Omdia
by informa techtarget

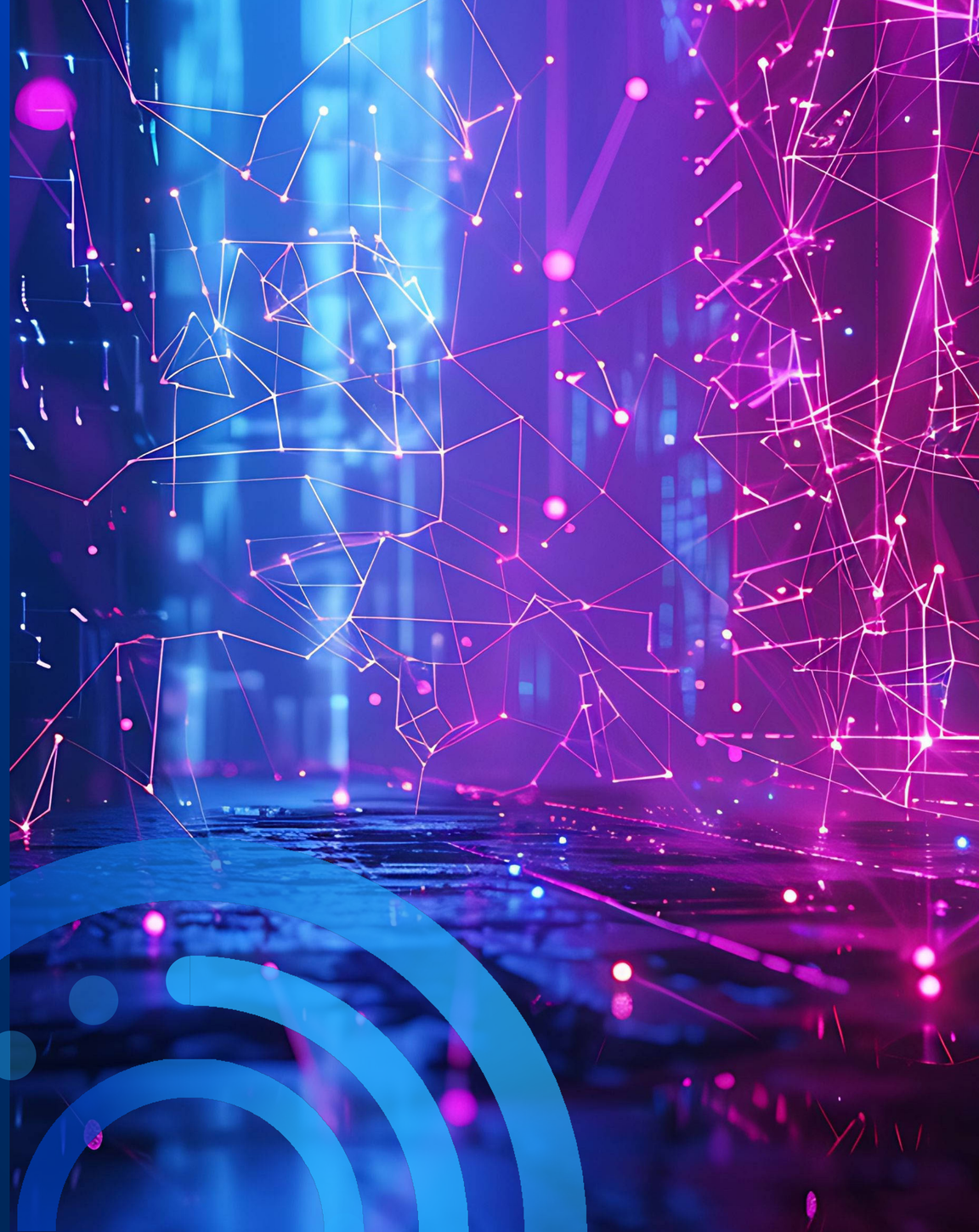
How Automated Data Lineage and Metadata Management Empower Enterprise AI

Cloudera Data Lineage on Microsoft Azure for Hybrid Data Governance

Stephen Catanzano | *Senior Analyst*

March 2026

This Omdia eBook was commissioned by Cloudera and is distributed under license from Techtarget, Inc.



Key findings

Back to Contents

Introduction

The race to AI is on, but your data may not be ready

Organizations worldwide are accelerating their AI initiatives, driven by the promise of improved operational efficiency, enhanced customer experiences, and competitive advantage. Yet beneath the excitement lies a sobering reality that most enterprises aren't prepared to deliver on AI's potential because their data foundations are fragmented, ungoverned, and poorly understood.

