

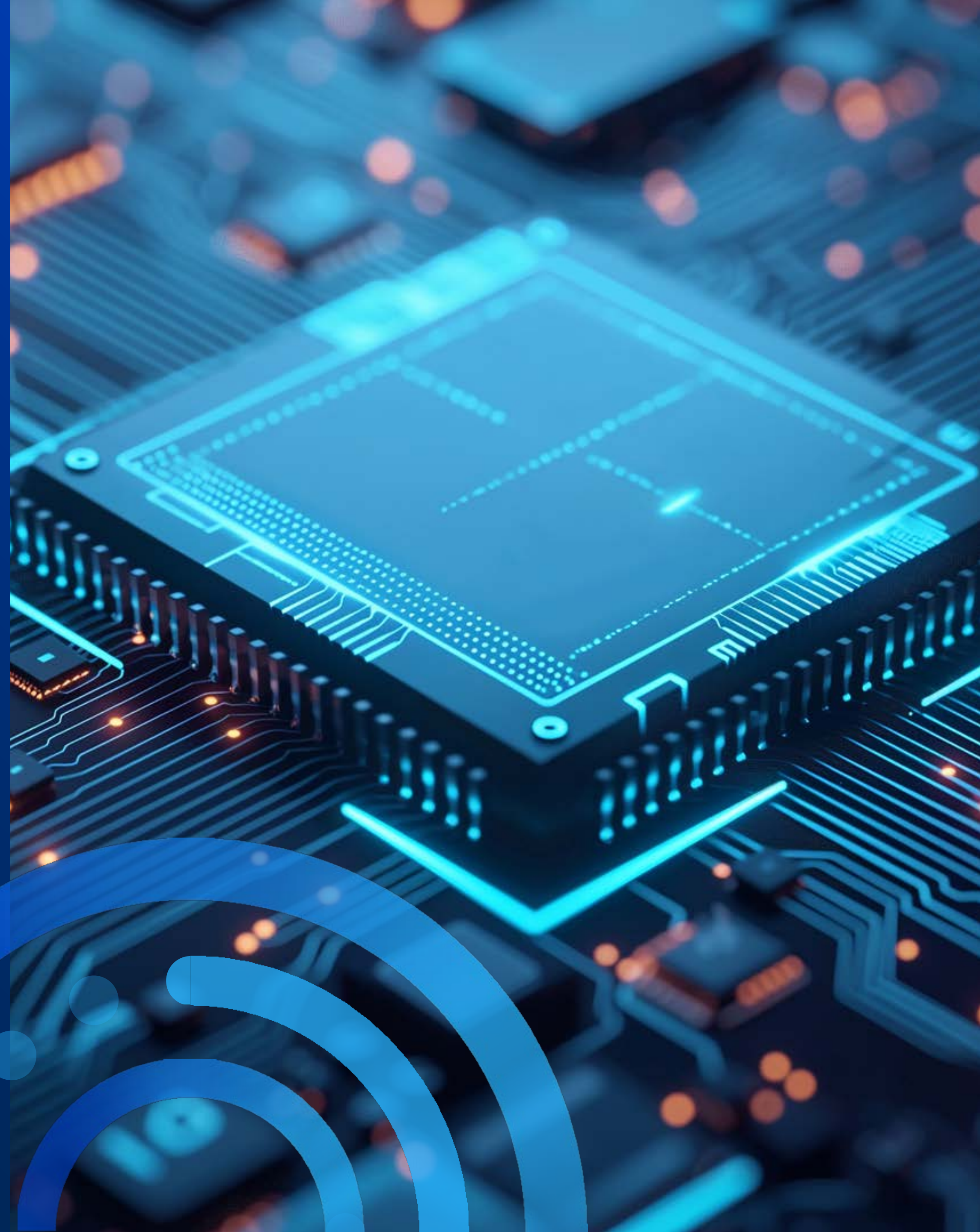
Getting Data Fit for AI – How Pentaho Intelligent Data Optimization Cuts Costs and Accelerates AI Readiness

Pentaho Data Optimizer for AI Success

Stephen Catanzano | *Senior Analyst Omdia*

February 2026

This Omdia eBook was commissioned by Pentaho
and is distributed under license from TechTarget, Inc.



Introduction - The AI paradox: massive data, massive waste

Organizations worldwide are accelerating their AI initiatives, driven by the promise of improved decision-making, operational efficiency, and competitive advantage. Yet beneath the excitement lies a sobering reality: Most enterprises are not prepared to deliver on AI’s potential because their data foundations are disorganized, duplicated, and poorly understood.

