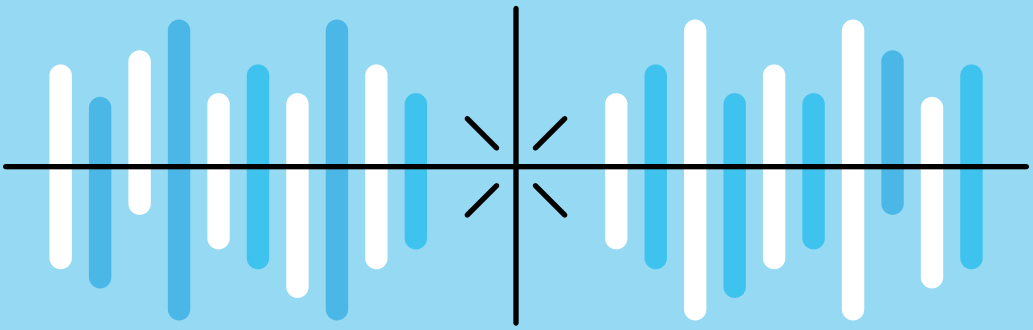
HARVARD BUSINESS REVIEW



Taming the Complexity of AI Data Readiness

HBR Analytic Services **PULSE SURVEY**

Sponsored by

CLOUDERA

The Model Is Not the Mission: Why Bringing AI to Your Data Defines AI Success

Sergio Gago
Chief Technology Officer
Cloudera

So far, the industry conversation about artificial intelligence (AI) has been mainly about models. But as any AI architect will tell you, what determines successful enterprise AI isn't the algorithm—it's the data. Data is the bedrock upon which all transformative intelligence is built. When data is trusted, accurate, and timely, organizations can generate credible, tailored insights that help them see what is happening, understand the impact, and take actions that enhance differentiation within their marketplaces. Without a trustworthy foundation, even the most sophisticated models will fail to deliver reliable, scalable, and differentiated business value.

Our current data landscape is complex, messy, and constantly evolving. Ensuring data is AI-ready is immensely challenging; studies have found only a small fraction of organizations consider their data completely ready for AI adoption, and leaders overwhelmingly identify siloed data and the difficulty of integrating data sources as their greatest hurdles in preparing data for AI use.

To overcome these challenges, organizations must solve for "data gravity," the force that keeps data sets grounded in place. While typically associated with the size of data (we know that massive data sets often root themselves in their locations), data gravity also encompasses sovereignty, security, trust, and cost. Using modern architectures—and concepts like zero copy data access and containerized AI processing—organizations can bring AI to the data, wherever it resides, rather than moving the data to the AI model. Bringing AI to the data reduces latency and security risks and avoids potential costs and risks associated with data movement and duplication.

AI-ready data is fundamental to realizing transformative business insights and unlocking AI's full potential—the real business value obtained when proprietary data drives bespoke insight. The ability to leverage data that's sourced from proprietary, protected sources within an organization for use in AI models is essential for making sound business decisions and protecting credibility with customers, shareholders, and business partners.

This report, which Cloudera has sponsored in association with Harvard Business Review Analytic Services, delves into these critical themes. It offers a strategic view of how enterprises can apply cloud and agentic AI technologies to build a robust data pipeline, establish rigorous data governance, and ultimately unlock the competitive edge that only AI-ready data can provide. The time to build a trusted, scalable data foundation is now.

Taming the Complexity of AI Data Readiness

On the face of it, the data required for artificial intelligence (AI) can be analogized to oil. Both raw materials are plentiful and potentially valuable, but each requires complex refining to unlock its value. One critical difference is that oil can sit in storage almost indefinitely, while valuable data has a shelf life. Customer behaviors shift. Market conditions evolve. Regulatory landscapes change. The longer the data sits unprocessed, the less accurately it assesses the present or forecasts the future. Accurate and timely AI data is far from being a commodity.

OIL REFINING IS a mature industry. But solving AI data challenges remains a work in progress for many organizations, and quality requirements differ markedly across business units and industries. Yet regardless of operational stakes or use-case differences, organizations with poor AI data quality face common risks. An organization may compile terabytes of business data, but if the data doesn't help answer important questions, it is merely expensive storage. Raw data often contains hidden biases, sampling errors, duplicate records, or myriad inconsistencies. Siloed or inaccessible data perpetuates knowledge gaps. AI trained on unprocessed, incomplete, or poorly processed data will produce unreliable outputs, squandering time, talent, and infrastructure resources.

