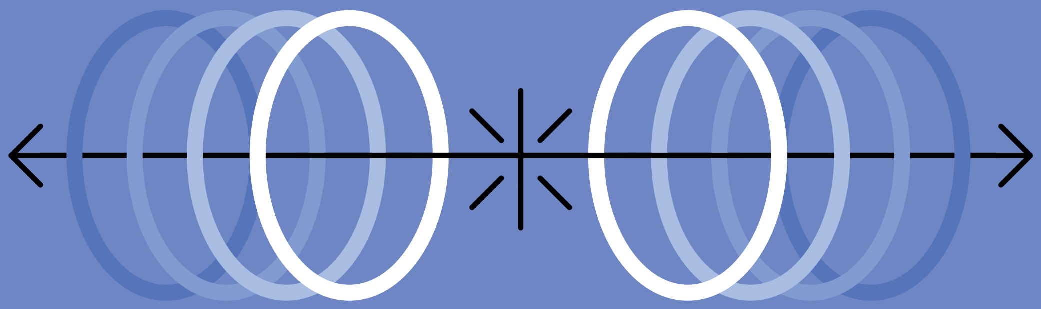
HARVARD
BUSINESS
REVIEW



Accelerating Data-Driven Transformation in the Hybrid Cloud

HBR
Analytic
Services
**WHITE
PAPER**

Sponsored by

CLOUDERA

SPONSOR PERSPECTIVE

Trusted Data Today for Tomorrow’s AI

The artificial intelligence (AI) revolution is upon us, and enterprises everywhere are looking to integrate AI into their mission-critical processes so they don’t get left behind in their respective fields. Data-driven enterprises are at the forefront of innovation, making real-time decisions based on high-quality data.

High-quality data. It’s the foundation of trustworthy AI, and it’s what ensures that organizations are getting value out of their data-driven decision making. At the same time, concerns around the quality and accuracy of data remain a top factor preventing enterprises from rolling out generative AI.[1]

So how can you ensure that you’re working with high-quality data? Simply stated, you need to have full control over and visibility into all your data, over the entire data life cycle and across distributed environments. Having mastery over your data estate is no small feat, and for virtually all data-driven organizations, the right combination of people, data, and technology makes all the difference.

At Cloudera, we’ve seen firsthand how trusted AI based on high-quality data can revolutionize industries, facilitating innovations including autonomous vehicles, prescriptive medical treatments, and seamless fraud prevention.

The next wave of innovation will come from agentic AI, and we’re seeing a massive increase in organizational use of AI agents.[2] Importantly, to maximize ROI, agentic architectures need a standard integration layer like Model Context Protocol (MCP). Servers built on MCP provide a universal gateway to govern enterprise data, closing the context gap and ensuring data quality.

This research by Harvard Business Review Analytic Services, in association with Cloudera, examines organizational maturity regarding AI initiatives and outlines several best practices that business leaders should follow to advance their journey to becoming a data-driven enterprise.

The report underscores the absolute necessity of having high-quality data and robust data governance as the foundation of your data management strategy. Additionally, it shows that organizations are increasingly adopting a hybrid cloud approach to capitalize on the best of both worlds, namely the operational and security advantages of on-premises servers and the cost and scalability benefits of the cloud. It also importantly outlines the architectural components that many data-driven organizations share that enable them to make real-time decisions based on trusted, high-quality data: open data lakehouses, frictionless data flows, and a unified data fabric.

Abhas Ricky
Chief Strategy Officer,
Cloudera

We sponsored this report because we’re passionate about helping customers realize their goals of becoming data-driven—making real-time decisions based on trusted, high-quality data and delivering competitive advantage and differentiation for their business success.

Now is the time to recognize the role that high-quality data plays in informing enterprise AI use cases, and ensuring that you have a modern data architecture—one that’s unified and interoperable, supports open discovery, and provides robust data governance—in place to deliver actionable, trustworthy insights. With the right data, analytics, and AI partner, the sky’s the limit as to on what your organization can accomplish!

[1] Bain & Co., “AI Readiness Survey,” July 2024 and December 2024.

[2] Cloudera, “Survey Report: The Future of Enterprise AI Agents,” April 2025.

Accelerating Data-Driven Transformation in the Hybrid Cloud

Even in the hypervelocity of the digital age, caution often dictates enterprise decision making. Organizations will intelligently test the waters and evaluate technology before investing millions of dollars in talent and infrastructure to become more data driven. But the test won't focus solely on whether data can be well managed in the cloud—these days, that's a given—although it's not without complexity. Less obvious is whether the organization stands ready to integrate data from disparate sources and exploit a bounty of insights.

