will you spend more on data storage in 2024 v. 2023?

- Yes **59%**
- Stay the same **31%**
- No, less **10%**
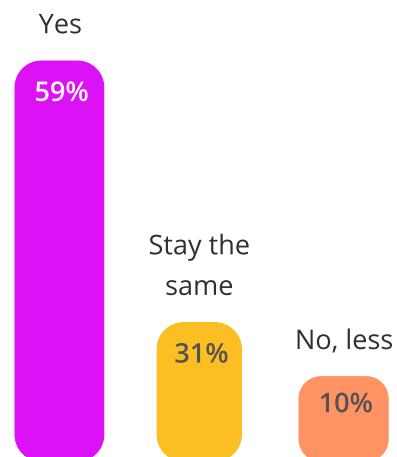
Insight:

The need to optimize storage costs is greater than ever today.

- Enterprises will need to balance data center costs with the complexity of managing data and apps in the cloud.
- Planning for AI is adding to the conundrum, given the expense of investing in AI-ready infrastructure and applications.
- Storage-agnostic data lifecycle management will be critical to right place data into the appropriate storage tier for cost savings, plus deliver intelligence on duplicate or zombie data that can be deleted altogether.

Top Challenges in Unstructured Data Management

This year, our survey asked participants to share both their technical challenges and business challenges for managing unstructured data. As in 2023, moving data without disruption to users and applications is a top technical challenge (54%), as IT leaders are always looking to deliver superior performance to their internal customers and avoid conflicts and unnecessary calls to the help desk.

> Moving data without disruption to users and applications is a top technical challenge.

But what do we mean by disruption? Primarily, this challenge pertains to the common problem when migrating or archiving data to new storage. Users are often left in the lurch, asking IT, where is their data now? Also, applications can break if the data is no longer in its original location or otherwise disconnected. These issues are caused by the proprietary methods that some storage and legacy data management technologies use to move data.

Read more about intelligent, non-disruptive data tiering strategies in the **Komprise Data Tiering Guide**.

The second leading technical challenge is using AI to classify and segment data (48%). Data classification, which involves identifying and labeling unstructured data, is required to aid user search across petabyte-scale distributed environments. Leveraging data for new value in analytics tools and managing data for compliance needs is difficult without a way to quickly find the right data sets based on keyword search. AI-enhanced data classification is a new, highly efficient way to do this, but best practices are still emerging here.

Managing data lifecycles across storage tiers and clouds (45%) is vexing for many organizations. They lack deep visibility into not only what data is stored where but the associated costs. In the cloud, services are changing constantly. This complexity is partly behind the trend in recent months to repatriate cloud workloads back on premises—yet that may not be the best answer when it comes to maximizing savings.
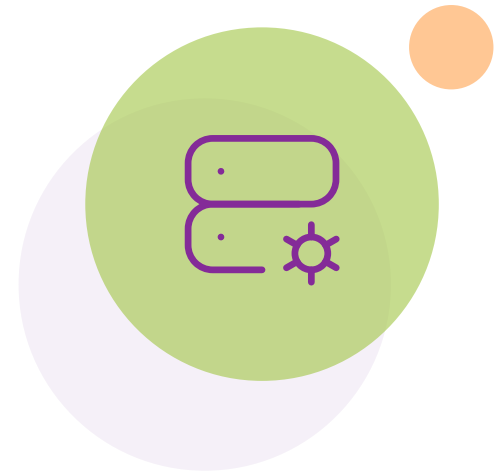
What are your top technical challenges with unstructured data management today?

Challenge	Percentage
Moving data without disrupting users and applications	54%
Using AI to classify and segment data	48%
Managing data lifecycle across storage and cloud tiers	45%
Empowering users to safely use data in AI tools	37%
Lack of visibility into what we have and what we can move/tier/migrate	30%
Finding and managing sensitive data	23%
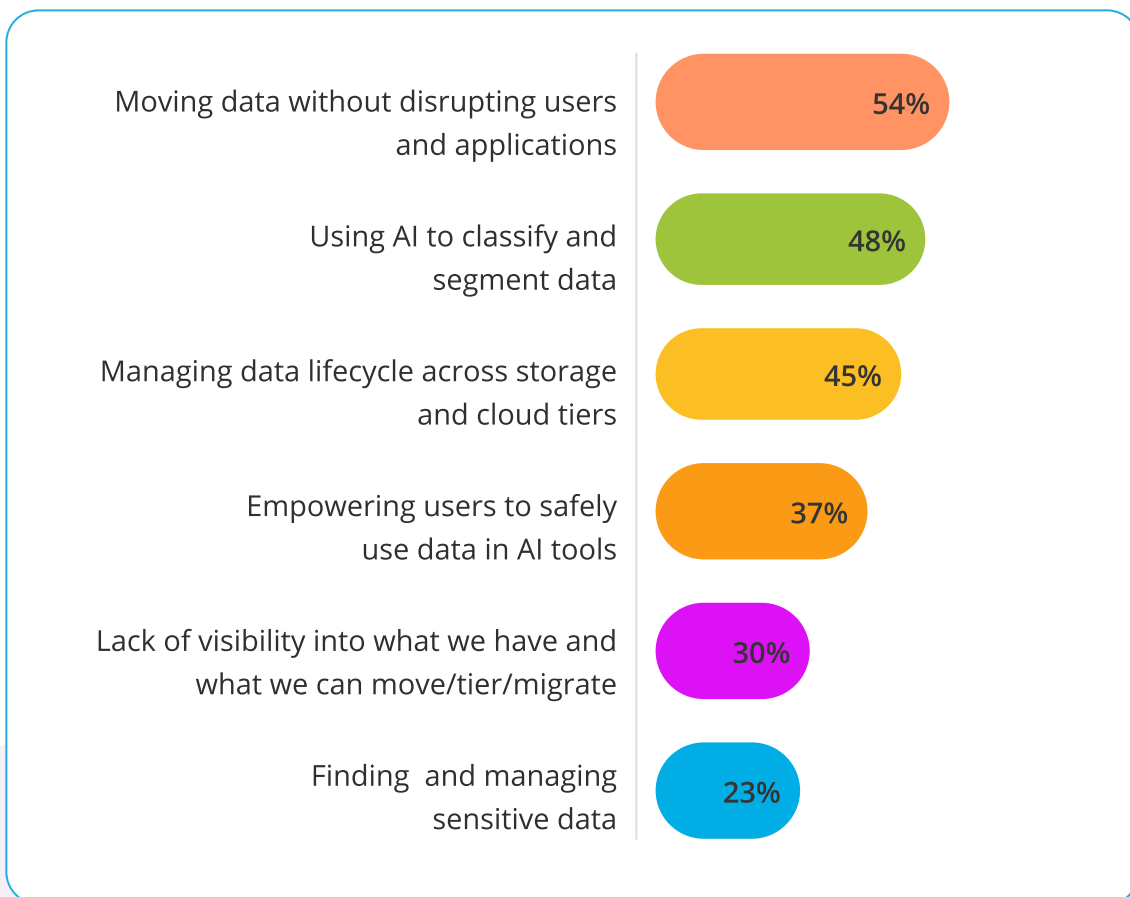
Top Business Challenges of Unstructured Data