"Most companies have too much data, and yet they can't find the trusted data when they need to answer their questions," says Teresa Tung, who leads the global data practice for Accenture, a Dublin-based management consultancy. Getting to the desired state of "data readiness," she says, "means that you can access data to see an accurate view of what's happening in your business and what you can do about it."

Trust is a high bar that many organizations have yet to reach. A new Harvard Business Review Analytic Services study of 231 respondents from the *Harvard Business Review* audience (all involved in their

HIGHLIGHTS

73%

of respondents agree that their organization should prioritize artificial intelligence (AI) data quality more than it currently does.

73%

of respondents agree that their organization has found the processing and preparing of data for AI to be challenging.

65%

expect many of their organization's business processes will be augmented or replaced by agentic AI in the next two years.

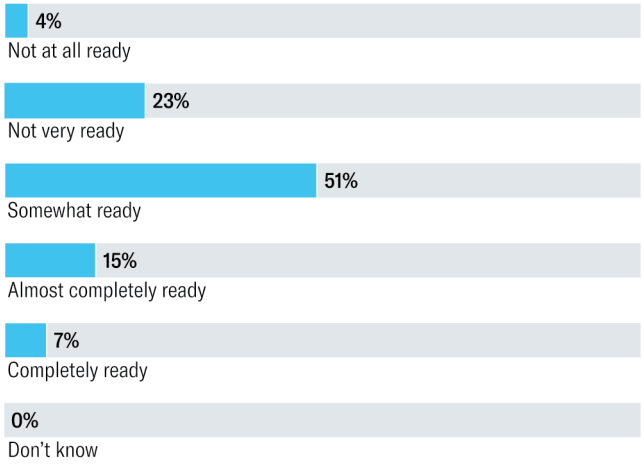
Due to rounding, some figures in this report may not add up to 100%.

FIGURE 1

Enterprise Data Isn't Ready for AI Adoption

But nearly two-thirds of respondents believe they're edging closer to it

Overall, how ready is your organization's data for AI adoption?



Base: 231 respondents

Source: Harvard Business Review Analytic Services survey, October 2025

organization's data decisions and at organizations at least exploring the use of AI) finds that 42% of them somewhat or strongly agree that their organization has high trust in its AI data. The majority of respondents acknowledge the inherent dangers of poorly prepared AI data. Yet paradoxically, nearly three in four (73%) somewhat or strongly agree their organization should prioritize AI data quality more than it currently does.

"We are nowhere near having high-quality, good data available," explains Beena Ammanath, a managing director and executive director of the Global Deloitte AI Institute at Deloitte Consulting LLP, the U.S. consulting arm of London-based Deloitte LLP. "Legacy data is one of the top issues holding companies back. It's the number one challenge most organizations are working with, depending upon where they are in their AI journey."

Just 7% of the study's respondents indicate that their organization's data is completely ready for AI adoption; another 15% say it's almost completely ready; the bulk of respondents, 51%, say their data is somewhat ready; and 27% say it's either not very or not at all ready. FIGURE 1

Despite growing awareness of the challenges and potential consequences of poor AI data, there's much to fix. "Data is still messy, still siloed, still not governed properly, but there's so much data sitting inside enterprises that's not being used," says Sesh Iyer, managing director and senior partner at Boston Consulting Group (BCG), a Boston-based management consultancy. He adds that "what is slowing [enterprises] today is the readiness, accessibility, and auditability of the data, as well as the data governance."

Many organizations remain stymied by their inability to build, scale, or automate their "data pipeline" to efficiently integrate, filter, and normalize internal and third-party data for analysis by AI or analytics tools. Skills gaps also compound the challenges in an organization's AI data transformation efforts. Yet growing use of data pipeline automation tools, such as agentic AI, promises to help organizations better manage the spiraling costs and daunting complexities of harvesting timely, accurate AI data insights. "As automation tools [begin] to reduce human error," says Tony Palmer, principal analyst and practice director at Omdia, a market research and advisory group in Newton, Mass., organizations will "close those gaps, and they're going to be able to minimize mistakes, and they're going to get more value faster."