TODAY, the accepted wisdom in company boardrooms is that deeper, faster insights boost profitability by surfacing ways to achieve greater operational efficiency with incremental or strategic product improvements. Nearly four in five (79%) companies acknowledge that they must integrate artificial intelligence (AI) in mission-critical processes to become more competitive, according to November 2024 data provided by Enterprise Strategy Group, a Newton, Mass.-based tech market research firm.[1]

A company's AI data platform decisions will determine how efficiently it can achieve scale and solve complex AI deployment challenges. According to a September 2024 study by 451 Research, a New York-based technology research firm, titled "AI & Machine Learning Infrastructure," 83% of organizations project their AI workflows will increase in the next two years, while two-thirds anticipate a need for infrastructure upgrades to meet these future demands. Whether organizations opt to run AI workloads in their data center or the cloud, many experience data preparation challenges, according to a January 2024 Harvard Business Review Analytic Services survey.[2] Half of the AI decision makers who responded said they have "difficulty integrating diverse data sources into a unified format." Another 44% complained of "poor data quality—their data isn't reliable or standardized."

Establishing trustworthy AI data quality also requires organizations to know the provenance of the data used for analytics or to contextualize large

HIGHLIGHTS

Companies today face crucial and complicated decisions regarding aggregating data for artificial intelligence (AI)-based data analytics.

Organizations are increasingly adopting the hybrid cloud model to create a secure and scalable data pipeline that generates consistently trustworthy AI data insights in a regulatory-safe manner.

Proper data governance throughout the AI data life cycle enables organizations to maximize the value of their AI investments, manage risk, and improve their business decisions.

> “How do you embed data into every decision, every interaction, and every process within the organization? Getting value out of your data-driven organization is far beyond data and technology.”
>
> Sanjna Parasrampuria, a partner in the Singapore office of McKinsey & Co.

language models (LLMs), chatbots, and generative AI (gen AI). It’s all hands on deck. Everyone who touches the data—not just the data scientists, engineers, cloud architects, and storage and operations teams—must be vigilant about data quality. “Being data driven is going to require today’s and tomorrow’s workers to have a greater understanding of the source and path of the data,” says Stuart Sim, a partner at Bain & Co., a Boston-based consultancy. He asserts that workers “put too much faith in the face value of the data or information supplied to us through end reporting systems.”

Given these varied trust issues, many companies prioritize solving their data production challenges—the thorniest part of becoming data driven—by establishing scalable, secure data pipelines on premises, in the cloud, or as a hybrid. However, data consumption is sometimes an afterthought; companies mistakenly assume their employees fully trust the data and are ready to exploit the new insights. It’s yet another facet of becoming data driven that requires planning and resources.

“You can build a fantastic analytics or decision support system, but there is no guarantee that people will use it,” cautions Sudipta Ghosh, a partner in the Mumbai office of PwC India, a management consultancy. “Why? Because they either don’t trust that application or are fearful that they might lose their jobs, or they might feel that they know better than what the application tells them because they have 35 or 40 years of experience in that sector. So there could be a healthy skepticism, denial, or fear. There could be all sorts of emotions, which should not be brushed under the carpet or ignored completely because that will impact how people use it.”

This Harvard Business Review Analytic Services report will seek to understand and examine how companies can overcome these challenges to accelerate and achieve their AI-driven data transformation objectives. By unifying their data in a lakehouse, a storage and processing architecture that supports structured and unstructured data in a hybrid cloud data operations model, organizations can get the right information to the right people at the right time with powerful security and control.

“How do you embed data into every decision, every interaction, and every process within the organization?” asks Sanjna Parasrampuria, a partner in the Singapore office of McKinsey & Co., a strategy and management consulting firm based in New York. “Getting value out of your data-driven organization is far beyond data and technology. It requires rewiring your organization, which starts with having a business-led roadmap of where value sits within all the company functions,” and, she adds, identifying the “biggest opportunities.”

Data Governance Groundwork

Enterprises widely deployed AI-based applications long before the November 2022 introduction of ChatGPT, the first popular gen AI “chatbot,” a technology many believe enables faster and better business decisions. Yet those AI applications employing predictive analytics, pattern detection, machine learning, and natural language processing—invaluable in many ways, especially to science—never ignited the imaginations of businesses, investors, and consumers in quite the same way. Until then, AI was widely experienced in algorithms, often for browsing, maps, or streaming, but it wasn’t easy for mainstream computer users to access or manipulate.

In its “The Data-Driven Enterprise of 2025” report, McKinsey predicted in January 2022 that amid “rapidly accelerating technology advances,” the characteristics of this new organization would include “data embedded in every decision, interaction, and process.”[3] Many companies unhesitatingly sought to deploy gen AI, contextualize LLMs, and deploy other forms of AI, such as predictive analytics, in a quest to modernize their business intelligence and analytics programs, mine company data for customer insights, innovate product offerings, and improve their operational efficiency.