Given the tangled security, privacy, ethical and legal considerations of AI in both business and society, it makes sense that **preparing for AI is the leading business challenge for unstructured data management (57%).** Navigate to the AI Plans section below to read more about AI concerns and tactics.

A lack of departmental visibility into storage spend (36%) is the third-leading business challenge, in an age when enterprises are instituting chargeback policies to their departments for shared accountability on IT spending.

Compliance with industry regulations and corporate policies (44%) is another hot topic. IT organizations are battling costly ransomware threats while also needing to adapt to a growing number of state and federal privacy regulations. Both threats relate directly to the way data is stored, shared and accessed. Data security is no longer a siloed activity in the enterprise, but one requiring collaboration among many different teams including data storage and network professionals.

What are your top business challenges with unstructured data management today?

Preparing for AI

57%

Compliance with industry regulations and/or internal policies

44%

Department and users lack visibility into their storage spend and data usage which makes alignment difficult

36%

Growing too fast and the need to control costs

34%

Ransomware protection & security requirements for data

31%

Generating new value/use from our data

21%

Artificial Intelligence Plans

Everyone's talking about AI but there is no winning formula yet for how organizations should deploy it. As always, best practices depend on your organization's needs and goals, risk profile and your industry. For now, participants are employing a fairly hands-off approach to how employees use AI with corporate data.

Only 13% restrict the data and AI services employees can use. Most (60%) have no restrictions at all or just restrict which users can train AI models.

Only **13%** restrict the data and AI services employees can use.

More than 30% say they have zero restrictions on the data, users or apps for AI. And, even with widespread concerns about AI from a safety, ethics and accuracy standpoint, only 5% completely disallow AI in the workplace.

How do you plan to use AI with your corporate data?

33%
Only authorized users can train AI models using corporate data

31%
Any employee can use AI to enhance their productivity with no restrictions

18%
Any employee can use any data but for approved AI services only

13%
Only some corporate data can be shared and only with approved AI services
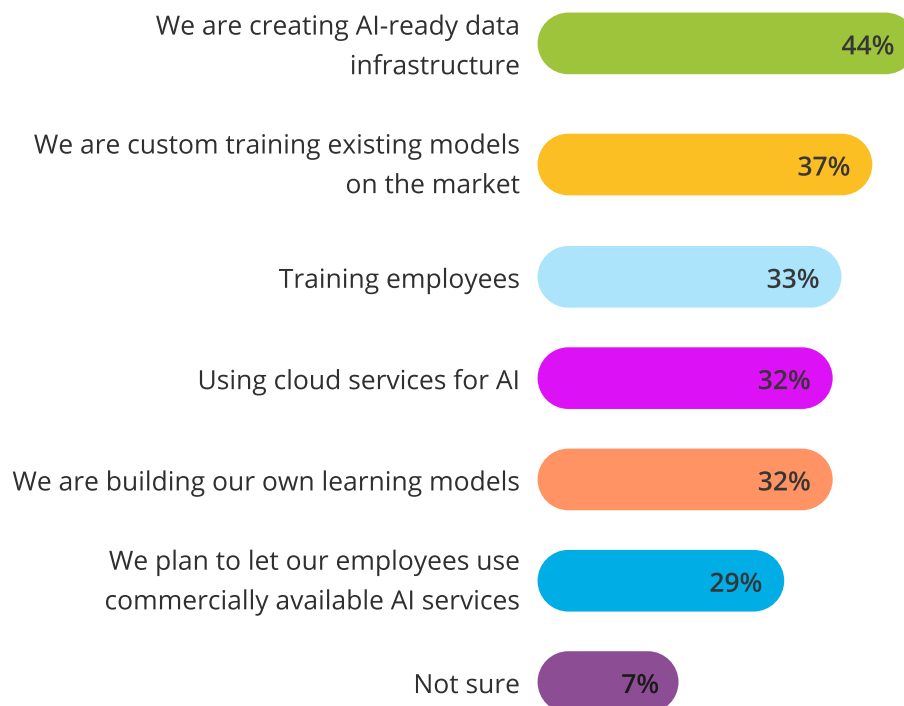
5%
We don't allow any use of AI

AI Strategies

There are many technical decisions that go into supporting AI—from which LLMs to use, where to deploy it, what infrastructure is required and how to train employees. IT leaders agree that creating AI-ready data infrastructure is most important, as noted by nearly half (44%). After that, the trends are less clear from developing AI models internally to using commercial services and/or the cloud to support AI.

While pre-baked infrastructure, public models and cloud services offer cost and ease-of-use benefits, IT organizations are also weighing the benefits of keeping AI in-house for better controls around data governance, data transparency and data security.

Nearly half of IT organizations are creating AI-ready data infrastructure.

How do you plan to leverage AI?