This report aims to understand how organizations can unlock AI's full potential by applying cloud and agentic AI technologies to accelerate, scale, and improve the data quality that's fundamental to creating transformative business insights. The report will also explore how organizations overcome obstacles to data readiness, such as organizational or regulatory constraints, by leveraging containers and data catalogs, as well as new techniques, such as virtual private clouds that bring AI processing capabilities directly to their data.

State of Data Unreadiness

Think of a data pipeline as a staging area that's often a landing zone for raw data, whether on-premises or in the cloud. Once AI data is ingested, data teams perform quality checks and tap orchestration and workflow management tools to manage

> **“Data is still messy, still siloed, still not governed properly, but there’s so much data sitting inside enterprises that’s not being used.”**
>
> Sesh Iyer, managing director and senior partner, Boston Consulting Group

and secure data. Normalizing data is often convoluted—far from a plug-and-play process.

Many companies tap public cloud services to consolidate, store, and refine their business data, knowing they lack a full stack of in-house technical experts that include data scientists, developers, storage and security specialists, along with cloud architects who can match infrastructure to workload needs.

Yet an inability to generate credible insights trusted by management, customers, and employees inhibits organizations from scaling their AI initiatives beyond the proof-of-concept stage. To prevent costly mistakes and foster trust in AI outputs, many organizations embark on a multifaceted and expensive journey that entails strategic planning and expanding investments in business processes, data governance, data talent, services, and infrastructure. The AI fervor hasn’t crested. There are few, if any, indications that AI spending has become as commoditized as, say, e-commerce. In August, UBS Investment Bank forecasted that global AI spending would reach $500 billion in 2026, up 33% from 2025.[1]

Until this decade, which has witnessed explosive demand for generative AI tools and cloud-based services, few organizations sought to pull insights from the murky depths of typically siloed data sources such as customer interactions, financial transactions, internet of things (IoT) sensors, video streams, and social media posts. Few would tackle the integration headaches of uniting and analyzing structured and unstructured data without believing that these investments would yield insights that make a difference.

For instance, companies look to gain deeper customer insights, a so-called 360-degree customer view, by analyzing transactional and behavioral data from multiple touchpoints. They can obtain market insights by dissecting video streams and social posts. They can also tap into IoT sensor data for predictive maintenance or operational insights, such as energy consumption.

All of that’s possible with viable data. But that data is often not available. “There’s additional work before AI comes into the picture,” explains Deloitte’s Ammanath. AI insights may generate myriad opportunities, but mining them may reveal foundational concerns. “Companies have to deal with additional complexities that have been introduced by AI, or really just bubbled up with AI, in addition to the legacy data quality issues that have always existed.”

According to the Harvard Business Review Analytic Services study, respondents say the biggest challenges of preparing data for AI use concern siloed data and the difficulty of integrating data sources (56%). The next tier of challenges includes the lack of a clear data strategy (44%) and data quality/bias issues (41%). FIGURE 2 Amid these concerns, 73% of respondents agree somewhat or strongly that their organization has found the processing and preparing of data for AI to be challenging.

Chief among the many worries plaguing organizations dealing with inadequately prepared AI data, 52% of respondents choose inaccurate/biased AI results as one of their top three concerns. Loss of security or intellectual property at 40% and unanticipated operational costs at 30% were the other most common answers.

Technology constraints aside, some organizations are still gathering the necessary elements to gain deeper insights into their business operations. “If you are a legacy manufacturing company, some of your engineering drawings are likely on paper, right?” says Ammanath. “They may not even be digitized.” She believes AI data teams must ask the questions, “Do we have the right data, and are we providing all possible data components to the algorithm?”

Even after organizations have digitized and cleansed their data of potential inaccuracies, they must determine whether it is theirs to use. If organizations cannot prove the origin of their data, they become vulnerable to potential intellectual property violations. “So now, if I am using data, where did the data come from?” asks BCG’s Iyer. “Was that the correct data, especially in regulated industries, to ensure non-repudiation? I have to be able to audit and show that it came from the right place.”