The challenge is not a lack of data; it is the opposite. Organizations are drowning in data they cannot find, cannot classify, and do not need. Storage costs are spiraling as data volumes grow, yet much of what organizations pay to store provides little to no business value. This is data debt, and it is the hidden barrier to AI success.

AI quality depends entirely on data quality. Organizations cannot train effective models or deploy reliable AI agents on data that is fragmented, duplicated, or inaccessible. Before asking, “How do we implement AI?” organizations must first ask, “Is our data fit for purpose?”

This eBook explores the data fitness imperative and highlights how Pentaho Data Optimizer can help organizations cut waste, reduce risk, and transform their data into a strategic asset ready for AI and core business initiatives.

KEY DATA POINTS:

61% of organizations said AI and machine learning initiatives are creating more data, making the management challenge even greater.¹

17% cited ensuring data quality and consistency as the single most significant challenge in managing data for AI.²

27% reported prioritizing data and analytics for improving decision-making and strategic planning, the top priority cited.³

¹ Source: Enterprise Strategy Group (now Omdia) Research Report, [The Critical Role of Storage in Building an Enterprise AI Infrastructure](#), September 2025.

² Source: Enterprise Strategy Group (now Omdia) Research Report, [Navigating Build-versus-buy Dynamics for Enterprise-ready AI](#), January 2025.

³ Source: Omdia Research Report, [Optimizing Cloud Analytics Costs in an Agentic AI Future](#), October 2025.



CONTENTS



The background is a dark blue gradient. It features faint, semi-transparent binary code (0s and 1s) scattered across the surface. A stylized bar chart with multiple bars of varying heights is also visible, rendered in a light blue and white color scheme. The overall aesthetic is digital and data-oriented.

The data growth crisis

Data is growing faster than organizations can manage

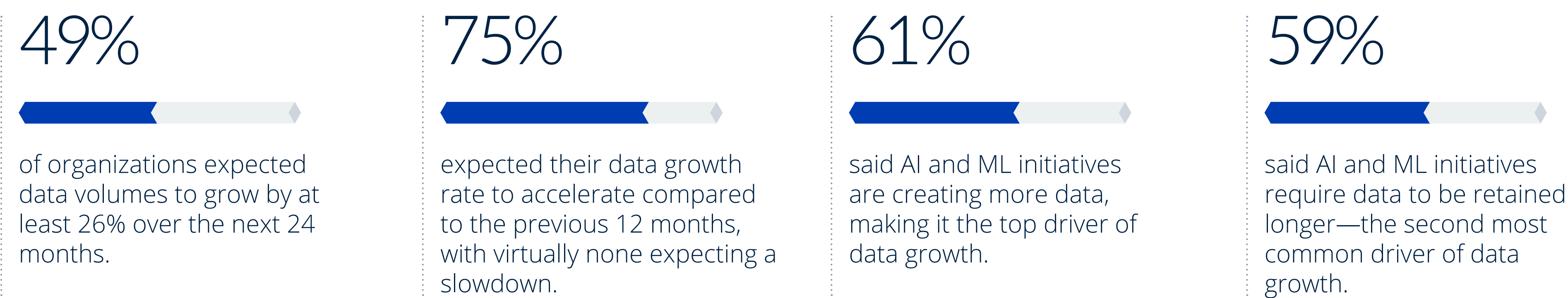
The data deluge is not slowing down; it is accelerating. Nearly half of organizations expect their total data volumes to grow by at least 26% over the next two years, with nearly one in six expecting growth of more than 50%.⁴ This is not a future problem to plan for; it is a current reality that demands immediate attention.

What is driving this explosive growth? AI itself. Organizations reported that AI and machine learning initiatives are the primary drivers of increased data volume and longer data retention requirements.⁵ The very technologies organizations are racing to adopt are compounding the data management challenges the organizations already face.

This creates a difficult cycle. AI requires vast amounts of quality data to function effectively, but AI initiatives generate even more data that must be stored, managed, and governed. Without intelligent systems to classify, tier, and optimize this growing data estate, storage costs will continue to spiral while the percentage of data providing actual business value shrinks.

The organizations that will succeed with AI are those that get ahead of this growth curve now, implementing solutions, such as Pentaho Data Optimizer, to ensure data is properly stored from the moment it is created through its entire lifecycle.