We are creating AI-ready data infrastructure	44%
We are custom training existing models on the market	37%
Training employees	33%
Using cloud services for AI	32%
We are building our own learning models	32%
We plan to let our employees use commercially available AI services	29%
Not sure	7%

Data Concerns for GenAI

There have been more than a few public cases of generative AI gone wrong since ChatGPT launched in the fall of 2022. Risks include exposing sensitive internal data into commercial language learning models (LLMs) which are then available to any external user, the ownership legalities of derivative works, the risks of private data such as PII being inadvertently fed into AI, and a lack of data transparency leading to false, misleading and/or biased outcomes.

> Corporate data leakage and privacy and security violations of data are the leading concerns for Generative AI.

Regarding GenAI, participants are equally worried about corporate data leakage (27%) and privacy and security violations of their data (26%). Unpredictable AI costs (14%) came in at a distant third. Given how easy it is to use GenAI, only 4% worry that their employees don't have the right skills or knowledge.

What is your top data management concern regarding using generative AI in your organization?

27%
Leaking of corporate data into the vendor's language learning model (LLM)

26%
Violation of privacy and security of our data

14%
Potential liability of using GenAI outputs that contain others' PII or IP
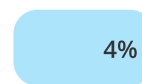
14%
Unpredictable AI costs

10%
Data provenance: lack of data source transparency and/or risk of inaccurate or biased data in vendor LLMs

5%
We don't have concerns or are unsure

4%
Our employees don't have the right skills/knowledge

Insight:

Data risks for generative AI came out of nowhere, with many organizations caught flat-footed as their employees jumped onto the GenAI bandwagon:

- There are five key areas of AI data governance to consider across security, privacy, lineage, ownership and governance of unstructured data for AI.

- **This blog** delves into the practices, starting with gaining deep visibility and search capabilities across all unstructured data so that you can track its usage, find and segregate sensitive data, monitor data.

Budgeting for AI

We asked participants how they would pay for AI and the results suggest that economizing in other areas will be imperative. Only 30% will increase the IT budget for AI, while 34% will use existing cloud funds to pay for it. Merging AI into the cloud budget makes sense on one level, since the major cloud providers are investing heavily in AI and ML technologies.

On the other hand, cloud spending has been under the microscope in the post-pandemic era as enterprise IT organizations wound up with much larger bills than expected. AI funding struggles are echoed by the Spiceworks 2024 State of IT research[3], showing that only 3% of software budgets are earmarked for AI.

Only **30%** will increase the IT budget for AI, while **34%** will use existing cloud funds to pay for it.

14

How are you budgeting for AI?

AI will be part of our budgeted cloud spend	**34%**
We are increasing our budget to add AI spending	**30%**
We will use efficiency gains from AI to fund more AI initiatives	**19%**
Unpredictable AI costs are a concern, and we don't know how to budget for it yet	**17%**
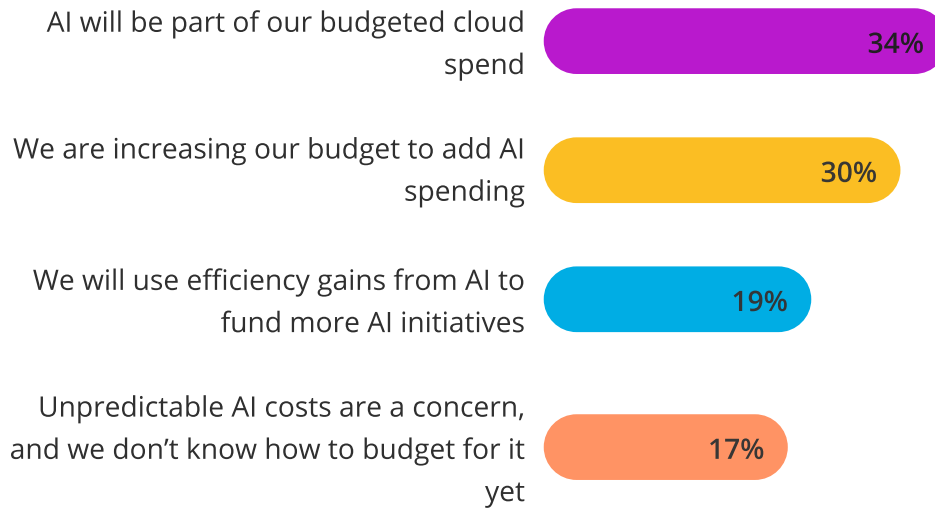
Insight:

Cost optimization is a central theme this year. By optimizing data management and storage, investing in energy-efficient data center technologies and being strategic with cloud spending versus lift and shift, organizations can find budget to support revenue-generating AI initiatives. AI vendors, too, will need to market their products and services with cost-efficiency in mind.
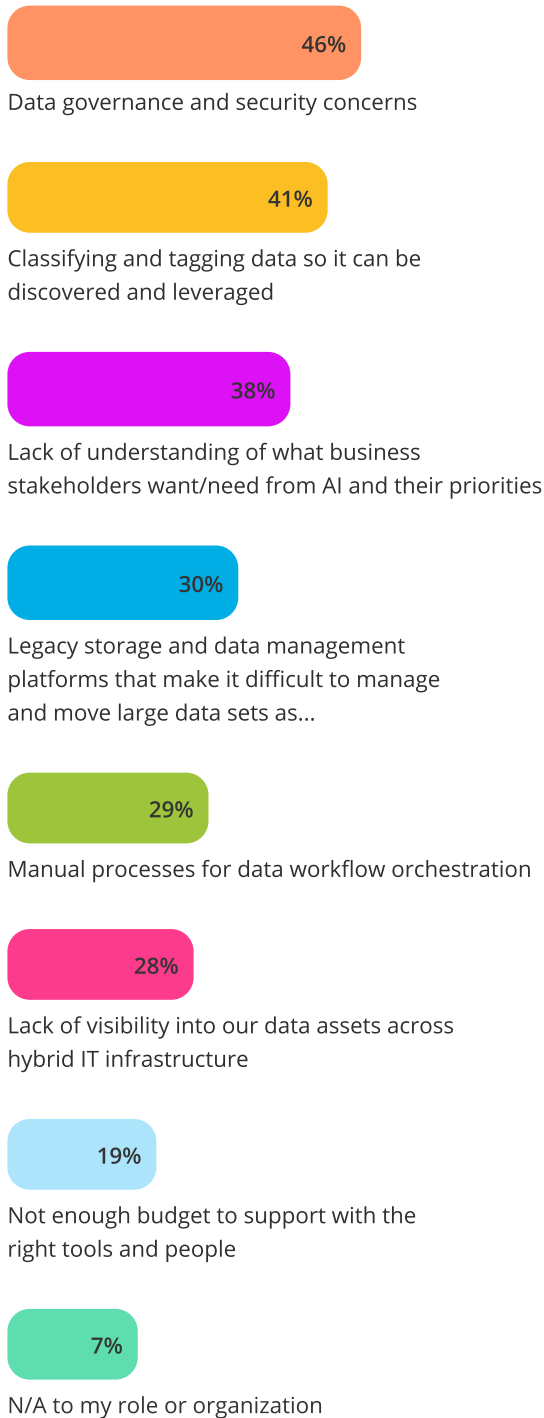
Top Challenges of Prepping Data for AI