The cloud was the prominent platform for this new investment because of its lower operational costs of large-scale capabilities and “massive benefits” for data team productivity by focusing on delivering use cases instead of managing overly complex data systems, notes Parasrampuria. In its report, McKinsey surmised that companies need a “cloud-enabled data platform to meet future data and analytical needs” and get AI to scale.

While harnessing the talent, tech infrastructure, and code to become data driven presented formidable challenges, data became the most common stumbling block. Companies today face crucial and complicated decisions regarding aggregating data for AI-based data analytics. Company data is often dispersed in multiple public or private clouds or software-as-a-service applications, or it is siloed for privacy purposes in various regional, on-premises data centers. As a result, many companies lack a comprehensive view of their data assets, stymying their analytics efforts.

“One of the most significant technical constraints when establishing a data-driven enterprise is the lack of comprehensive, integrated infrastructure,” says Chris Thomas, principal and U.S. cloud leader in the Cleveland office of Deloitte Consulting LLP. “While tools like AI, machine learning, and cloud platforms are often seen as the solution, companies sometimes focus too narrowly on one area without considering the broader system.”

Though many organizations have utilized AI for years, there’s no doubt that generative AI in particular has lit a fuse. A June-July 2024 Deloitte study called “Focusing on the Foundation,” tracking the investments of business and technology leaders, found that 40% of them are “investing in the foundations for a robust data estate—data architecture, data management, and data insights—57% are investing in cloud platforms and 60% in generative AI,” adds Thomas.[4] “The question remains whether these priorities are in the right order, as a generative AI solution is only as good as the data one curates.”

Though many organizations have utilized AI for years, there’s no doubt that generative AI (gen AI) in particular has lit a fuse.

Getting Data Right

Businesses have long understood and respected the importance of high-quality data. Bad data impedes understanding, and trusted data has always been the core of successful business intelligence and analytics programs. Still, the probabilistic and interpretative nature of AI-driven business insights necessitates specialized approaches to data governance and cleaning, posing significant challenges for AI adopters. Flawed and biased data undermines AI’s interpretive capabilities, causing inaccuracies such as false correlations, ethical risks, and amplification of existing biases.

The possibility of AI data hallucinations—often wildly incorrect or inappropriate outputs—keeps risk, compliance, and data privacy teams up at night. Consequently, many organizations have embraced new frameworks and tools for evaluating data reliability, understanding confidence metrics, and validating AI-generated insights for decision making.

As companies look to aggregate multiple data sets with structured data such as customer records and financial transactions and combine them with even more unstructured data inputs such as call center transcripts, social media interactions, or streaming data from devices (via the internet of things) and other equipment, they require amassing significant compute and data storage resources along with data operations expertise to “normalize” these heterogeneous data formats, often containing both structured records and unstructured data, which could include video, audio, and driving routes, among myriad other possibilities. The result is usually a massive dataset that is complicated and expensive to maintain, model, and mine for fresh data insights.

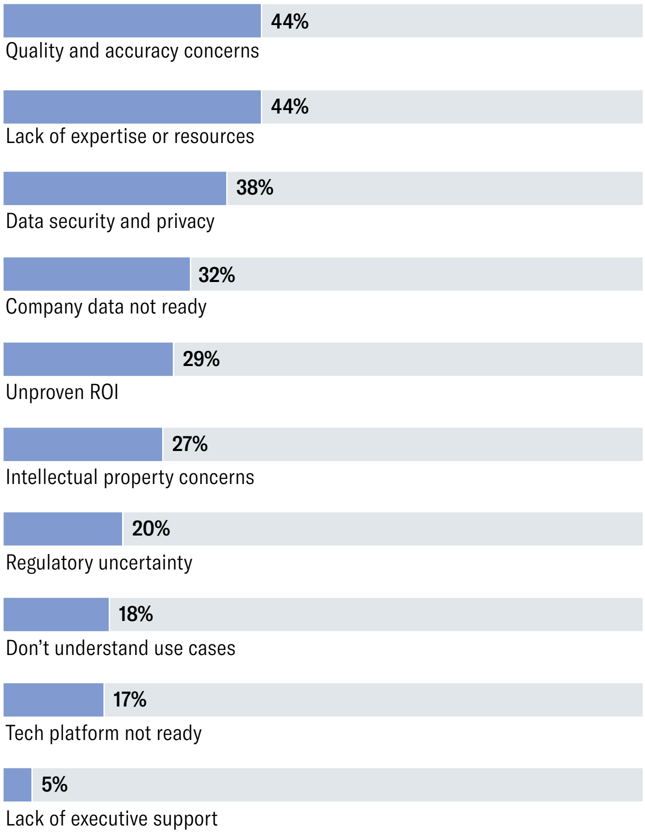
PwC India’s Ghosh asserts that companies need “a reality check” about their data. “If there’s a problem with the data, there will be a problem with the analysis,” he explains.