Key data points:⁶



⁴ Source: Enterprise Strategy Group (now Omdia) Research Report, [The Critical Role of Storage in Building an Enterprise AI Infrastructure](#), September 2025.

⁵ Ibid.

⁶ Ibid.

Much of enterprise data is redundant, obsolete, or trivial

Here is an uncomfortable truth: The majority of data that organizations pay to store provides little or no business value. Industry analysis consistently shows that redundant, obsolete, and trivial data, known as ROT, accounts for a majority of typical enterprise storage. A global [Splunk and Enterprise Strategy Group study](#) cited by IBM revealed that 60% of business leaders believe half or more of their organization’s data is “dark,” meaning they do not know what it contains or whether it has value. Nearly one-third said three-quarters of their data fell into this category. Most enterprises today manage somewhere between five to 10 petabytes of data, while large enterprises exceed approximately 65 petabytes.⁷ This is not merely an efficiency problem; it is a financial drain hiding in plain sight. Organizations are paying premium storage rates for data that no one accesses, that is duplicated across multiple systems, and that should have been purged years ago. Meanwhile, this clutter makes it harder to find the high-value quality data that actually drives business decisions and AI initiatives.

The problem extends beyond cost. Obsolete data often contains sensitive information, including personally identifiable information and protected health information, which creates compliance risk. Every forgotten file with customer data represents potential regulatory exposure. Reducing sensitive data does not just save money; it reduces risk.

According to our research, 52% of enterprises reported that they use or plan to use unstructured data in their AI workloads,⁸ including data residing in PDFs, videos, emails, and documents that are difficult to classify. This unstructured data is where much of the ROT data accumulates, growing silently while organizations focus on more visible structured databases.

Key data points:



⁷ Source: “[Modern Data Infrastructure Dynamics](#),” Hitachivantara.com, 2023.
⁸ Source: Enterprise Strategy Group (now Omdia) Research Report, [The Critical Role of Storage in Building an Enterprise AI Infrastructure](#), September 2025.

Data quality is the top challenge for AI success

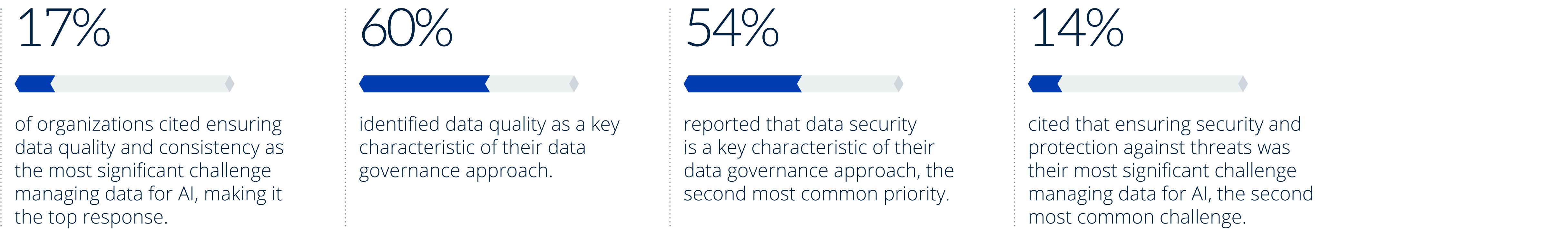
When organizations were asked to identify their most significant challenge in managing data for AI initiatives, one answer rose above all others: **ensuring data quality and consistency**.¹⁰ This should not be surprising. AI models are only as good as the data they are trained on. If an organization feeds an AI system poor-quality data, it will get poor-quality outputs.

Data quality issues manifest in multiple ways. Inconsistent formatting across sources makes integration difficult. Duplicate records create confusion and skew analysis. Outdated information leads to decisions based on conditions that no longer exist. Missing data leaves gaps that models must work around, often poorly.

The consequences extend beyond model performance. When business users cannot trust the insights generated by AI systems, adoption stalls. Teams waste time manually verifying outputs instead of acting on them. The efficiency gains that AI promises begin to evaporate when every recommendation requires human double-checking.

Organizations recognize this challenge. Omdia research shows that data governance efforts focus heavily on data quality, with 60% of organizations identifying it as a key characteristic of their governance approach.¹¹ Yet, recognition and resolution are different things. Achieving consistent data quality across hybrid environments with data scattered across cloud, on-premises, and edge systems requires automated, policy-driven approaches that scale data growth.