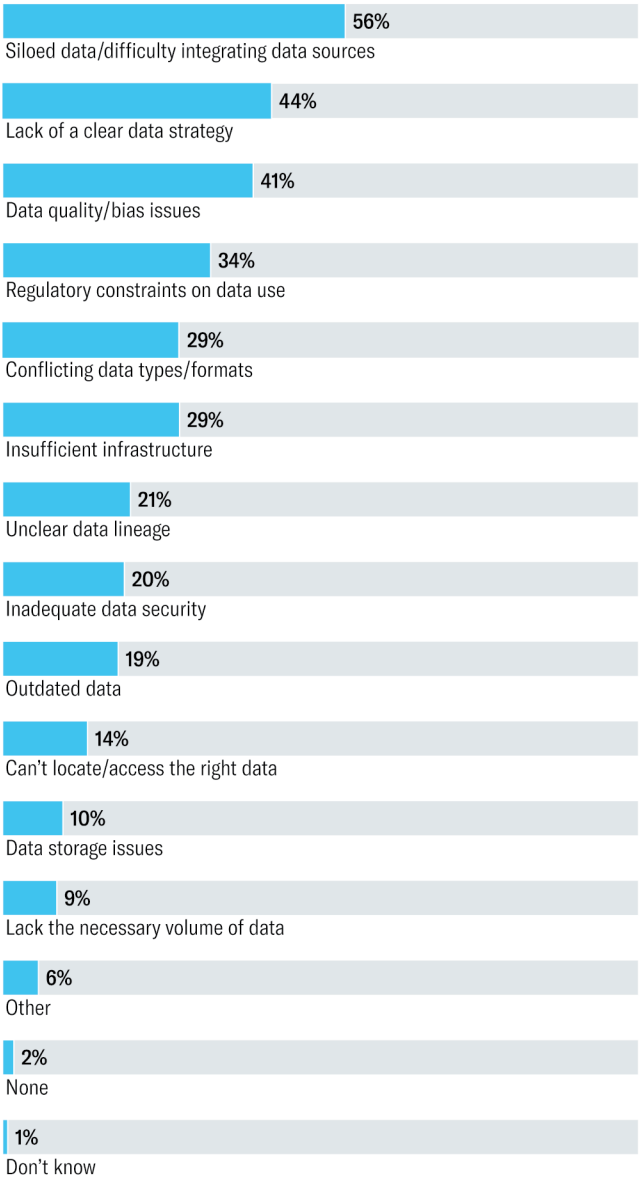
Data trust goes beyond knowing where the data comes from and can also encompass practical considerations, such

FIGURE 2

Challenges Impeding AI Data Preparation

Data silos and lack of clear data strategy are top obstacles

What data challenges, if any, are making it difficult for your organization to prepare its data for AI use? *Select all that apply.*



Base: 231 respondents

Source: Harvard Business Review Analytic Services survey, October 2025

as how to use it. Inventory is a case in point. "It's not enough to know that [an item] price is $10," says Accenture's Tung. "I'd better make sure I actually know what it means. Is this in U.S. or Australian dollars? Is it a unit or bulk price, and when is the price valid? That's the sort of experience that we need for AI to use data at scale."

The Cloud as Data Factory

When faced with the enormous and complex problem of AI data quality, organizations must overcome multiple obstacles across infrastructure, strategy, and talent. They may stumble trying to address everything at once. Lacking a viable plan, organizations are unlikely to successfully harvest large volumes of fresh, accurate, and relevant AI data.

Yet for most organizations, AI data strategies remain a work in progress. Data strategy adoption depicts a classic bell curve. About one in four respondents' organizations (23%) have created a data strategy for AI adoption. However, in a sign of growing awareness of its importance, 53% of respondents say their organization is developing one. The remaining quarter of respondents say their organization either hasn't started one (22%) or they don't know (3%). FIGURE 3

Producing valid data isn't just a matter of following a five-star recipe, because the inputs can be fluid at times. What steps should organizations prioritize to counter their wide-ranging AI data management issues, such as eliminating data silos and fixing fragmented or poorly vetted data? "I don't think there's a standard way," notes Omdia's Palmer. "I think that's part of the problem. There are integrated platforms and tools that combine that kind of lineage tracking with metadata management capability, and really what we're recommending is that organizations should invest in tools that can provide end-to-end data governance."