Preparing for AI is the leading business challenge for unstructured data management and prepping data is a significant part of this endeavor. Naturally, governance and security is the leading challenge here, as expressed by 45%, followed by classifying and tagging data (40%) and understanding business needs (38%). A large majority (87%) also expressed challenges that relate to a lack of the right tools—citing visibility, legacy technology and manual processes.

How to address these challenges?

We also asked IT leaders how they will handle these issues, and they ranked modernizing data storage and data management platforms and creating internal task forces as top solutions. Close behind is the tactic of using cloud-based AI services, which automate processes such as creating LLMs, provisioning the right hosting resources, and filtering data. **AI data workflows** are an emerging concept. This technology can discover, classify and automate the movement of data to AI tools and enrich the metadata of unstructured data.

What are your top challenges in preparing data for use in AI tools?

46%
Data governance and security concerns

41%
Classifying and tagging data so it can be discovered and leveraged

38%
Lack of understanding of what business stakeholders want/need from AI and their priorities

30%
Legacy storage and data management platforms that make it difficult to manage and move large data sets as...

29%
Manual processes for data workflow orchestration

28%
Lack of visibility into our data assets across hybrid IT infrastructure

19%
Not enough budget to support with the right tools and people

7%
N/A to my role or organization

How do you plan to address these challenges?

53%
Working to modernize and upgrade our data storage and data management platforms for AI

41%
Creating internal task force of IT security, legal and others to create a strategy

37%
Using public cloud services to help speed up, host and automate data prep processes