Key data points:¹²



¹⁰ Source: Enterprise Strategy Group (now Omdia) Research Report, [Navigating Build-versus-buy Dynamics for Enterprise-ready AI](#), January 2025.

¹¹ Ibid.

¹² Ibid.

The data fitness foundation

Data must move for AI, but moving data is risky

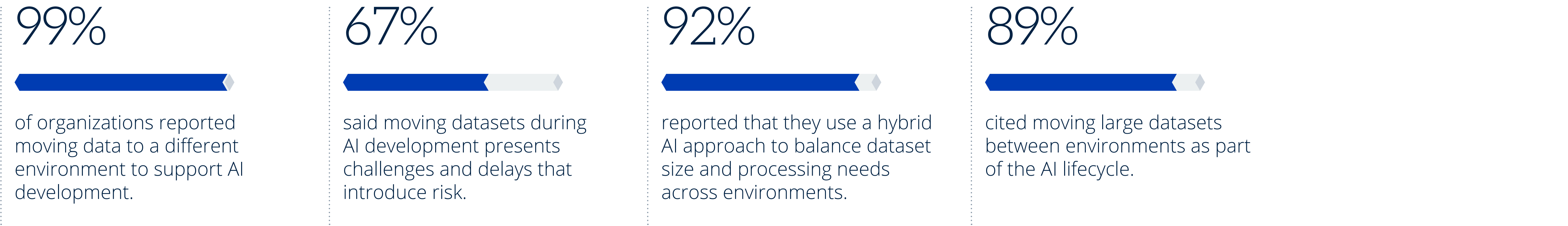
There is an inescapable reality in AI development that data must move. Nearly all organizations reported that they must move data to different environments to support their AI initiatives.¹³ This movement is essential for consolidating relevant data into centralized locations where it can be cleaned, transformed, processed, and prepared for use in model training and deployment.

The reasons for moving data are compelling. Organizations need to integrate more effectively with other systems, support real-time AI applications, ensure compliance with privacy requirements, and leverage advanced AI tools available in target environments. A hybrid approach has become standard, with the vast majority of organizations balancing data across public cloud, on-premises, and edge locations.

But moving data is hard, and most organizations struggle with it. Two-thirds of organizations reported that moving datasets during AI development poses challenges and delays that introduce risk.¹⁴ Data movement can break pipelines, create inconsistencies, and expose sensitive information if not handled properly. Without proper governance and traceability, organizations cannot be confident that the data they move will maintain its integrity and compliance or that they are eliminating ROT data prior to a move, which increases costs and risk.

This tension—the need to move data combined with the difficulty of doing so safely—is precisely why intelligent data lifecycle management from Pentaho has become critical. Organizations need systems that make data movement traceable, reversible, and policy-driven.

Key data points:¹⁵



¹³ Ibid.
¹⁴ Ibid.
¹⁵ Ibid.

The hidden costs of data sprawl

Modern enterprise data is scattered across departments, systems, and geographies. It lives in cloud platforms, on-premises data centers, file servers, SharePoint sites, data warehouses, and countless other locations. This sprawl makes it extraordinarily difficult to understand what data exists, identify what is redundant, and determine what still matters for the business.

The challenges are numerous and interconnected. How do organizations determine the freshness of data when it is spread across dozens of systems? How do they find duplicate data assets consuming disk space in multiple locations? How do they enforce retention policies consistently when data lives in different environments with different tools? How do they identify low-value data sitting in high-cost storage tiers?

Without visibility across the entire data estate, optimization is impossible. Organizations cannot tier data intelligently if they do not know what they have. They cannot enforce retention policies if they cannot find all instances of regulated data. And they cannot eliminate duplicates if they cannot identify them across systems.

Organizations have attempted various approaches to manage this complexity. Nearly half reported using a centralized data management system, while others said they rely on automated pipeline orchestration (45%) or data governance frameworks (44%). Yet a significant portion still depends on manual data management and oversight (43%),¹⁶ an approach that simply cannot scale with the volume and velocity of modern data growth.

Key data points:¹⁷



¹⁶ Ibid.
¹⁷ Ibid.

What is data fitness?

Data fitness is a Pentaho concept with profound implications. It means having data that is organized, classified, cleansed, accessible, properly tiered, and ready for business use, including AI initiatives. Data-fit organizations know what data they have, understand its value, and store it in the right place at the right cost. They can find specific data when needed and trust that it is accurate and compliant.