The challenge isn't a lack of data; it's a lack of visibility. Enterprise data teams face mounting pressure to modernize and adopt AI, but inconsistent governance, the divide between on-premises and cloud environments, and a limited understanding of data lineage makes progress slow, risky, and expensive. Without knowing where data comes from, how it is transformed, and who has the right to use it, organizations cannot build the trustworthy AI systems their businesses demand.

This eBook explores the data readiness challenge and how Cloudera Data Lineage and Microsoft Azure together deliver a unified, governed foundation for enterprise AI.

Key data points:

47% of organizations cited improved operational efficiency as a top motivator for AI adoption.[1]

70% ranked data quality as a high or very high priority for AI-driven initiatives.[2]

80% of organizations said AI agents are a high or top priority compared to other AI initiatives.[3]

[1]Source: Enterprise Strategy Group (now Omdia) Research Report, Navigating Data Governance in the Age of AI, September 2024.
[2]Ibid.
[3]Source: Enterprise Strategy Group (now Omdia) Research Report, AI Agents: The Game-changing Generative AI Use Case, August 2025.

The enterprise data challenge

Data complexity is growing faster than organizations can manage

Today's enterprises operate in increasingly complex data ecosystems. The majority of organizations collect data from hundreds of sources daily, spanning on-premises systems, multiple clouds, SaaS applications, and edge devices.[4] This explosion of data sources, combined with the proliferation of AI initiatives, has created unprecedented challenges for data teams.

The complexity isn't just about volume. Organizations must manage structured and unstructured data across hybrid environments while ensuring quality, security, and compliance. More than half of enterprise data is now unstructured, including documents, images, and logs that are critical for AI but notoriously difficult to govern.

As data sprawl continues and AI adoption accelerates, the gap between data collection and data readiness widens. Organizations need scalable solutions to streamline data ingestion, ensure quality, and maintain governance across their entire data estate.

Key data points:

64% of organizations reported collecting data from 100 to 499 sources daily.[5]

54% of enterprises have 1 PB or more of data under management, with an average of 2.94 PB.[6]

51% of data under management is unstructured.[7]

21% of organizations experienced more than 40% annual data growth.[8]

[4]Source: Enterprise Strategy Group (now Omdia) Research Report, Data Readiness for Impactful Generative AI, April 2025.
[5]Ibid.
[6]Source: Enterprise Strategy Group (now Omdia) Research Report, Achieving Cyber and Data Resilience, September 2024.
[7]Ibid.
[8]Ibid.

Back to Contents

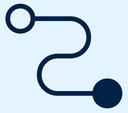
Most organizations don't understand their data's journey

Data lineage is about understanding how data originated, how it changes over time, and the rights any individual or group within an organization has to use it, which is fundamental to building trustworthy AI. Yet our research revealed a striking gap: The vast majority of organizations have limited visibility into their data's journey from source to insight.[9]

This lineage gap creates significant risks. Without understanding data origins and transformations, organizations cannot ensure the accuracy of AI model outputs, comply with regulations requiring data traceability, or confidently modify data pipelines without triggering downstream failures.

The consequences are tangible: Data errors delay decisions and can cost organizations hundreds of thousands of dollars per hour in downtime, cloud migrations fail at alarming rates due to hidden dependencies, and compliance audits become lengthy, manual exercises that drain resources and introduce risk.

Key data points:



80% of organizations understand 50% or less of where the data for their AI processes comes from (i.e., its data lineage).[10]

42% of cloud migrations fail due to hidden data dependencies, according to Cloudera.

The average cost per hour of downtime from data errors causing decision delays is $300,000, according to Cloudera.

Manual processes extend audit preparation time by 72%, according to Cloudera.

[9]Source: Enterprise Strategy Group (now Omdia) Research Report, Navigating Data Governance in the Age of AI, September 2024.
[10]Ibid.

Back to Contents

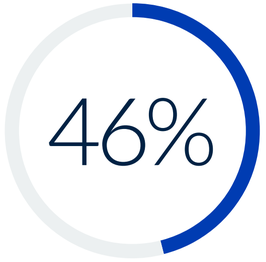
When you can't trust your data, you can't trust your AI

Trust is the foundation of data-driven decision-making. When business users question whether the data they see is accurate, complete, and up to date, productivity suffers. Employees waste time verifying information, second-guessing insights, and hedging decisions that should be straightforward.

This trust deficit becomes especially critical in the context of AI. Generative AI and AI agents amplify whatever data they use and run on, accurate or not. If organizations feed AI models with data of uncertain provenance and quality, they risk automating and scaling poor decisions across the enterprise.

The research is clear that organizations recognize data quality and integrity as essential to AI success. Yet nearly half reported only partial trust in the data given to employees for decision-making.[11] Closing this trust gap requires complete transparency into data origins, transformations, and quality, exactly what automated data lineage provides.