FIGURE 1

Generative AI Impedance

Organizations assess the factors that keep them from rolling out generative AI

What's preventing your company from moving faster?



Source: Bain & Co. survey, July 2024

"Most of the problems stem from the fact that the data is not complete, not consistent, not accurate, not uniformly maintained, or sitting in silos." Ghosh believes that the combination of data aggregation, advanced data management, and data governance techniques can help solve many issues. "A good amount of focus must be there to understand, identify, manage, collate, and store data more consistently." Yet mastering the often arduous process can lead to "a single version of truth, or at least the most consistent version of truth in one place," he says.

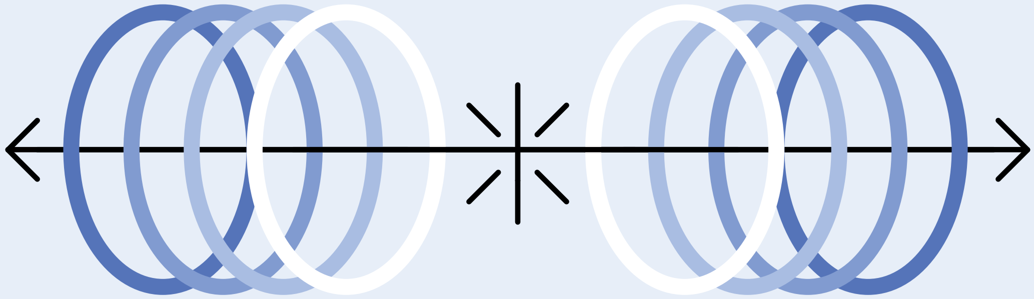
Another problem impeding data-driven initiatives concerns modernizing data centers with faster computing and storage for demanding AI workloads. "I think the biggest challenge most large companies face is they're working on legacy systems and siloed data—that's the biggest roadblock," explains McKinsey's Parasrampuria. "This has accentuated the problem of people in the same organization not necessarily trusting their data."

According to the July 2024 Bain & Co. "AI Readiness Survey" that polled executives on what's preventing their organizations from moving faster into gen AI, 32% conceded that their company data is not ready and 44% said they had quality and accuracy concerns. FIGURE 1 In the past three times Bain fielded that question in its quarterly study, respondents' attitudes about their company's data readiness have only worsened.

Apart from the sheer mass of AI data sets spanning multiple terabytes or petabytes of information, managing data in the AI age is far more complicated than traditional business intelligence reporting, Bain's Sim contends. That's due partly to the artificial nature of "derived data sets and synthetic data," he says, referring to the complicated process of validating insights about patterns and trends. "You're creating multiple copies of different versions of data using a scientific method and you're not entirely sure where your source of truth either starts or ends. For that reason, you need very sophisticated configuration management and change management around your data sets."

Many companies further complicate lineage questions by acquiring third-party data, such as public records or weather data, to enrich the value of their AI data sets. Solving the data trust issue requires chief data officers (CDOs) to "track the entire lineage of data from the source" to where it's "applied in decision making," says Sim. "Knowing the complete data life cycle enables CDOs to determine the source of truth."

The stakes are rising for companies working with gen AI and other forms of AI. A July-September 2024 Deloitte study called "The State of Generative AI in the Enterprise" highlighted growing business awareness of the factors that may impede gen AI adoption, most notably mistakes and errors with real-world consequences, cited by 35% of respondents, a shortage of high-quality data (30%), and general loss of



"You're creating multiple copies of different versions of data using a scientific method and you're not entirely sure where your source of truth either starts or ends."

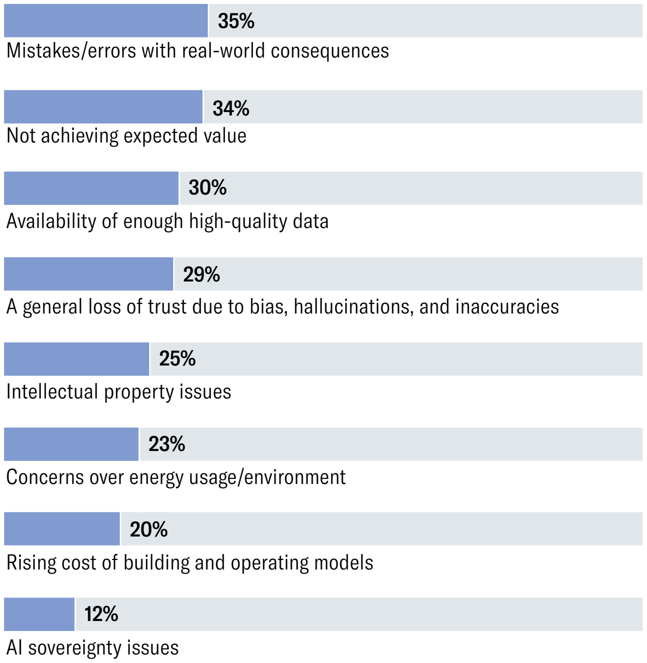
Stuart Sim, a partner at Bain & Co.