The opposite of data fitness is data debt. Data debt accumulates when organizations store data without understanding its value, when duplicates proliferate across systems, when obsolete files linger in premium storage, and when no one can answer basic questions about what data exists and where. Data debt costs real money and creates real risk, yet it rarely appears on balance sheets.

Achieving data fitness requires answering three fundamental questions:

- 1. **What data do we have?** This requires discovering and cataloging capabilities that span all data sources, structured and unstructured.
- 2. **What is the data worth?** This requires classification, data quality assessment, and scoring based on business value, access patterns, and regulatory requirements.
- 3. **Where should the data live?** This requires intelligent tiering that matches storage costs to data value.

The concept of data temperature also provides a useful framework. Hot data is frequently accessed and of high value, making it a prime candidate for premium storage. Warm data sees occasional access and can be stored in lower—but easily accessible—tiers. Cold data is rarely accessed and should be stored in the archive. Lastly, frozen data has no value and should be a candidate for purging.

Key capabilities for data fitness include the following:



Discovery and cataloging:
Understand all data across structured and unstructured sources.



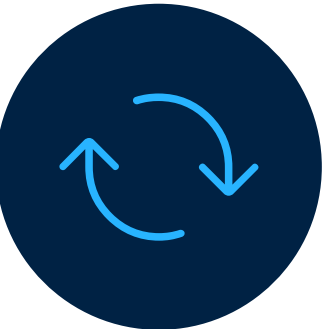
Classification and scoring:
Determine business value, sensitivity, and compliance requirements.



Data quality: Determine the quality and overall trust of the data to be used for AI and decision-making.



Intelligent tiering: Match storage costs to data value with hot, warm, cold, frozen, and archive tiers.



Lifecycle automation:
Policy-driven movement, retention, and purging at scale.

The background is a solid dark blue color. Overlaid on this background is a faint, abstract pattern of white dots and lines. The dots are arranged in a way that suggests a network or a data flow, with some lines connecting the dots. The overall effect is a modern, technological aesthetic.

Pentaho for smarter storage management

Pentaho Data Optimizer for smarter storage

Pentaho Data Optimizer brings smart simplicity to data storage decisions, helping organizations get data fit by cutting waste, reducing risk, and ensuring data is optimized for core business and AI initiatives. As part of Pentaho’s unified platform for intelligent data management, Pentaho Data Optimizer automates how organizations classify, tier, move, archive, and purge data based on its value, usage, and risk profile.

The solution provides clear visibility into what data drives decisions, what is stale, and what is simply costing money without providing value. It works across the full spectrum of enterprise data environments, including Hadoop, file servers, document stores, relational databases, data warehouses, and cloud platforms. This broad connectivity ensures organizations can optimize their entire data estate, not just isolated portions.

What sets Pentaho Data Optimizer apart is its policy-driven approach to lifecycle management. Rather than requiring manual intervention for every data decision, organizations define policies based on business rules, regulatory requirements, and value assessments. The system then enforces these policies automatically and continuously, ensuring data is always in the right place at the right cost.

Integration with Pentaho Data Catalog extends these capabilities further, enabling organizations to discover, understand, and govern data across their enterprise. Combined with AI-powered classification capabilities, organizations can automatically categorize data, including unstructured content, to inform tiering and retention decisions.

Key capabilities of Pentaho Data Optimizer include the following:



Policy enforcement: Automated data lifecycle management is based on business rules and regulations.



Identification of sensitive data: Risks and costs are reduced when sensitive data is identified and removed.



Intelligent tiering: Move data to appropriate storage tiers based on actual access patterns and value.



ROT detection: Identify and eliminate ROT data.



Full traceability: Every action is transparent, auditable, and reversible.

How Pentaho Data Optimizer works: Five steps to data fitness

Pentaho Data Optimizer follows a proven, five-step process that takes organizations from data chaos to data fitness. This structured approach ensures comprehensive coverage while delivering quick wins along the way.

The process begins with **Define**, in which organizations establish terminology for data freshness, duplicates, and obsolescence. Policies are created to manage the data lifecycle, address legal holds, and meet regulatory retention requirements. This foundation ensures all subsequent actions align with business and compliance needs.

Next is **Scan**, where the system connects to enterprise data sources and captures metadata. Content can be evaluated to assess sensitivity, and AI-powered classification helps automatically categorize data. This discovery phase creates the visibility required for optimization.