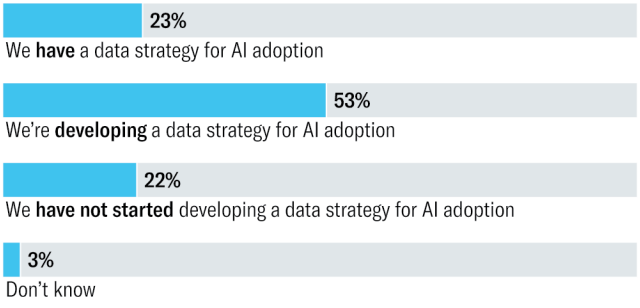
Unsurprisingly, data protection is the most common element of AI data strategies. That's not just a matter of good cloud hygiene, like backing up data; it's also about keeping

FIGURE 3

AI Data Strategies Are a Work in Progress

More than half say their organization is developing an AI data strategy

Does your organization have a data strategy for AI adoption?



Base: 231 respondents

Source: Harvard Business Review Analytic Services survey, October 2025

sensitive data out of competitors' or cybercriminals' hands. Accordingly, respondents at organizations that have or are developing an AI data strategy say that, overall, the three most critical elements of these plans are security, including protecting sensitive data and privacy (59%); developing data quality that's clean, consistent, and usable (46%); and applying data governance (41%). Behind these, the next-most-critical AI data strategy elements include alignment with the overall business strategy (36%), data integration and accessibility (29%), and ethical AI and bias management (28%). By any measure, that's a daunting checklist of addressable issues.

Palmer believes that if organizations don't prioritize "ensuring quality and data security in every decision that they make, their [AI] models are going to underperform and their initiatives are going to generate inaccurate results. They're not going to be able to satisfy compliance requirements."

The study indicates that, for many, AI data preparation occurs in the cloud. While 51% of respondents say the cloud, including public, private, or multicloud configurations, is their organization's primary data storage choice for the majority of its AI data, another 28% say they spread their AI data across an equal mix of the cloud and on-premises infrastructure. Only 11% perform these tasks exclusively on-premises. The cloud may play an even larger role in the future, as 77% of respondents indicate that their organization is increasing cloud storage for AI data over the next 12 months. In comparison, only 22% say they're boosting on-premises AI data storage.

Apart from the cloud's practical and technical advantages for AI data handling, such as automatic scaling, managed caching services, and data life cycle automation, Tung also sees other strategic considerations. Organizations should strive to productize their data, treating the cloud as a data factory. "Companies need to think about their data—where it's coming from and that it's ethically sourced and of good quality," she says. "Building efficient cloud capability is like creating a factory to take in data, transform it, and apply AI." She extends the analogy, adding that "the data products I create and the efficiency of my data supply chain become part of my differentiation."

Preparing data for AI processing also includes ensuring it can be interpreted accurately. "When you're looking to use data in the context of building intelligence, you're trying to drive shared meaning across the enterprise," says Iyer. "The intent is to ensure that there's no ambiguity." He adds that with "consistent data definitions," your AI model works when you query your data and get "a well-defined, standardized answer," no matter where the query or the data comes from.

Ultimately, good governance demands that those who know it best own it, according to Tung. "The data needs to be owned by the business," she says, "so you're rightsizing the investment, and you're ensuring that the data is described properly, both in terms of how you can use it and what you can use it for. The business is the one that's ultimately going to be the best authority to capture the value."

Changing the Laws of Data Gravity

Training AI models on massive data sets introduces new data challenges. Copying data between servers eats up time and computing power. This challenge, known as data gravity, explains why vast—and often still expanding—data

> **When you're looking to use data in the context of building intelligence, you're trying to drive shared meaning across the enterprise. The intent is to ensure that there's no ambiguity.**
>
> Iyer at Boston Consulting Group

repositories tend to attract critical applications and services and why, rather than moving them from one data center to another, they are often supported in place.

Now, data strategy is evolving as technology enables organizations to run AI inference or model training without moving the data from one location to another. While data gravity has long constrained organizations from repurposing their data for discrete tasks, recent advances in distributed AI frameworks now enable processing where data resides. Those advances trim processing costs, but more significantly, they reduce the latency, bandwidth costs, and security risks associated with moving big data sets to centralized locations.

Moving processing to the data is essential for enormous data sets, Iyer says. "We are starting to see the emergence of multi-petabyte workloads, and in those cases, you have to find a way to bring the algorithm to the data, because it's physically impossible to move the data because of its size and scale."