FIGURE 2

Slowing Generative AI Adoption

Organizations assess the factors that could slow the overall marketplace adoption of generative AI

Which factors could most slow market adoption of generative AI in the next two years?



Source: Deloitte survey, July-September 2024

trust due to bias, hallucinations, and inaccuracies (29%).[5] FIGURE 2 The authors noted, "For broader gen AI adoption to occur, the technology's reliability, accuracy, and trustworthiness will need to improve."

The real-world consequences cited by Deloitte survey respondents could include egregious privacy leaks. Such incidents run afoul of global and domestic privacy regulations such as the General Data Protection Regulation (GDPR), a European Union data privacy regulation, and the California Consumer Privacy Act (CCPA), which grants rights to state residents over how businesses use their personal information. GDPR fines can hit €20 million (approximately $21.6 million) or 4% of annual sales, whichever is higher. CCPA penalties are more modest, at $2,500 to $7,500 per violation, depending on whether the mistake was intentional, but there is no cap on total penalties.

Although the well-known federal Health Insurance Portability and Accountability Act of 1996 was installed long before gen AI, maintaining compliance with it remains a pressing concern, particularly for the health care industry. Like customer service functions in various industries, health care providers deploy chatbots to streamline responses to patient queries, including appointment requests and routine questions about procedures and medications. They can also facilitate patient data collection, raising the stakes for data security and privacy procedures.

For these reasons, among many others, data compliance teams run a battery of AI safety checks, including adversarial testing on data sets during all development phases, to evaluate the model's efficacy and gauge how it handles confidential data. The checks help ensure that personal health care information or, in the case of financial information, personally identifiable information is anonymized or scrubbed. The checks also prevent the disclosure of company secrets, including intellectual property, to avoid reputational damage and a loss of investor confidence. Biased data that violates antidiscrimination statutes in the U.S. and Europe, including the new EU AI Act, can result in heavy fines.

"In addition, companies are recognizing the need for transparency in their AI models," says Deloitte's Thomas. Government auditors, employees, and customers increasingly seek explanations about how AI models work. "Clear communication about how data is being used, how models make decisions, and the measures taken to ensure fairness and accuracy can significantly increase stakeholder trust. Moreover, cross-department collaboration is critical to aligning AI systems with broader business objectives, ensuring that data is used consistently and responsibly across the organization."

What Hybrid Delivers

Any discussion of the pros and cons of running data workloads in the cloud versus on premises typically suggests that they're better together—each offering the operational or cost advantages lacking in the other. That recognition fuels robust demand for hybrid clouds, which thrive by enabling companies to port workloads and data between on-premises data centers and public or private clouds. Companies maintain some workloads on premises for compliance, security, or performance reasons and tap public cloud services for cost, scalability, risk or compliance management, and business continuity. Taking a hybrid approach to running AI workloads on premises and performing compliance management in the cloud offers companies the scalability and advanced tools of the cloud for managing compliance while maintaining security and control over sensitive data on premises. Similarly, hybrid cloud models help organizations satisfy a bevy of global regulatory restrictions—especially data sovereignty mandates—and help address trust concerns by allowing firms to manage sensitive data in a regulatory-compliant and secure manner.

Nowhere is this flexibility more evident than in how organizations use hybrid environments to facilitate transparency and explainability in their AI decision making. Many companies store sensitive data on premises and tap cloud servers for processing, resulting in traceable AI data flows. AI models can run in the cloud before data teams move them into production on premises. Hybrid environments also excel at federated model training, a process that allows geographically dispersed teams to collaboratively train a single global model using their local data—without needing to aggregate sensitive information in one location.

“A hybrid cloud-based data platform enables organizations to achieve scale, security, and governance while undergoing a data-driven transformation,” says Thomas. “Scalability is a key advantage, as workloads can dynamically expand while keeping latency-sensitive operations on premises. Security and compliance can also benefit from this model, allowing organizations to store sensitive or regulated data in private environments while utilizing public cloud infrastructure for high-performance analytics.”

Thomas also promotes hybrid to “enhance business agility, allowing companies to move workloads between environments without vendor lock-in. This flexibility supports collaboration across dispersed teams while maintaining centralized governance.”

> “A hybrid cloud-based data platform enables organizations to achieve scale, security, and governance while undergoing a data-driven transformation,” says Chris Thomas, principal and U.S. cloud leader at Deloitte Consulting LLP.