The **Measure** phase uses a no-code interface to generate rules that evaluate data against ROT criteria. Data is scored based on multiple attributes, including size, usage rate, access patterns, and business purpose. This scoring informs tiering and retention decisions.

Enforce is where action happens. Using the same no-code interface, organizations move, delete, archive, or re-tier data based on policies and scores. Every action is tagged with the reason the action was made, and teams are notified. Data can always be restored if needed. This traceability ensures compliance and confidence with automation, where an organization monitors and sets policies for automating ongoing optimizations.

Finally, **Monitor** provides ongoing visibility through dashboards that visualize storage and costs. Continuous monitoring combined with automated enforcement ensures the organization stays data fit as new data arrives and business needs evolve.

Key differentiators of the Pentaho Data Optimizer solution include the following:



No-code interface: Build rules and policies without programming expertise.



Traceability and reversibility: Every action is logged and can be undone if needed.



Application cataloging: Assess impact on applications and reports before moving data.



Continuous automation: Event- or schedule-driven enforcement keeps data optimized.

From data debt to AI-ready

Data optimization is a first and vital prerequisite for AI success. Organizations cannot build effective AI on disorganized, duplicated, inaccessible data. Before organizations can realize the promise of AI and AI agents, they must first ensure their data foundation is solid.

Pentaho Data Optimizer helps transform data from a liability into a strategic asset and is also part of a larger overall strategy from Pentaho to get data AI-ready.

When data is properly organized and cataloged, AI systems can find and access what they need. When data is deduplicated and cleaned, model training produces more reliable results. When data is properly tiered, organizations can afford to retain the historical data that AI requires without breaking budgets.

The integration of Pentaho Data Optimizer, Pentaho Data Catalog, Pentaho Data Quality, and AI capabilities in the Pentaho platform provides a comprehensive foundation for enterprise AI. Pentaho Data Catalog provides discovery and governance. Pentaho Data Quality enables teams to identify and resolve missing, inconsistent, and inaccurate data before it impacts downstream systems and decision-making. Pentaho Data Optimizer ensures data is properly managed and stored. AI-powered classification helps categorize content automatically. Together, these capabilities ensure that data is not only stored but also ready for use.

This readiness extends to emerging use cases for AI agents. AI agents require governed data access, trusted sources, and traceable lineage to operate safely and effectively. Organizations that lack visibility into their data cannot deploy AI agents with confidence. Those that achieve data fitness can embrace these capabilities, knowing their foundation is solid.

Perhaps most importantly, the cost savings from optimization can directly fund AI initiatives. Getting data fit can free up budget for AI projects that previously seemed unaffordable.

Pentaho provides the following AI-readiness requirements:



Organization and catalogs: Data that can be found and understood.



Accessibility and queryability: Data that AI systems can reach when needed.



Quality and trust: Data that is ready for AI systems and decision-making across the organization.



Proper tiering: Data stored at appropriate cost levels based on value.



Government and traceability: Data with clear lineage and access controls for AI agent use cases.



Delivering business value

Case study: How one CFO found budget for AI

A CFO asked her team an interesting question: “What is our largest preventable cost?” She had already optimized office space, renegotiated supplier contracts, and streamlined operations. She was looking for remaining opportunities.

The answer surprised her: data storage. Her organization’s AWS storage bill totaled £2.4 million annually, while its London office lease cost £2.1 million. The organization was spending more on data storage than on providing a physical office space for its people, and much of that data could not be searched, categorized, or even fully cataloged.

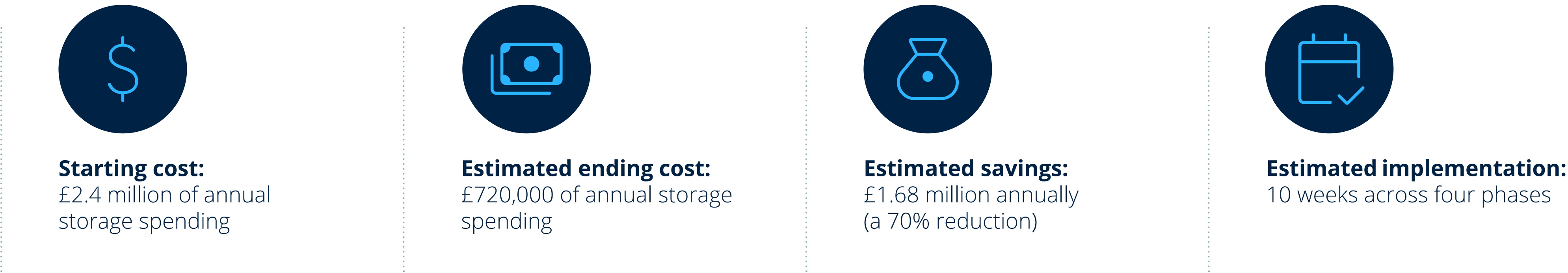
A comprehensive assessment using Pentaho Data Optimizer revealed the scope of the problem: The majority (65%) of stored data had not been accessed in more than two years. Additionally, 23% of the data was duplicated across storage tiers. Premium storage was hosting archived emails that no one read. In total, £1.6 million of annual spending went to data providing no business value.