Key data points:



of organizations distrust or only somewhat trust the data given to employees for decision-making.[12]

expected AI agents to improve their efficiency and automation,[13] which can't happen with strong data trust.

of organizations feed 21% to 50% of their data into AI models,[14] indicating room to optimize data utilization.

[11]Ibid.
[12]Ibid.
[13]Source: Enterprise Strategy Group (now Omdia) Research Report, AI Agents: The Game-changing Generative AI Use Case, August 2025.
[14]Source: Enterprise Strategy Group (now Omdia) Research Report, Data Readiness for Impactful Generative AI, April 2025.

Back to Contents

The foundation for AI-ready data

Metadata management: The key to knowing your data

Metadata is data about the data and provides the critical context organizations need to understand, govern, and trust their information assets. For AI initiatives, metadata management is especially crucial since it enables teams to identify relevant training data, ensure compliance with usage rights, track model inputs for explainability, and maintain audit trails.

Yet despite its importance, metadata management remains a significant gap for most organizations. Research shows that only 41% of enterprises comprehensively classify their data with metadata,[15] leaving them blind to critical information on data origins, ownership, sensitivity, and quality.

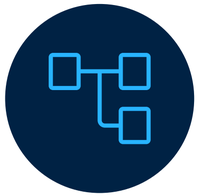
Organizations that do invest in metadata management see clear benefits. Those using both metadata analysis and content examination achieve the most comprehensive understanding of their data. The combination enables precise data discovery, accurate impact analysis, and confident governance, all prerequisites for trustworthy AI.

Key data points:

81%
of organizations reported leveraging data classification tools or processes.[16]

54%
reported using both metadata analysis and file examination for classification, which is considered a best practice.[17]

Only 41%
reported that the majority of their data is classified by metadata.[18]



36%
said they use metadata management tools for lineage tracking.[19]

[15]Source: Enterprise Strategy Group (now Omdia) Research Report, Navigating Data Governance in the Age of AI, September 2024.
[16]Ibid.
[17]Ibid.
[18]Ibid.
[19]Source: Enterprise Strategy Group (now Omdia) Research Report, Data Readiness for Impactful Generative AI, April 2025.

Back to Contents

How organizations track data lineage today and where they fall short

Organizations employ a variety of tools and methods to track data lineage, ranging from sophisticated AI-powered solutions to manual spreadsheets. The most forward-thinking enterprises leverage automated tools and cloud provider solutions to scale their lineage tracking, while others rely on collaborative management between data and compliance teams.[20]

However, significant gaps remain. Lower adoption of foundational capabilities like metadata management and annotation tools[21] indicates that many organizations lack the building blocks for comprehensive lineage tracking. The result is fragmented visibility that spans some systems but leaves critical blind spots in others.

The opportunity is clear. Organizations need integrated platforms like Cloudera Data Lineage that combine automated lineage tracking with enhanced metadata capabilities to support end-to-end data governance. Such solutions must operate across hybrid environments, spanning on-premises systems, multiple clouds, and diverse technologies to provide the unified visibility that AI initiatives demand.

Key data points:

57%

of organizations said they use AI tools for data lineage management, making it the top cited method.[22]

48%

reported using cloud provider-supplied tools.[23]

47%

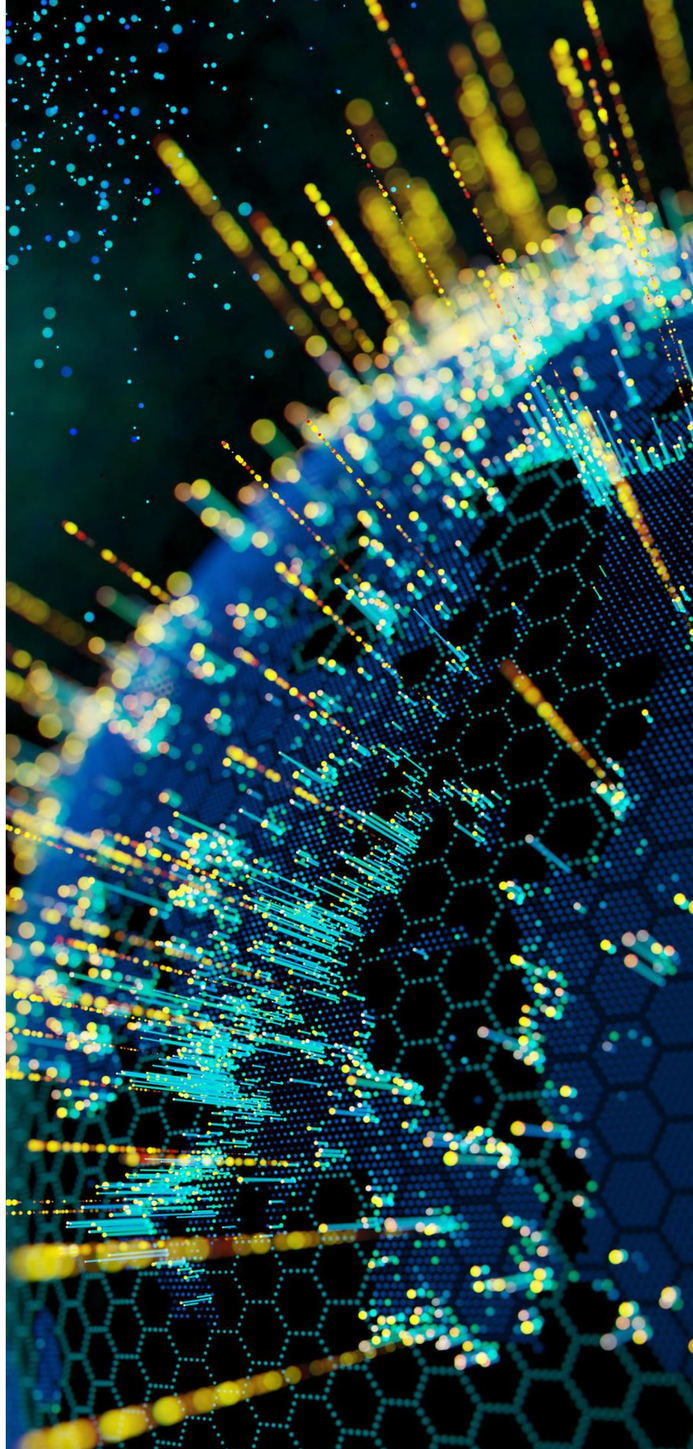
said they rely on automated tools.[24]