Thomas believes a data integration strategy should establish “a strong foundation for technological innovation.” Companies can scale resources up or down with a hybrid cloud to meet changing business operations and satisfy regulatory requirements. He adds, “This forward-looking approach can enable organizations to evolve with the ever-increasing volume of data instead of merely reacting to it.”

The Unified Approach

Data lakehouses have become the “it” cloud data platform because of their success in managing enormous amounts of structured and unstructured data—something neither data lakes nor data warehouses do as well or as cost-effectively. Adopters tout their “unified storage,” which integrates data lake storage with data warehouse processing capabilities. Lakehouses store raw data in its native format—lowering storage costs and reducing the need for extract, transform, load (ETL) processes, a common but relatively cumbersome and expensive means of integrating data stored in lakes and processed in warehouses. Unifying structured and unstructured data in lakehouses ultimately improves data quality by employing schema enforcement that disallows inconsistent data types and reduces errors—steps that minimize data movement between systems.

As a result of these advantages, “lakehouse architecture is becoming more and more prominent,” contends Matteo Colombo, principal and a global data, cloud, and AI leader in the Seattle office of KPMG, a management advisory firm based in Amstelveen, Netherlands, who believes the most significant impact is on cost management. “We have seen a 40% to 60% total data infrastructure cost reduction, which is certainly very positive.” He describes the mix of data lake and data warehouse features as “the best of both worlds,” which can eliminate data silos. A centralized user interface provides “value for all types of user personas, such as data scientists, engineers, and business analysts.”

Lakehouse architecture pairs well with a hybrid cloud, especially for its data and security consistency and resource optimization. “A well-implemented lakehouse architecture empowers enterprises to manage growing data volumes efficiently while ensuring analytical flexibility, governance, and cost control—critical in today’s cloud-first environment,” says Thomas. “The singular platform is important to increase agility so that business leaders and data scientists can all access the same data without requiring multiple transformations or complex data movement.”

Unlike a data warehouse, “lakehouses leverage cloud-native storage and compute separation, allowing organizations to scale storage independently of processing power and optimize costs,” adds Thomas. “Other potential benefits include improved query performance and real-time processing, which enable real-time analytics use cases.” He cites Deloitte’s “The State of Generative AI in the Enterprise” study, which found that “70% of respondents are investing in such data management capabilities.”

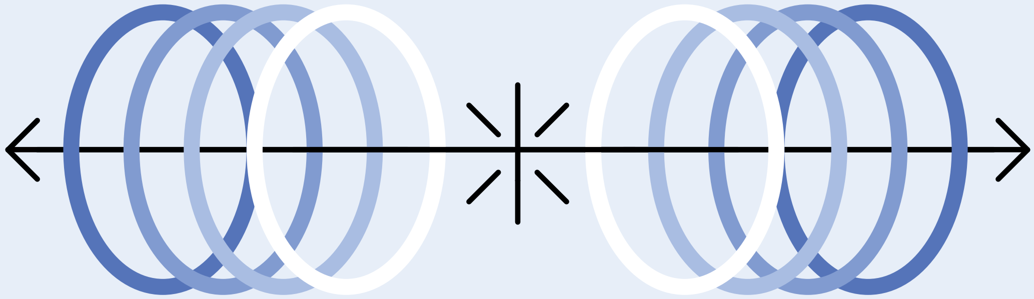
As organizations weigh the strategic and practical advantages of deploying lakehouse architectures for managing data and facilitating analytics, Ghosh notes that “usually a decision has to be made whether a data lakehouse should sit on prem or in the cloud.” Lakehouses typically run AI workloads in the cloud when organizations need robust compute and storage scalability, but for real-time edge processing, on premises can be a better option from a cost and performance standpoint. “Very rarely [do] you have a situation where you just keep the data fragmented in multiple places,” adds Ghosh. He warns that “joined-up reporting”—also called joint analysis—will be slow and complex in that situation. “You have to ensure that the data comes together in one place, most likely in the cloud.”

As the quest to improve decision making drives corporate investment in hybrid cloud data platforms, organizations must avoid derailing their transactional systems—the enterprise resource planning (ERP) or customer relationship management (CRM) systems that run their business—notes Ghosh, who advocates lakehouses for that reason. “The first advantage is decoupling the system of records from the system of analytics,” he says. “You don’t want to slow down that system [of records] when pulling data from it for reports and analytics. You pull the data only once, put it in the lakehouse, and then you can query that multiple times, so you reduce the back-and-forth between your sources and your consumption.”

Ghosh says companies often need to combine data from multiple sources to get a “cross-sectional functional analysis” of their business. “So, suddenly, you have CRM, ERP, external data, HR, and at least four or five different sources of information that need to be combined for any meaningful joint analysis. The data lakehouse is the place where you combine the data. You build a data model where the data is interrelated and combined so that you can get your analysis faster.”