37%
Developing new tools in-house

29%
Hiring outside consultants, partners or experts

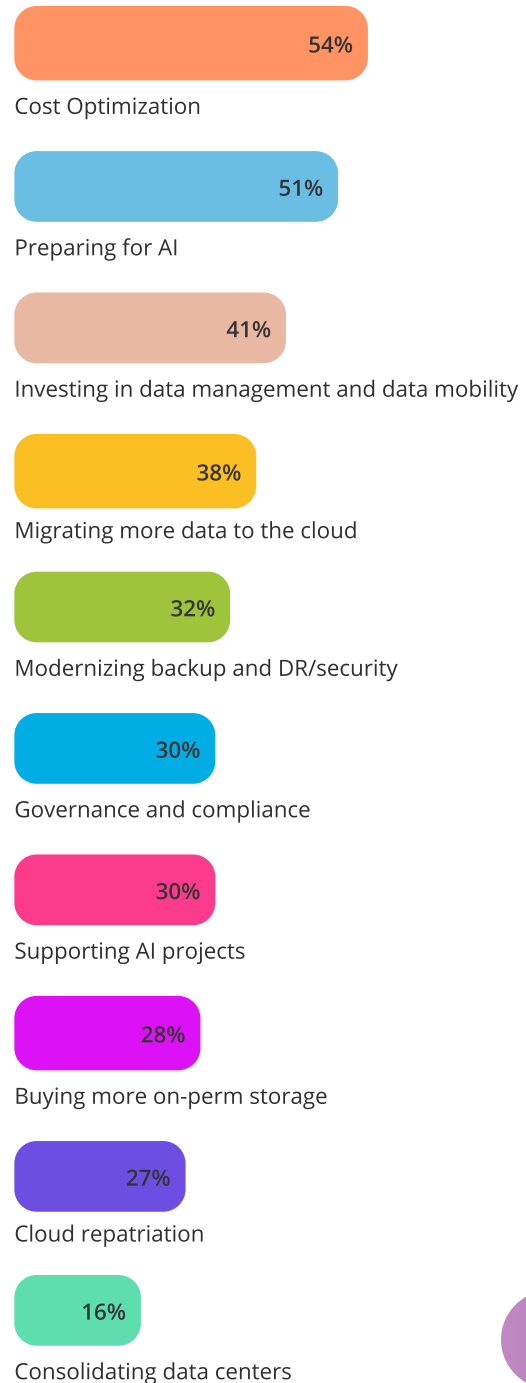**9%**
We don't have a plan yet; unsure

Data Storage & Unstructured Data Priorities

Managing costs is job #1 today for IT directors and executives overseeing data storage. Even with the cost of hardware continuing to decline and many classes of affordable object storage in the cloud, it is still tricky to manage the cost of data. After all, most organizations don't delete much or any data for fear that someone, someday, will need it. Plus, backups and disaster recovery requirements create multiple copies of data—which sit on expensive NAS devices.
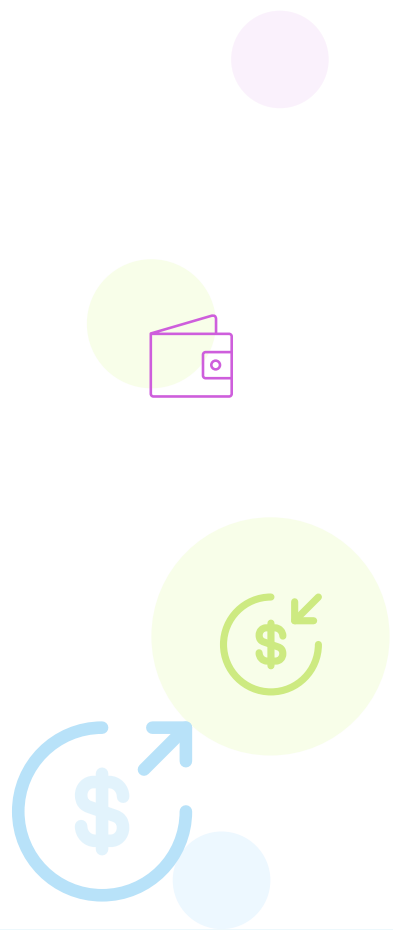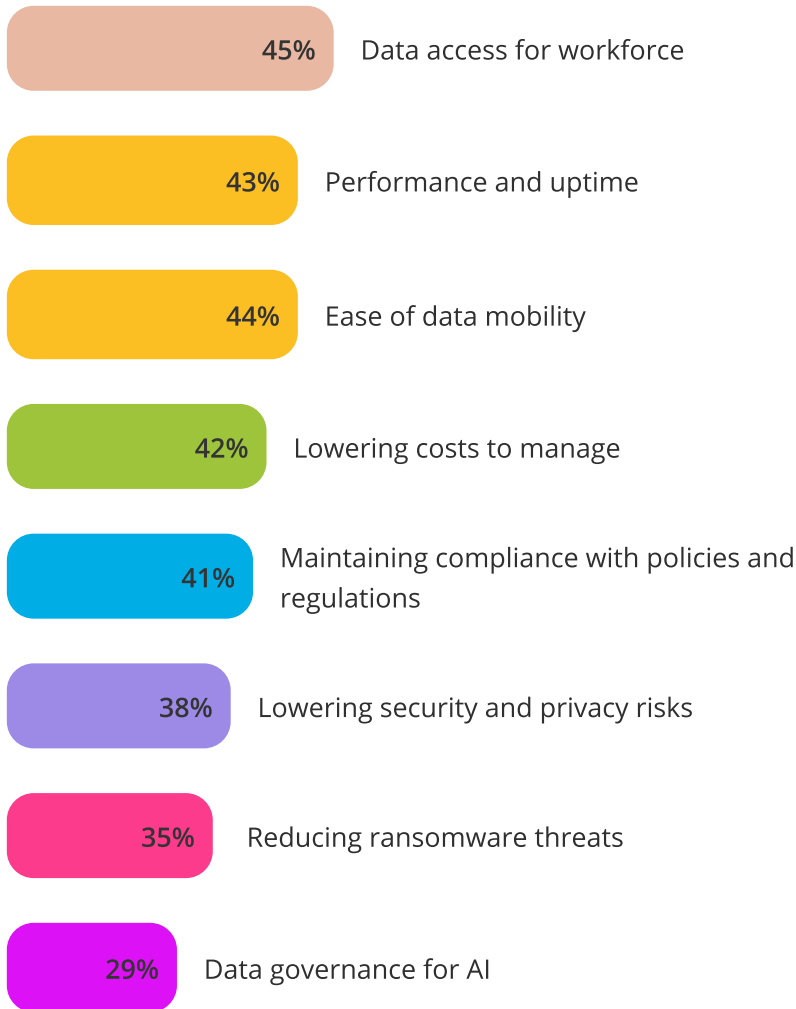
Amid these pressures, prepping for AI—the top data storage priority of 2023—came in at a close second, followed by investing in data management and mobility technologies. Less than **30%** will buy more on-premises storage.

What are your top data storage priorities in the next 12 months?

Priority	Percentage
Cost Optimization	54%
Preparing for AI	51%
Investing in data management and data mobility	41%
Migrating more data to the cloud	38%
Modernizing backup and DR/security	32%
Governance and compliance	30%
Supporting AI projects	30%
Buying more on-perm storage	28%
Cloud repatriation	27%
Consolidating data centers	16%

We also asked participants what their bosses cared about most. The answers focused squarely on meeting the needs of end users: data access, ease of mobility and performance and uptime.

What does your IT leadership care about most per unstructured data?

45% Data access for workforce

43% Performance and uptime

44% Ease of data mobility

42% Lowering costs to manage

41% Maintaining compliance with policies and regulations

38% Lowering security and privacy risks

35% Reducing ransomware threats
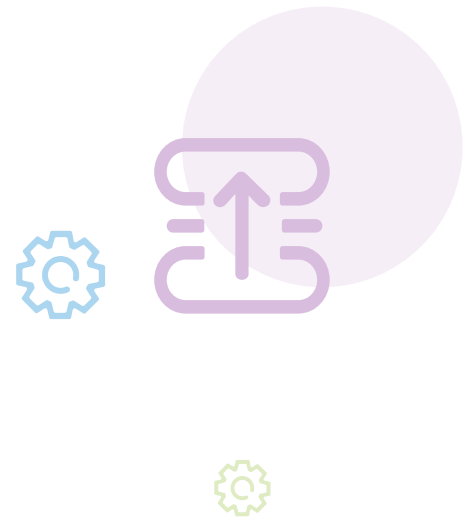
29% Data governance for AI

Insight:

Getting ready for AI remains important, but not at the expense of fiscal responsibility.

- IT organizations must first keep data protected, available and usable in the most efficient manner possible.

- Managing data more strategically can help on several fronts including right-placing data for the best economics, enriching and classifying data for AI and ML and ensuring great performance for end users.

Future capabilities for unstructured data management

Data security threats continue to breathe down the necks of IT professionals, creating the need for more and more layers to keep hackers away from systems that house corporate data. Data storage technologies often feature built in security today and it's logical to expect unstructured data management software to also support security, privacy and governance needs.

AI data governance and security (47%) is the top future need today—indicating maturing plans for AI in production. This is up from 6th place and 28% in 2023.