Only 30%

reported using annotation and tagging tools.[25]

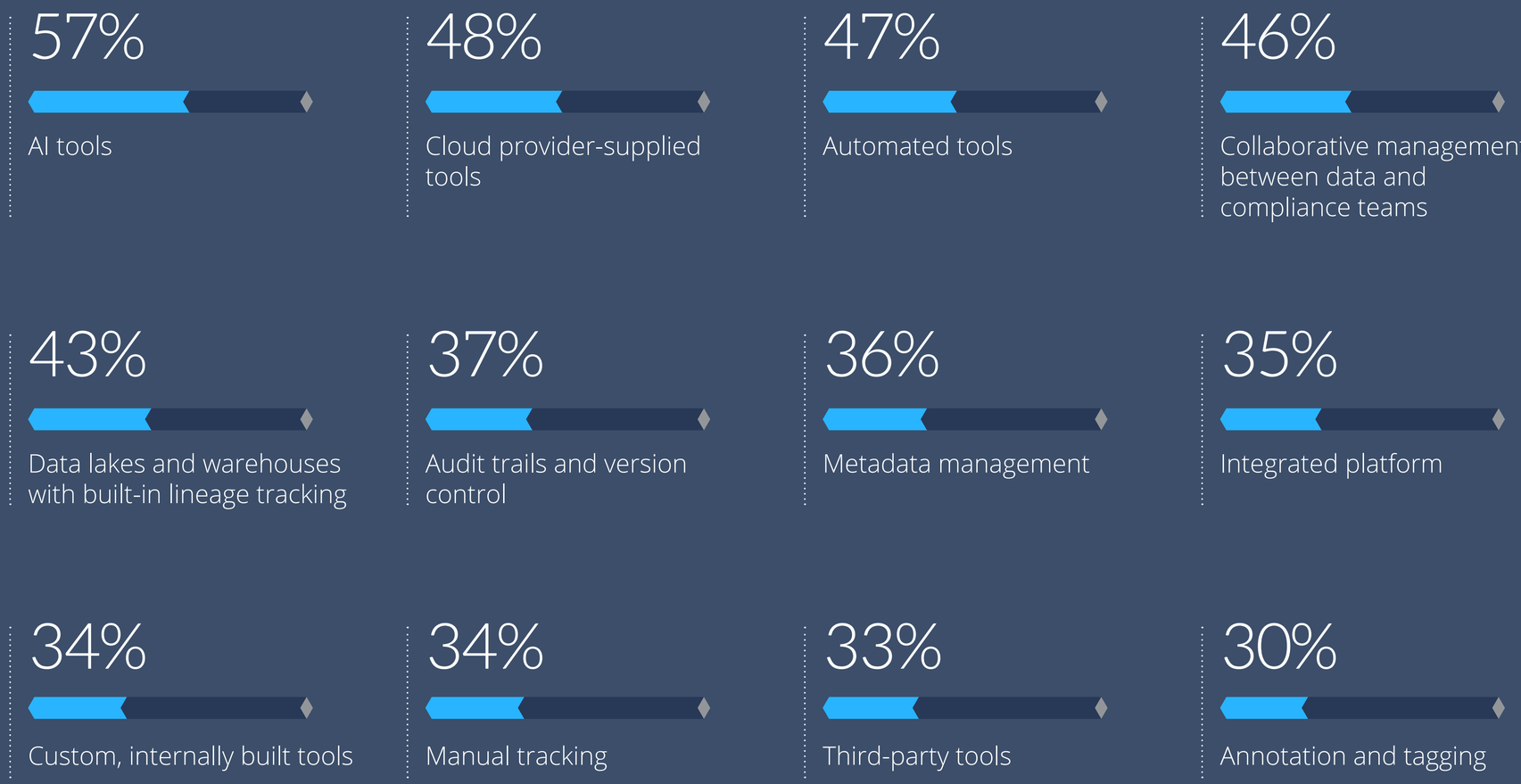
14+

hours/week are lost by teams waiting for subject matter experts due to tribal knowledge silos, according to Cloudera.



[20]Ibid.
[21]Ibid.
[22]Ibid.
[23]Ibid.
[24]Ibid.
[25]Ibid.

Data lineage tools highlight transparency efforts



Back to Contents

The rise of AI agents makes data governance more critical than ever

AI agents are autonomous systems that can execute tasks, make decisions, and interact with other systems without human intervention, and they represent the next frontier of enterprise AI. Organizations are prioritizing AI agents for their potential to dramatically improve efficiency, automate complex workflows, and unlock new capabilities.

But AI agents also raise the stakes for data governance. These systems access, process, and act on data across organizational boundaries, requiring unprecedented levels of trust, traceability, and control. Organizations that lack visibility into their data lineage will struggle to deploy AI agents safely, ensure compliance, and maintain accountability for automated decisions.

The integration challenges are already emerging. Most organizations reported facing significant hurdles connecting AI agents to existing systems and processes. Success requires a governed data foundation that provides clear lineage, consistent metadata, and robust access controls, regardless of whether data resides on premises or in the cloud.

Key data points:

[26]Source: Enterprise Strategy Group (now Omdia) Research Report, AI Agents: The Game-changing Generative AI Use Case, August 2025.
[27]Ibid.
[28]Ibid.
[29]Ibid.
[30]Ibid.

Back to Contents

Cloudera and Microsoft: A unified solution

Cloudera Data Lineage: Complete visibility for hybrid data ecosystems

Cloudera Data Lineage is the leading automated data lineage, discovery, and catalog solution and enables enterprise data teams to find, understand, and manage their metadata with unprecedented efficiency. The solution provides complete visibility across complex hybrid data environments, helping teams execute changes with confidence, ensure trustworthy data, and accelerate strategic initiatives.