As companies look to scale up their AI data analytics program to proliferate their data insights, they deploy lakehouses in a hybrid cloud to manage data ingestion’s increasing volume and velocity. Lakehouses support near-real-time data processing and provide centralized data management that can improve the accuracy of AI data models. Many companies use the hybrid cloud to process sensitive data on premises and run AI analytics in the cloud as allowed under regulatory conditions. Cloud cost-containment skills and strict data management policies are essential practices.

One tradeoff is that the analysis isn’t real time but near real time, explains Ghosh. “The data can be moved to a



“We have seen a 40% to 60% total data infrastructure cost reduction, which is certainly very positive.”

Matteo Colombo, principal and a global data, cloud, and AI leader in the Seattle office of KPMG

> **“Data fabric provides a centralized integration layer. In contrast, data mesh takes a decentralized approach, assigning data ownership to domain teams, ensuring localized security controls and federated governance.”**
>
> Deloitte Consulting’s Thomas

lakehouse within a few seconds. There will be a lag. You can reduce the lag to a small time limit.” But that’s not an issue, he says, because “more often than not, from a senior management perspective, they’re not looking for decisions every second.”

Data Journey Progressions

Modern enterprise data architecture aims for frictionless data flows—reducing data management overhead and improving data governance to help companies better manage a complex data landscape. Just as lakehouse architecture reduces data transfer from one location to another, the time- and resource-consuming ETL exercise, two distinctly different approaches to data management, aims to build an even more agile data ecosystem.

Data fabric, which helps automate data integration and management processes, sprang to life nearly 25 years ago to help companies master the complexities of distributed data management. It can provide data access without copying it, saving on storage requirements. Data mesh is a decentralized data architecture that emerged just five years ago that organizes data management around business domains, avoiding the bottlenecks that plague centralized approaches.

Data mesh “builds on existing data governance frameworks such as data products and data dictionaries to provide a single source of truth across complex data ecosystems that might cover multiple cloud services and hybrid deployments,” says Sim. Together, he says, the combination of data fabric and mesh “forms a powerful platform to build trust in today’s complex data ecosystem.” They go about it in different, if complementary, ways. He adds, “Data fabric builds on the data lake construct by including unstructured data and broader support for modern data governance tooling.”

These architectures strengthen data trust and bolster security, contends Thomas. “They can enable organizations to scale and trust their data while maintaining security and regulatory standards.” He explains how they achieve these goals in different ways. “Data fabric provides a centralized integration layer, automating metadata management, security policies, and compliance while ensuring data quality,” says Thomas. “In contrast, data mesh takes a decentralized approach, assigning data ownership to domain teams, ensuring localized security controls and federated governance.”

Ghosh says data mesh offers organizations significant savings on storage duplication and migration. “If something new gets added, you can work seamlessly because the mesh can read it and give you the analysis almost immediately,” he says. However, Ghosh cautions that data mesh requires “strong data governance, management, and cataloging. Understanding the data becomes very critical.”

Mesh and fabric, especially when combined with hybrid cloud and lakehouse technologies, offer organizations greater business flexibility and data governance options. They can also help organizations struggling to scale their data programs with a limited number of data science and data team specialists. Parasrampuria says, “Mesh is less about tech and more about a data operating model for domain-based data management and federated data governance.” She adds, “If you have a data lakehouse, progressing toward mesh seems like a natural move—it’s the maturity of your data journey.”

After Data Modernization

As companies gain proficiency in managing their data life cycle, they integrate diverse data sources and types, automate data preparation, and process data sets across on-premises and cloud environments. This progression enables them to leverage a broad range of AI capabilities, including machine learning algorithms, natural language processing, data visualization, and pattern recognition, to uncover impactful insights that drive business decisions.

But building a data-driven enterprise isn't merely mastering data modernization and optimizing the data production cycle, explains KPMG's Colombo. "Where I think organizations are falling short once the data is available is in understanding how to use it to drive change into their business. Many organizations invest hundreds of millions of dollars to standardize their data and production processes, but they completely ignore the consumption cycle, thinking that the moment they have quality data, they can enable a self-service mode and everything will be OK, but it's not necessarily the case."

Colombo believes that the organizational challenges of becoming data driven are often cultural. "Whether it's in terms of trust in the data or understanding the value of the data, we see a bit of a gap there," he says. "How do you solve that gap? Typically, you would say governance is the answer. The first thing that I think organizations need to address is building a strong data culture." Colombo argues it's not just about establishing roles and responsibilities and driving consistent policies. "If these individuals don't fully understand the ramifications of using data and the evolution into AI products, they will struggle with performing this role." He believes many companies "underinvest in the consumption cycle—educating their business on using the data, how to drive value from data, and building a culture of data."

Building a robust data culture is a process that often starts with establishing the value of the data and then developing data champions, explains Colombo, who spends "a lot of time educating business leaders on the potential of data in their context. Some of those leaders can become data product owners and articulate the value of those data products."