While the cloud remains a great place to build data products, much of the data a company needs to use isn't going to be colocated, says Tung. Increasingly, organizations employ a "zero copy data access" architecture enabled by containerized applications running across different computing environments, such as the cloud. This enables organizations to optimize where and how they process AI data without duplicating data. This system allows organizations to apply their data to new AI use cases, adds Tung. "For reasons such as cost, sovereignty, and regulations, I think it'll be more likely to move a container of the AI processing to the data as opposed to moving the data to the app."

The only reason companies need to make data set copies now is for "performance or regulatory reasons," Tung adds. Increasingly, organizations must keep data where it resides to comply with sovereignty laws, most notably the European Union's General Data Protection Regulation, which restricts the cross-border movement of personally identifiable information. "We're not going to be able to bring [data together] physically in the same place," she adds. "For most cases, [data will] remain in a multi-partner and multisystem environment. Instead, we will find and access data where it sits."

Artificial Intelligence to AI's Rescue

Autonomous actions have emerged as one of the most highly anticipated AI-powered capabilities. Agentic AI systems, which consist of code, AI models, and connectors that act on other systems, don't just recognize problems—they initiate actions to solve them. While there are many promising ways to use autonomous assistants, such as for making dinner or travel reservations, it's particularly appealing for manual-intensive processes.

Take, for example, data management, where data flows into an organization from multiple, often disparate, sources and must be scrubbed, often at great expense, before it is deemed trustworthy for insights and analysis. There is a surprising number of steps and variables at play in this process. It's not just about identifying data problems; the big win is solving them without operator intervention.

The technology is winning converts who perceive agentic AI as a solution to their AI data quality issues. About half of the survey respondents (47%) agree that their organization believes that agentic AI can solve its data quality issues. More broadly, nearly two-thirds (65%) expect many of their organization's business processes will be augmented or replaced by agentic AI in the next two years. FIGURE 4

Iyer says that agentic AI "really changes the way enterprises can unlock the value that is resident in enterprise data. You're reducing manual grunt work and doing things like mapping, documentation, quality checks, pipelines, triage—all things that are difficult to do manually."

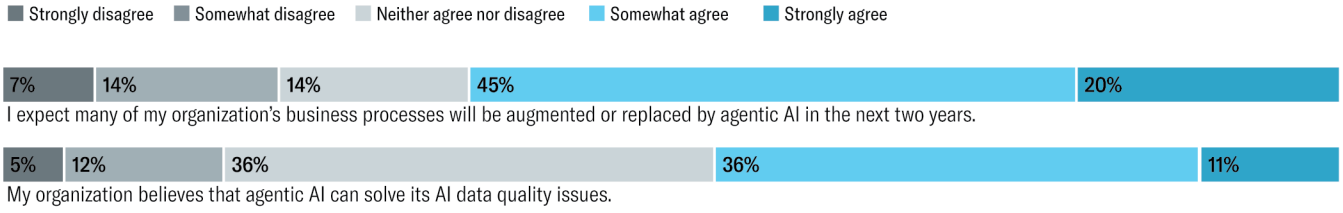
He says that his organization is building ontologies—shared vocabularies for data and how it connects—orchestrated by agentic AI, which "ensures that you're able to have the rules and assertions that actually make data semantically relevant. AI agents give you the ability to do this at

FIGURE 4

Searching for Data Quality Improvement

Many agree that agentic AI can solve such issues and will impact business processes

Rate the extent to which you agree or disagree with the following statements.



Base: 229 for first statement and 204 for second statement; both exclude "Don't know."

Source: Harvard Business Review Analytic Services survey, October 2025

scale—much faster—and get to a desired level of quality. And then, if I also need to maintain the asset, these agents can monitor data health, data drift, and ensure that things [happen] in a very responsible way, an operationally correct way."

Tung believes agentic AI benefits from widely available open-source and vendor tools and that organizations that have "invested in things like data governance, data standards, and metadata requirements" understand why these building blocks are "needed for getting these agents to do something on your behalf." For example, she says they can "learn from your data what they need to do—figure out patterns and certain things that don't work or things that do."

She believes agentic AI is also adept at preparing your data, including generating ad hoc insights. "So rather than pre-building reports and applications, if I'm able to use agents dynamically," she says, "[I] don't have to ever build a system to just present that view. I might use agentic AI to generate the code pipelines and the data transformations to access that data, and then I'll use it to generate the tests and the test data. You still have the human loop, but the agents are doing all the mundane tasks."