What sets Cloudera Data Lineage apart is its comprehensive, multi-dimensional approach to lineage. **The solution tracks data flows at three levels: cross-system lineage, providing end-to-end visibility across an organization's data landscape; inner-system lineage, revealing column-level transformations within processes; and end-to-end asset lineage, connecting source systems to final consumers.**

With 100% automated metadata extraction and integration and 60+ technologies out of the box, Cloudera Data Lineage delivers value in hours, not weeks, according to Cloudera. The SaaS-based solution requires no resource overhead, supports both legacy on-premises and modern cloud systems, and delivers immediate time to value.

Key data points:

+60	100%	3-48	90%	Up to 75%
source integrations out of the box.	automated metadata extraction.	hour time to value.	time savings on root cause analysis, impact analysis, and migration.	reduction in cloud migration time.

Back to Contents

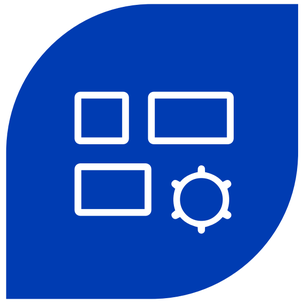
See your entire data journey with multidimensional lineage

Traditional lineage tools provide only partial visibility, showing connections between systems but missing the critical transformations that occur within processes. Cloudera Data Lineage revolutionizes metadata management through comprehensive, multidimensional lineage that illuminates data flows at every level.

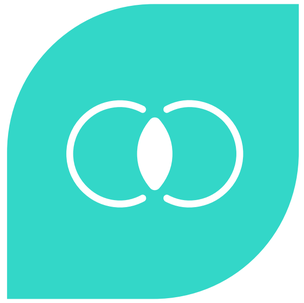
- Cross System Lineage provides end-to-end visibility from source systems to final reports, enabling impact analysis and technology flow mapping across an organization's entire data landscape.
- Inner System Lineage reveals column-level detail and field-level transformations within ETL processes, semantic layers, and BI tools, exposing all expressions and logic applied to an organization's data.
- End-to-End Asset Lineage connects every asset from the source system to the end consumer, providing complete traceability.