A similar situation occurred in digital transformation over the past decade. Organizations learned the importance of change management strategies to reach their goals. "Building a data-driven culture certainly comes with challenges, primarily aligning the data strategy with the company's broader business objectives," says Thomas. "Gaining buy-in from both internal and external stakeholders is essential for successful transformation. Cross-department collaboration is essential to eliminate silos and ensure data is easily accessible across teams. Effective change management should focus on fostering a continuous learning environment, helping employees adapt and integrate data into their daily roles for long-term success."

Companies shouldn't push AI agendas without carefully considering "grassroots" concerns, explains Ghosh. "If there are concerns, they need to be addressed, and organizations have to be prepared for a slightly longer duration for that adoption. The real benefit of any of these applications only comes if people use that application. More often than not, projects fail not because the technology fails, not because of the competence of the people who have built that application, but because the change management has failed."

Ghosh recommends beginning change management at the outset of the AI work. "If you get people on your side from the beginning, it actually improves the quality of the product," he adds. Ghosh cites the example of a development team building a data dashboard without consulting the customer. "The person has to be emotionally invested in that." Even if it is the "best dashboard, if I have not been involved from the beginning, my first reaction will be, well, I don't need it or I don't want to use it because this is not what I want."

Recent data indicates that AI transformation efforts have been stalled as much—if not more—by consumption and training issues as by technical hurdles like building AI data pipelines. According to the October 2024 "Where's the Value in AI?" report by Boston Consulting Group Inc. (BCG), a Boston-based management consultancy, a successful AI transformation pins the effort allocation at 10% to algorithms, 20% to tech, and 70% to people and processes.[6] The BCG report found that companies fall short of upskilling staff to enable them to adapt to the demands of an AI data-driven environment. A key finding is that "less than one-third of companies have upskilled one-quarter of their workforce to use AI. That's better than a year ago but far from where

Despite many rapid and remarkable AI advancements, notably in gen AI, most companies and workers are far from fully tapping its transformative capabilities.

companies need to be for workers to feel comfortable with such a job-threatening technology."

For example, workers, many in technology roles, must learn to adapt to the broader use of "agentic AI," which involves using software agents and automation to perform autonomous tasks such as validating and auditing data before processing. As these intelligent agents graduate from pilot tests into production, data workflows will change and many workers will be asked to learn new skills. Parasrampuria estimates that businesses must "upskill, reskill, and redesign … up to 80% of the jobs in the organization."

Conclusion

Few companies become data driven or incorporate AI into their operations without attempting some form of data transformation. Organizations are increasingly adopting the hybrid cloud model to create a secure and scalable data pipeline that generates consistently trustworthy AI data insights in a regulatory-safe manner. Proper data governance throughout the AI data life cycle enables organizations to maximize the value of their AI investments, manage risk, and improve their business decisions.

Despite many rapid and remarkable AI advancements, notably in gen AI, most companies and workers are far from fully tapping its transformative capabilities. Organizations tend to prioritize the production—rather than the consumption —of AI insights. This preference has meant that resources tend to flow more freely into building and managing AI data pipelines than into educating the employees whose decision making and daily workstreams depend on operationalizing those insights.

But early adopters, often in information-savvy domains such as finance, see AI in action. Ghosh is starting to see the payoff of AI-driven investments that deliver automated management reports and compliance paperwork for specific business functions. As a result, finance teams can focus less on compliance and control and more on performing higher-value tasks such as "business analytics, business partnering, and insights-driven" initiatives, he says.

"That's one small example of how organizations are questioning some of the traditional ways in which they used to work and taking advantage of the technological advancements in the field of data and analytics," adds Ghosh. Ultimately, the aim is to "redefine those functions and make [the workers] more future-ready and more successful."

Endnotes

[1] Enterprise Strategy Group, “Spending Intentions Survey,” December 2024.
https://research.esg-global.com/reportaction/515201716/Marketing.

[2] Harvard Business Review Analytic Services, “HBR-AS/Profisee Final Survey Results,” January 2024.
https://hbr.org/sponsored/2024/05/data-readiness-for-the-ai-revolution.

[3] McKinsey & Co., “The Data-Driven Enterprise of 2025,” January 2022.
https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025.

[4] Deloitte Consulting LLP, “Focusing on the Foundation,” June-July 2024.
https://www2.deloitte.com/us/en/insights/topics/digital-transformation/where-are-organizations-getting-the-most-roi-from-tech-investments.html.

[5] Deloitte, “The State of Generative AI in the Enterprise,” July-September 2024.
https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html.

[6] Boston Consulting Group, “BCG 2024 Global Study on AI and Digital Maturity,” October 2024.
https://www.bcg.com/publications/2024/wheres-value-in-ai.

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.