The organization sees its intelligent data optimization strategy being implemented in four phases: assessment and categorization to gain complete visibility, intelligent tiering to move inactive data to appropriate storage based on actual access patterns, deduplication to eliminate redundant copies while maintaining compliance, and automated policies to prevent future accumulation.

The team expects the project to transform the organization’s cost structure. Estimated annual storage costs would drop from £2.4 million to £720,000—a 70% reduction—saving £1.68 million annually.

The CFO’s reaction to the assessment captured the strategic impact: “We just found budget for the AI initiatives we thought were unaffordable.”

Case study results:



The measurable impact of data readiness

Organizations that invest in data readiness, including data quality, governance, and lifecycle management, see significant improvements across operational and customer-facing metrics. Research confirms that these initiatives deliver tangible business value that extends far beyond the data team.

Data readiness initiatives have significantly improved developer productivity, with more than eight in ten organizations reporting improvement.¹⁸ When developers spend less time hunting for data and more time building solutions, output increases and time to market accelerates. This productivity improvement directly supports AI initiatives that depend on engineering resources.

The strategic impact is equally notable. Organizations increasingly prioritize data and analytics to improve decision-making and strategic planning.¹⁹ This reflects a mature understanding that data capabilities are essential not only for operational efficiency but also for driving revenue and securing a competitive advantage.

The message from our research is clear: AI success depends on data readiness. Organizations that build strong foundations that include quality data, comprehensive governance, and intelligent lifecycle management position themselves to extract maximum value from their AI investments while minimizing risk and ensuring compliance.