This multidimensional approach, powered by Cloudera Knowledge Graph, uncovers hidden data relationships that traditional tools miss. The result is smarter decisions, faster troubleshooting, and confident governance across an organization's entire data estate, whether the data resides on premises, in Microsoft Azure, or across multiple clouds.

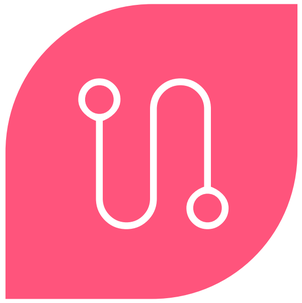
Key data points:



Cross system
End-to-end visibility, impact analysis, technology flow mapping.



Inner system
Column-level detail, script visualization, transformation transparency.



End-to-end
Asset-level lineage from source to consumer.

Knowledge graph
Unveils hidden data relationships for smarter decisions.

Discovery
Cloudera Data Lineage's discover module provides an automated, centralized visual map to instantly locate and identify metadata across an organization's entire data landscape.

Back to Contents

Cloudera and Microsoft: A unified foundation for enterprise AI

Enterprises struggle to modernize and operationalize AI because their data remains fragmented and inconsistently governed across on-premises and cloud environments. Cloudera and Microsoft together address this challenge by delivering a unified, governed “cloud anywhere” data foundation.

Cloudera brings its hybrid (cloud anywhere) platform with enterprise AI (AI anywhere) capabilities and open data lakehouse (data anywhere) architecture with unified security and governance with Cloudera Shared Data Experience (SDX), and petabyte-scale performance for mission-critical workloads. Microsoft Azure provides the trusted cloud foundation enterprises standardize on, with elastic economics, a broad native ecosystem, and enterprise-grade security and compliance.

Together, these capabilities unlock high-value on-premises data for analytics and AI on Microsoft Azure, build a unified governed foundation combining Microsoft Azure’s cloud-scale services with Cloudera’s SDX and Data Lineage, and enable real-time analytics and AI using streaming, lakehouse, and hybrid AI services. The result is a complete modernization path from legacy to hybrid to AI that reduces risk and accelerates value.

Cloudera/Microsoft Joint Value Pillars:

Back to Contents

Seamless integration with Microsoft Fabric and Power BI

As organizations adopt Microsoft Fabric and Power BI to generate insights from their data, the complexity of managing, governing, and ensuring data integrity grows. Power BI projects form the foundation of most BI strategies, but native tools often fall short in providing the depth of visibility needed for large, intricate dataflows spanning multiple systems and sources.

Cloudera Data Lineage is a robust solution that provides strong lineage across the Microsoft ecosystem. Cloudera Data Lineage tightly integrates with Power BI and transforms how business intelligence teams manage automated data governance, compliance, and impact analysis.

Some of the key integration capabilities of Cloudera Data Lineage include achieving transparency from data ingestion to final report; gaining full control over every dataflow, measure, and proactive insight to catch issues before they impact reports; and tracking display name changes and report-native measures that often mask true data lineage in BI environments.

Microsoft Ecosystem Coverage:

- Power BI Projects, Dataflows, and Composite Models
- Azure Synapse Analytics, Azure Data Factory
- Azure SQL, Databricks on Azure
- SQL Server (SSIS, SSRS, Analysis Services)
- Full support for semantic models and report-native measures

Back to Contents

Delivering business value

Immediate ROI built in from day one

Cloudera Data Lineage delivers measurable business value from the first day of deployment. By automating manual lineage tracking and documentation, organizations dramatically accelerate critical processes while reducing costs and improving data quality.

According to Cloudera, teams save over 90% of the time previously spent on root cause analysis, change impact analysis, migration planning, and decommissioning of legacy systems. Additionally, cloud expenditures decrease by at least 25% through continuous cleansing routines that combat "data obesity" and identify unused tables and redundant pipelines. Production stability also improves tenfold as proactive change impact analysis preemptively addresses potential issues before they cascade into outages.

These efficiency gains translate directly to strategic value. Data teams previously consumed by manual documentation and firefighting can pivot to high-value initiatives. Migration projects are also able to be completed in half the budgeted time and cost, according to Cloudera, and organizations can build the trusted data foundation required for AI initiatives that deliver competitive advantage.

Key ROI Metrics:

90%
time savings on root cause analysis, impact analysis, migration.

25%
cost reduction through optimized cloud spending.

10x
improvement in production stability.

Up to 99%
time reduction: "What took weeks now takes hours or minutes."

"We've seen **major time** savings of up to 99%. What took weeks in the past now takes hours or even minutes; it's insane!"

— Mark Horseman,
Alberta Institute of Technology

Back to Contents

Solving the critical challenges that block AI success

Every organization pursuing AI faces common challenges that slow progress and increase risk. Cloudera Data Lineage directly addresses these barriers with automated, comprehensive solutions.