Iyer cautions that a successful agentic AI implementation requires a "good team with a diversity of skills to do this well," including "deep data engineers." There's a need for talent who can write efficient data prompts, domain subject matter experts, and, he adds, "a product owner who can actually bring all these people together to ensure that they stay committed to the objective function that has to get delivered. Without the right skill sets, you are not able to get to the right output, and then you have to go reconfigure, so it becomes a painfully iterative play. Success versus failure is how you compose these teams and put them against a mission."

Conclusion

Poor data quality plagued enterprises long before the advent of the cloud or AI. Organizational intolerance for data problems is a sign of AI's expanding presence. Data governance is a higher priority amid the escalating costs of mining AI insights and the need to reduce the risk of data errors that can undermine an organization's credibility with customers, shareholders, business partners, and even employees.

Until recently, few organizations invested significant resources in unifying disparate data types or combining siloed databases to mine them for previously unavailable market, operational, or customer insights. As the study indicates, most organizations affirm the value of AI data preparation yet lack an actionable AI data strategy. Many organizations express concern about the difficulty of fixing data problems and grapple with multiple challenges in constructing and managing a data pipeline to refine their data in a robust, timely manner.

To satisfy these objectives, many respondents say their organization will improve AI data preparation through a mix of process and governance changes along with technology investments. When asked which data solutions their organization is focusing on to get their data more prepared for AI adoption over the next 12 months, half of respondents say their organization plans on integrating data sources/

> **“We’re finally changing the top-down view that data is a cost center. Today the top leaders want to see AI across their business, and that means solving for their data.”**
>
> Teresa Tung, global data practice lead, Accenture

breaking down silos, while nearly as many (46%) will be enhancing data governance to ensure that.

There is widespread belief that agentic AI tools can improve data quality and automate processes to mitigate human error and enhance AI outcomes. Agentic AI will be transformative if it autonomously finds and fixes data errors before humans notice—building trust in the data used to make business decisions. Ammanath adds, “I can definitely see agentic AI aiding data governance, whether it is tracking the validity of the data, tracking the lineage, [or] maintaining the currency of the data.”

While agentic AI tools may fill organizational gaps in data operations skills, they could also make a much broader organizational impact. A slight majority of the study’s respondents somewhat or strongly agree that their organization’s business processes will be augmented or replaced by agentic AI tools in the next two years.

The quest for high-quality AI data is changing organizational priorities. “We’re finally changing the top-down view that data is a cost center,” says Tung. “Today the top leaders want to see AI across their business, and that means solving for their data. It has never been as much of a big priority until now.”

Endnotes

1 UBS Investment Bank, “CIO Expects Global AI Spending to Hit USD 375bn This Year,” August 2025. https://www.ubs.com/us/en/wealth-management/insights/market-news/article.2515967.html.

METHODOLOGY AND PARTICIPANT PROFILE

Harvard Business Review Analytic Services surveyed 231 members of the *Harvard Business Review* audience via an online survey fielded in October 2025. Respondents qualified to complete the survey if they were involved in their organization's data decisions (including around how data is used/not used for AI) and their organization was actively practicing, piloting/testing, or exploring/considering AI use for business purposes.

ORGANIZATION SIZE

- 29% — 10,000 or more employees
- 32% — 1,000–9,999 employees
- 10% — 500–999 employees
- 16% — 100–499 employees
- 13% — 50–99 employees

SENIORITY

- 27% — Executive management/board members
- 42% — Senior management
- 21% — Middle management
- 10% — Other grades

INDUSTRIES

- 13% — Technology
- 12% — Financial services
- 11% — Education
- 11% — Health care
- 10% — Government/not-for-profit
- 10% — Manufacturing

All other sectors less than 10% each.

JOB FUNCTIONS

- 19% — General/executive management
- 10% — Administration
- 10% — IT
- 9% — Consulting
- 8% — R&D/innovation/product development

All other functions less than 7% each.

REGIONS

- 48% — North America
- 23% — Europe
- 19% — Asia Pacific
- 7% — Latin America
- 4% — Middle East/Africa

HARVARD BUSINESS REVIEW

VISIT US ONLINE

hbr.org/hbr-analytic-services

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject-matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticservices@hbr.org.

CRE74320326