Key data points:²⁰

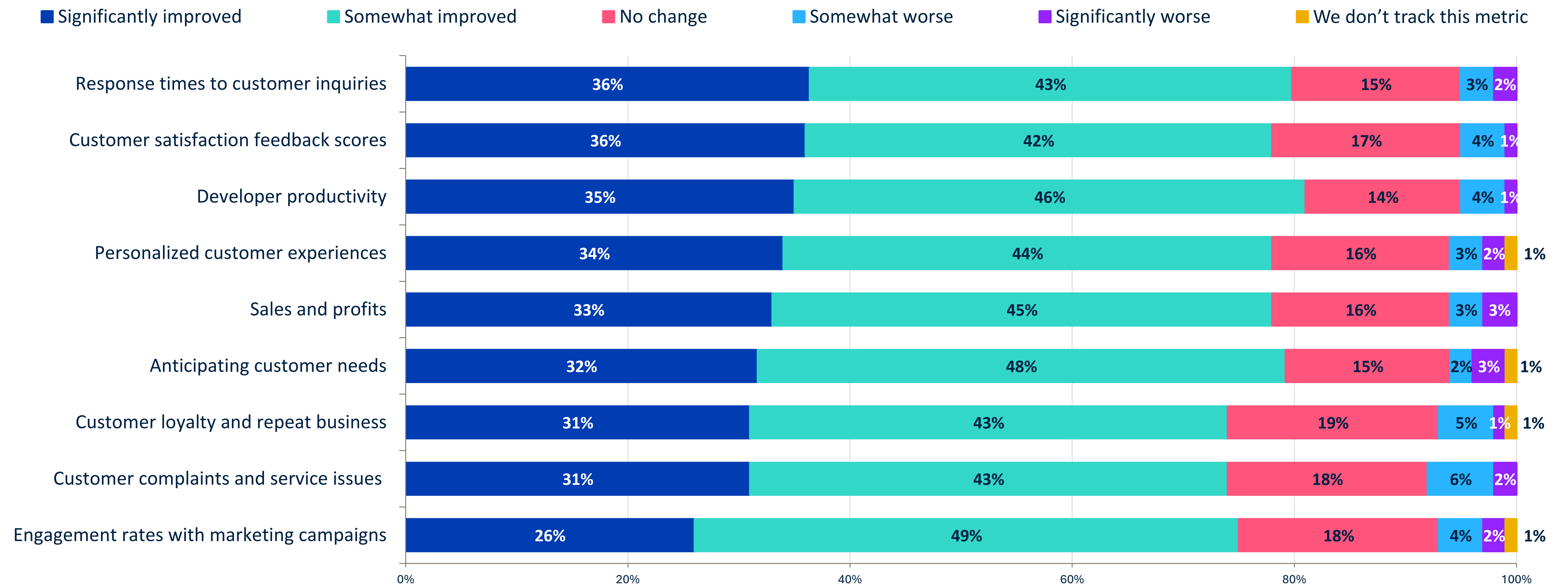


¹⁸ Source: Enterprise Strategy Group (now Omdia) Research Report, [Data Readiness for Impactful Generative AI](#), April 2025.

¹⁹ Source: Omdia Research Report, [Optimizing Cloud Analytics Costs in an Agentic AI Future](#), October 2025.

²⁰ Source: Enterprise Strategy Group (now Omdia) Research Report, [Data Readiness for Impactful Generative AI](#), April 2025.

The impact of data readiness initiatives shows impressive progress



Get data fit – Your path to AI success

The path to enterprise AI runs through data fitness. Organizations that understand their data, manage it intelligently throughout its lifecycle, and store it at appropriate cost levels will be the ones that unlock AI’s transformative potential. Those that continue accumulating data debt will struggle with accuracy, compliance, and trust, limiting their ability to compete in an AI-driven future.

Data optimization is not just about cutting costs, though the savings are substantial. It is about building the foundation that AI requires. Organizations cannot train reliable models on disorganized data. They cannot deploy AI agents safely without governance and traceability. They cannot afford to retain the historical data AI needs if most of their storage budget goes to waste.

Pentaho Data Optimizer provides intelligent lifecycle management that transforms data from a cost center into a strategic asset. With automated classification, policy-driven tiering, deduplication, and full traceability, organizations can finally get their data estate under control. The savings fund innovation, and the organization becomes AI-ready.

The question is not whether to invest in data fitness; it is how quickly an organization can begin.

Three questions every organization should ask are:

1.

What percentage of our data storage spending provides active business value?

2.

Can we find specific data when the business needs it?

3.

Is our data structured and ready for AI consumption?

Pentaho

ABOUT

Trusted by over 70% of the Fortune 100, Pentaho solutions enable the foundational and AI-grade data fitness that drives faster and more confident decision-making.

Pentaho Data Optimizer is part of a unified end-to-end data integration and intelligence platform. It helps reduce waste, ensure compliance, and prioritize high-value data for innovation and AI-readiness. Combined with Pentaho Data Catalog for discovery and governance and Pentaho Data Integration for data movement and transformation, organizations gain complete control over their data estate. Pentaho Data Quality enables teams to identify and resolve missing, inconsistent, and inaccurate data—before it impacts downstream systems and decision-making.

The platform works across the full spectrum of enterprise environments, including Hadoop, file servers, document stores, relational databases, data warehouses, and cloud platforms. Whether data is structured or unstructured, on premises or in the cloud, Pentaho provides the visibility, automation, and governance that modern data management requires.

In a world accelerated by AI, an organization’s trust in its data begins with managing it well. Pentaho gives organizations the confidence to act, knowing their data is accurate, secure, and worth the cost of keeping.

Ready to get data fit? See how Pentaho Data Optimizer can cut your storage costs, reduce risk, and prepare your data for AI.

Learn More

Request a Demo



©2026 TechTarget, Inc. d/b/a Informa TechTarget. All rights reserved. The Informa TechTarget name and logo are subject to license. All other logos are trademarks of their respective owners. Informa TechTarget reserves the right to make changes in specifications and other information contained in this document without prior notice.

Information contained in this publication has been obtained by sources Informa TechTarget considers to be reliable but is not warranted by Informa TechTarget. This publication may contain opinions of Informa TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent Informa TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, Informa TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of Informa TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at cr@esg-global.com.



Omdia provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

© 2026 TechTarget, Inc. All Rights Reserved. Unauthorized reproduction prohibited.