AI data governance includes not only protecting data from breaches or misuse, but ensuring data accuracy, compliance with industry regulations, managing biases in data, and ensuring that the data sets do not lead to false, misleading or libelous results.
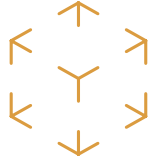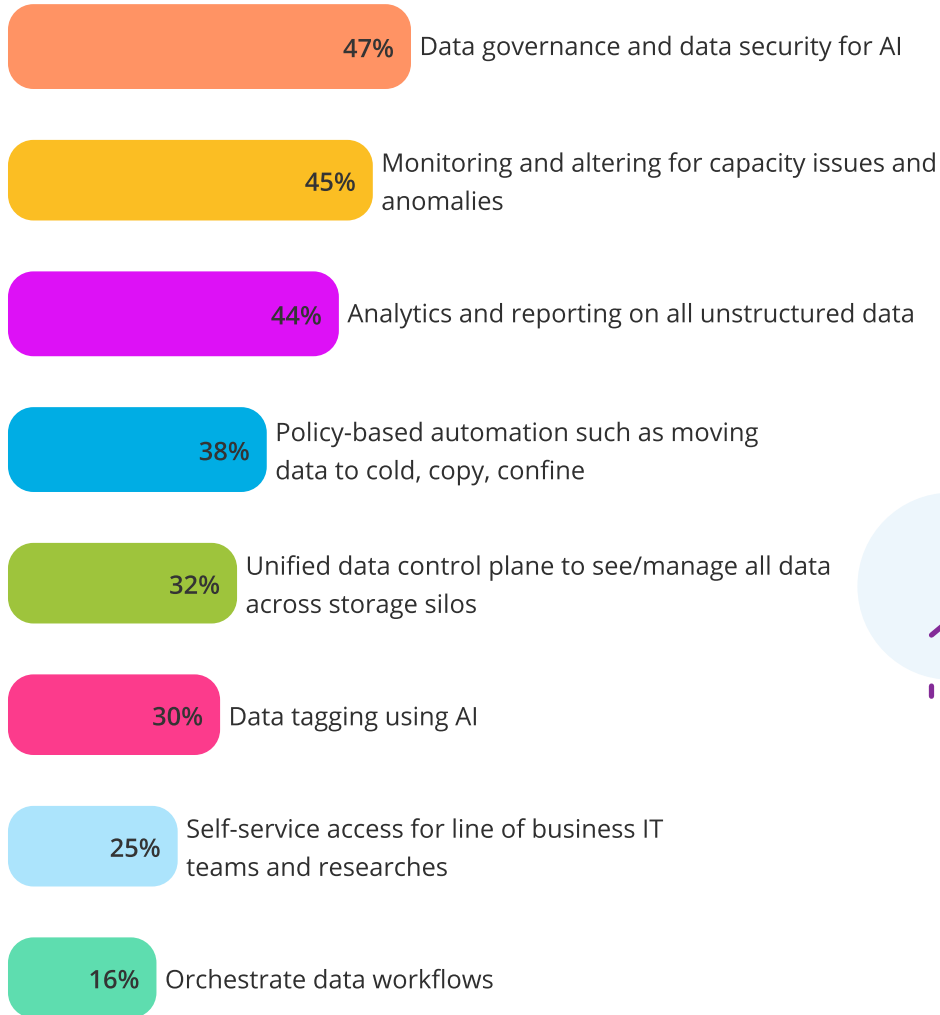
Monitoring and alerting for capacity issues or anomalies, last year's top pick, remains high again along with analytics and reporting.

Insight:

As unstructured data management evolves, organizations are looking for much more than a way to easily migrate or tier data to new storage and to analyze costs:

- They want easier ways to manage data risks, such as by giving users granular search to find PII data and automatically move it to secure locations according to corporate policies.

- They want to know if there is unusual activity on a user's share or if a storage array is reaching capacity.

What unstructured data management capabilities will be most important to you in future?