For cloud migration issues, the solution automatically maps cross-system dependencies, cutting down migration times and dramatically reducing failure rates caused by hidden dependencies. For business disruption, visual lineage tracing reduces root cause analysis from days to minutes, minimizing the high cost of data-related downtime.

Automated documentation breaks tribal knowledge silos and eliminates delivery bottlenecks. This automation helps organizations avoid compliance penalties, as automated lineage and governance transform metadata into audit-ready evidence, while eliminating much of the audit preparation time caused by manual processes.

For those experiencing unexpected outages, automated dependency mapping predicts impacts before deployment, preventing changes from triggering cascading failures. Cloud spending waste also decreases, as teams identify unused tables and redundant pipelines, which consume a high percentage of typical cloud budgets.

	Challenge	Solution
	Cloud migration	Map dependencies, cut migration time
	Business disruption	Root cause analysis: days → minutes
	Delivery bottlenecks	Automated documentation, self-service discovery
	Compliance penalties	Automated audit-ready lineage and governance
	Unexpected outages	Predict impacts before deployment
	Cloud spending waste	Identify orphaned and redundant resources

Back to Contents

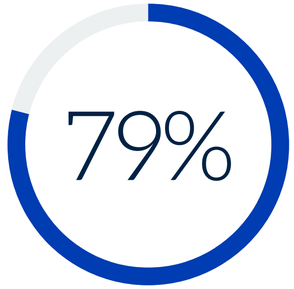
The measurable impact of data readiness initiatives

Organizations that invest in data readiness, including data lineage, metadata management, and governance, 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.

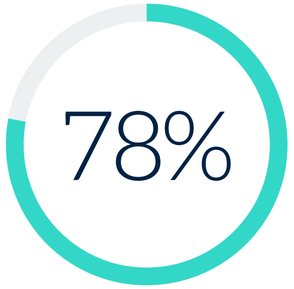
Response times to customer inquiries and customer satisfaction scores show significant improvement, driven by faster access to trusted data and more reliable insights. Developer productivity increases as teams spend less time hunting for data and more time building solutions. And organizations gain the ability to anticipate customer needs and deliver personalized experiences based on comprehensive, trusted data.[31]

The message is clear: AI success depends on data readiness. Organizations that build strong foundations, with complete lineage, comprehensive metadata, and unified governance, position themselves to extract maximum value from their AI investments while minimizing risk and ensuring compliance.

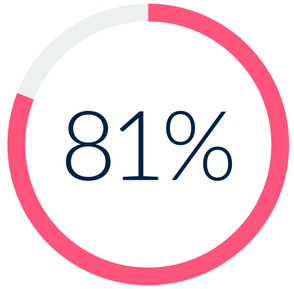
Key data points:



of organizations reported improved response times to customer inquiries (36% reported significant improvement).[32]

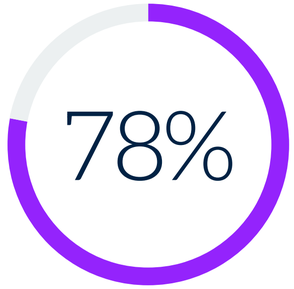


reported improved customer satisfaction scores (36% reported significant improvement).[33]



reported improved developer productivity (35% reported significant improvement).[34]

reported an improved ability to anticipate customer needs (32% reported significant improvement).[35]