47% Data governance and data security for AI

45% Monitoring and altering for capacity issues and anomalies

44% Analytics and reporting on all unstructured data

38% Policy-based automation such as moving data to cold, copy, confine

32% Unified data control plane to see/manage all data across storage silos

30% Data tagging using AI

25% Self-service access for line of business IT teams and researches

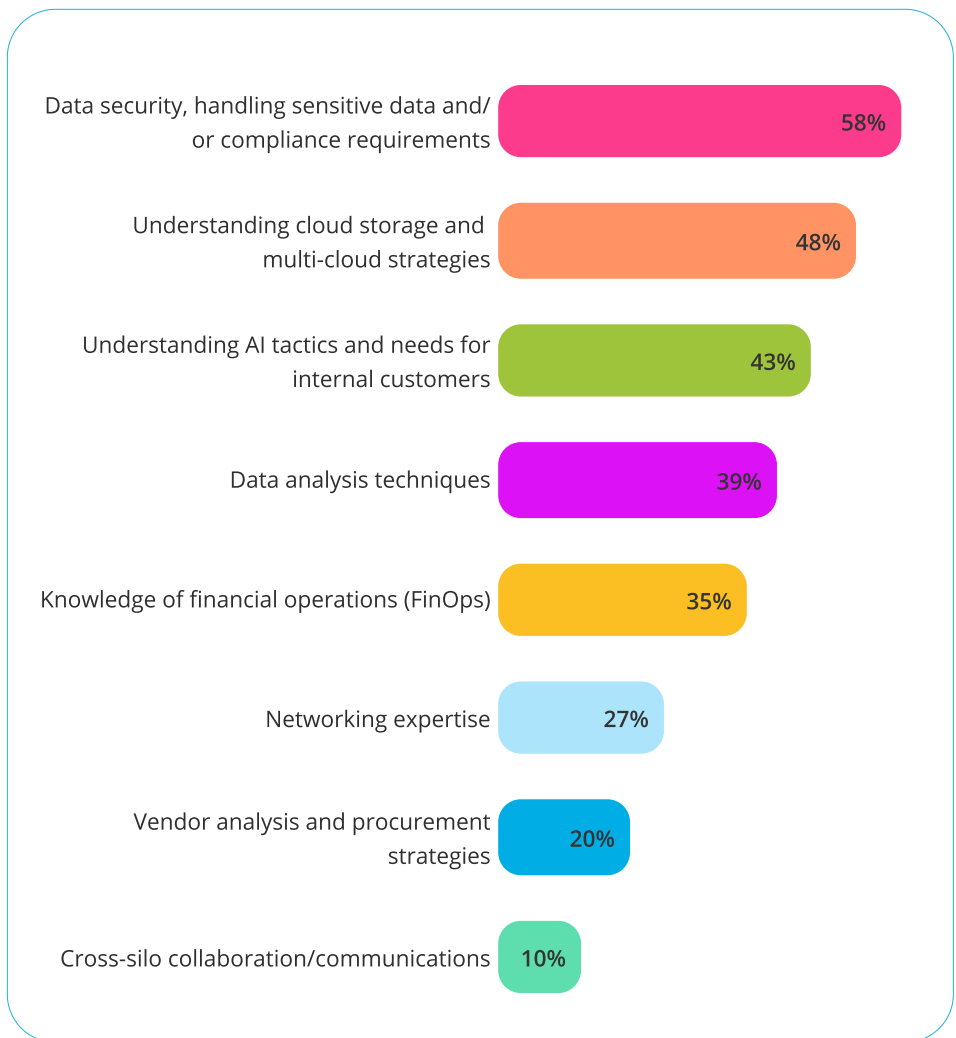**16%** Orchestrate data workflows

IT Skills & Organization

This year we introduced a new section to the survey focused on IT skills and organizational changes. **IT teams are struggling the most to fill skills and knowledge gaps in data security and compliance (58%).** Understanding cloud storage is the second priority, followed closely by AI tactics and needs for the business.
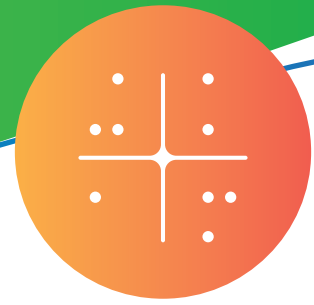
To compare, a recent IDC survey[4] of North American IT leaders found that AI skills were in most demand, followed by IT operations and cloud skills across architecture, data management and storage. These skills gaps are resulting in missed revenue growth objectives, quality problems, and a decline in customer satisfaction, IDC notes.

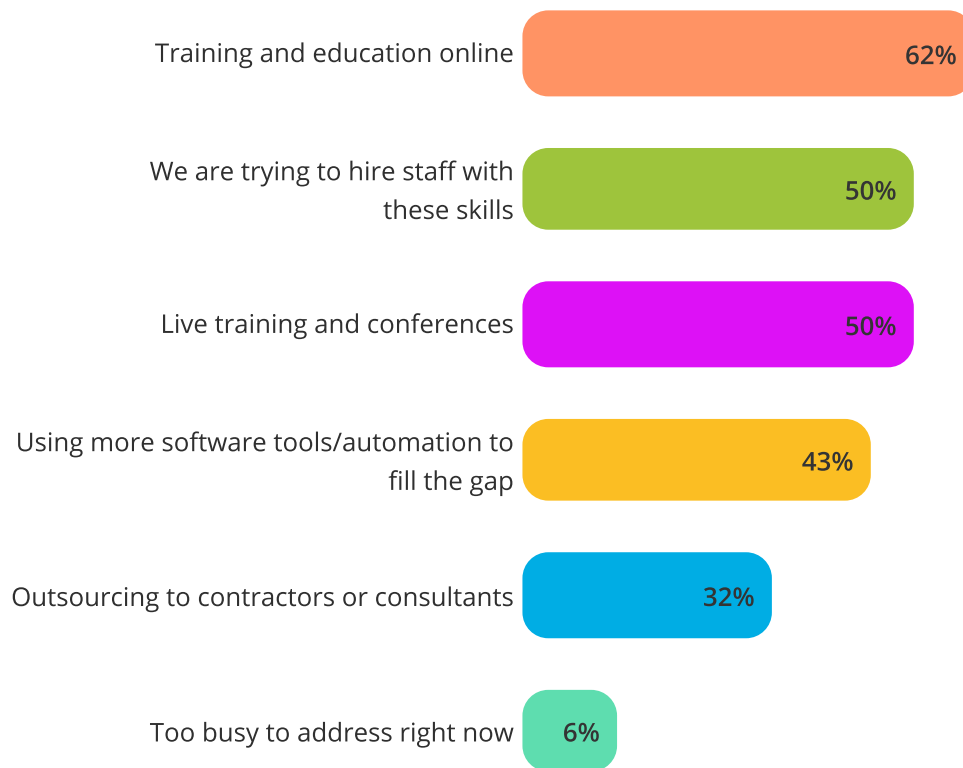**What are the top skills & knowledge gaps in your data management/ storage teams?**

Skill	Percentage
Data security, handling sensitive data and/or compliance requirements	58%
Understanding cloud storage and multi-cloud strategies	48%
Understanding AI tactics and needs for internal customers	43%
Data analysis techniques	39%
Knowledge of financial operations (FinOps)	35%
Networking expertise	27%
Vendor analysis and procurement strategies	20%
Cross-silo collaboration/communications	10%

Tactics to address skills gaps

When asked how they would address these gaps, participants in our survey prioritized online training (62%), followed by hiring and live training/events (a tie for 50%). This should be comforting news to people worried about AI taking over their jobs—although 43% are relying on automation to help fill the need.
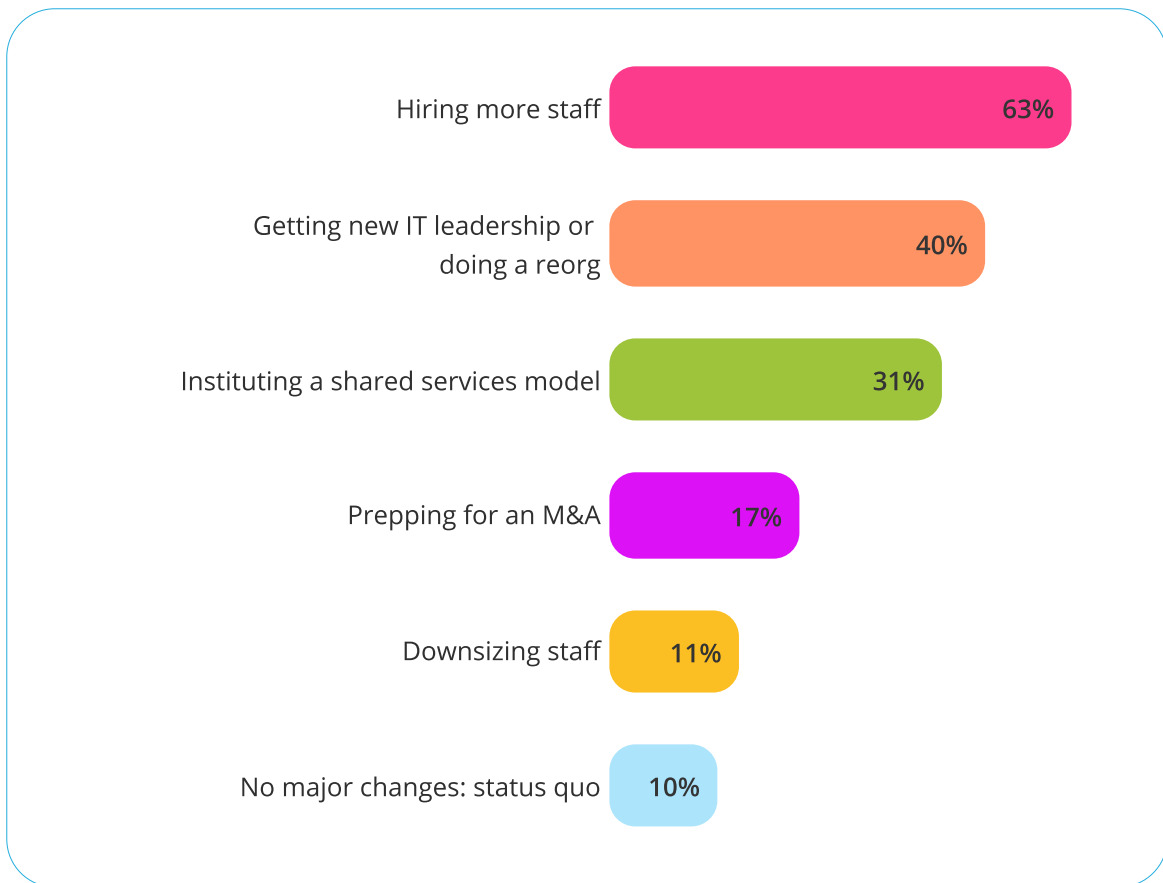
How are you addressing these gaps?

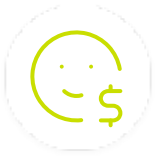| Category	Percentage
Training and education online	62%
We are trying to hire staff with these skills	50%
Live training and conferences	50%
Using more software tools/automation to fill the gap	43%
Outsourcing to contractors or consultants	32%
Too busy to address right now	6%

Planned IT Organizational Changes

We also asked what organizational changes IT will implement in the coming year, with a clear winner for hiring (63%), followed by an IT leadership change or reorganization (40%). A reorganization will include adopting a shared services model for 31%, a popular tactic to reduce costs by streamlining the procurement, delivery, measurement and billing for IT services. The coming year portends to be a turbulent time for IT, as teams manage many competing, shifting priorities with often limited resources.

What organizational changes are ahead for IT in the coming year?

Category	Percentage
Hiring more staff	63%
Getting new IT leadership or doing a reorg	40%
Instituting a shared services model	31%
Prepping for an M&A	17%
Downsizing staff	11%
No major changes: status quo	10%

Top 5 Takeaways

Cost Optimization Leads the Day

For the fourth year in a row, IT directors say they will spend more on storage this year than the previous. This year, the top data storage priority is cost optimization, over preparing for AI in 2023. Participants told us that their IT leadership cares most about making data accessible, highly available and easy to move as needed. They want these benefits while not disrupting users and applications—the top technical challenge for unstructured data management again this year.

Scrappy Times for AI

Getting ready for AI involves a potential laundry list of expenditures from upgrading storage and computing infrastructure, cleaning and preparing data, training or developing custom LLMs, acquiring new IT skills, and beefing up security tools. Paying for this will be tough, with only 30% saying they will increase the IT budget for AI. Strategies will in most cases involve carving out existing budgets—such as cloud, according to 34% – to fund AI. It will be imperative to optimize data management for savings while also investing in energy-efficient data center technologies and being surgical with cloud spending to avoid waste.

Security & Governance Concerns Hinder AI

Participants are equally worried about corporate data leakage and privacy and security violations of their data. IT teams will at a minimum need policies and tools to prevent sensitive data leakage into general purpose LLMs and to audit employee usage of AI to lower the incidence of false or misleading outcomes. IT will modernize data storage and data management platforms and create internal task forces as first steps to safely deploy AI.

Unstructured Data Management Evolves for AI Governance

Unstructured data management solutions are maturing far beyond giving IT users a way to easily migrate and tier data to new storage and to analyze costs. IT teams now want features supporting AI data governance and security, such as the ability to quickly find, tag and classify sensitive data and move it automatically by policy to secure storage where it can't be ingested into GenAI. Other capabilities include creating AI data workflows which integrate data management tools with AI tools to find and tag sensitive data sets like PII across large data estates.

AI Strategies Focus on Infrastructure

IT leaders are setting their sights today on creating "**AI-ready data infrastructure**." There are many pathways for this—between procuring and developing internal technology, using cloud services or combining those strategies in a hybrid model. IT teams are split on developing AI models internally versus using commercial services and/or the cloud. AI technology decisions will need to factor in internal expertise and resources to support these new technologies, budget and data security concerns.

About Komprise

Komprise is a provider of unstructured data management and mobility software that frees enterprises to easily analyze, mobilize, and monetize the right file and object data across clouds without shackling data to any vendor. With Komprise Intelligent Data Management, you can cut 70% of enterprise storage, backup and cloud costs while making data easily available to cloud-based data lakes and AI/ML tools.

Learn more at www.komprise.com