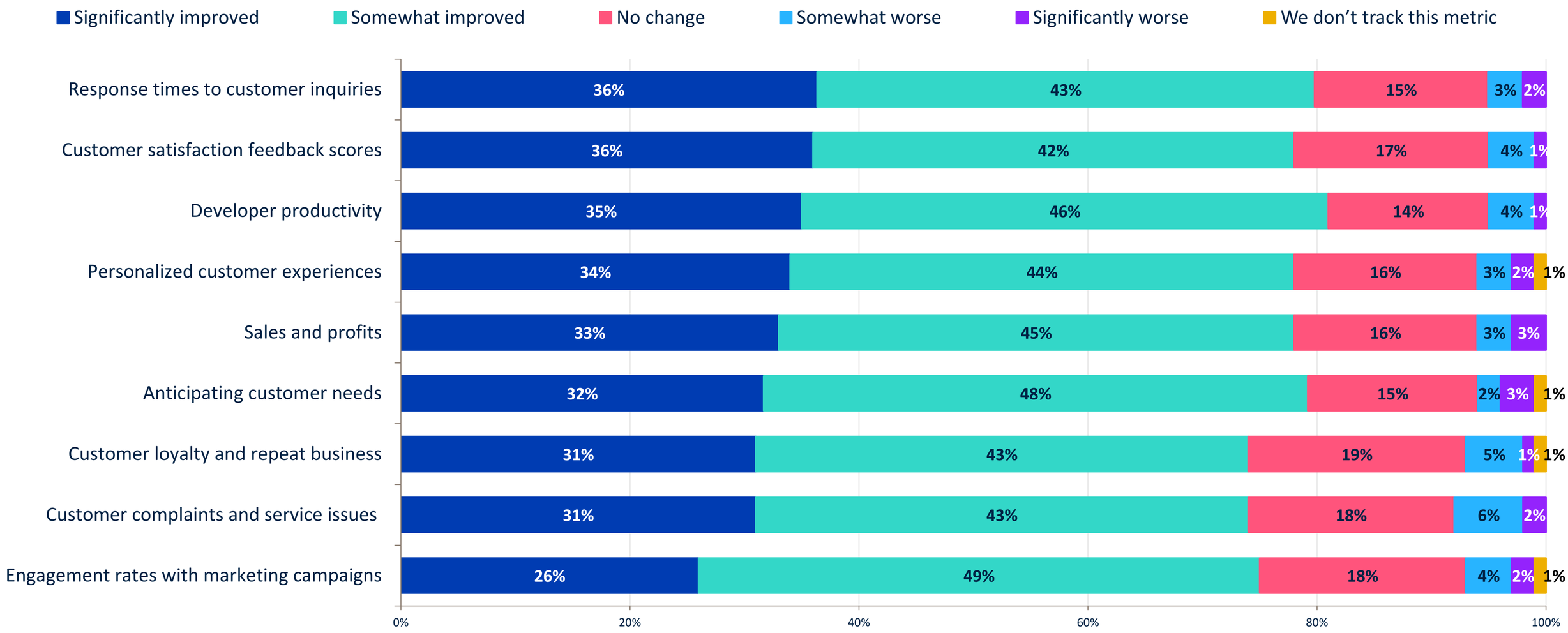


reported improved personalized customer experiences (34% reported significant improvement).[36]

[31] Source: Enterprise Strategy Group (now Omdia) Research Report, Data Readiness for Impactful Generative AI, April 2025.
[32] Ibid.
[33] Ibid.
[34] Ibid.
[35] Ibid.
[36] Ibid.

Back to Contents

Impact of data readiness initiatives shows impressive progress



Back to Contents

Build an AI-ready data foundation with Cloudera Data Lineage and Microsoft Azure

The path to enterprise AI runs through Cloudera Data Lineage. Organizations that understand their data's lineage, manage metadata comprehensively, and govern data consistently across hybrid environments will be the ones that unlock AI's transformative potential. Those that don't will struggle with accuracy, compliance, and trust, limiting their ability to compete in an AI-driven future.

Cloudera Data Lineage and Microsoft Azure together deliver the unified, governed foundation that makes AI success possible. With Cloudera Data Lineage's automated data lineage and metadata management running on Microsoft Azure, organizations can gain complete visibility across complex hybrid environments, execute changes with confidence, ensure trustworthy data, and accelerate strategic AI initiatives.

The technology is proven at scale, with Cloudera customers collectively managing nearly 30 exabytes of data, according to Cloudera. Additionally, the integration is seamless, with 60+ connectors working out of the box. The value is immediate, and teams see results in hours, not weeks. The question isn't whether to invest in data readiness; it's how quickly an organization can begin.

Key Takeaways:

- 80% of organizations do not understand their data lineage on most of their data—you do not have to be one of them.[37]
- Automated lineage eliminates manual tracking and tribal knowledge silos.
- Hybrid governance enables AI across on-premises and cloud environments.
- Immediate value with 3-48 hour time to deployment, according to Cloudera.
- Proven scale, with 30 exabytes under management across Cloudera customers, according to Cloudera.

[37]Source: Enterprise Strategy Group (now Omdia) Research Report, Navigating Data Governance in the Age of AI, September 2024.

Back to Contents

CLOUDERA

ABOUT

Cloudera is the only hybrid data and AI platform company that large organizations trust to bring AI to their data anywhere it lives. Unlike other providers, Cloudera delivers a consistent cloud experience that converges public clouds, on-prem data centers, and the edge, leveraging a proven open source foundation. As the pioneer in big data, Cloudera empowers businesses to apply AI and assert control over 100% of their data, in all forms, improving security, governance, and real-time and predictive insights. The world's largest brands across all industries rely on Cloudera to transform decision-making and ultimately boost bottom lines, safeguard against threats, and save lives.

Cloudera's open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring generative AI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to solve what seems impossible today and in the future.

Cloudera Data Lineage extends this foundation with automated metadata management and comprehensive lineage, ensuring organizations can trust, govern, and leverage their data for AI initiatives. Together with Microsoft Azure, Cloudera delivers the unified hybrid data platform that enterprise AI demands.

Ready to get your data AI-ready? See how Cloudera Data Lineage can deliver visibility, compliance, and cost reduction for your organization.

Learn More

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.