SPONSORED CONTENT | WHITE PAPER

Data architecture and strategy in the AI era

CIO

SPONSORED BY

CLOUDERA

Large organizations understand and appreciate the value of unifying the data lifecycle on a single platform as a springboard to advanced analytics and artificial intelligence (AI). Still, they struggle with managing data volumes and complexity, security concerns, governance issues, and a proliferation of data silos, according to a new study by Foundry on behalf of Cloudera.

Survey respondents agreed that cloud computing provides greater flexibility and shorter development times, even though nearly one-quarter of them still process their workloads entirely in their data centers. More than four in five agreed that having one place to run and manage all applications and data across clouds and on-premises infrastructure is critical.

A substantial majority said emerging architectures such as data lakehouses — which combine the flexibility of data lakes and the performance of structured data warehouses — reduce complexity, but many have yet to adopt them. However, the survey found that new concepts such as data mesh and data fabric are quickly catching on and will be in place at about half of large organizations within 18 months.

The AI imperative

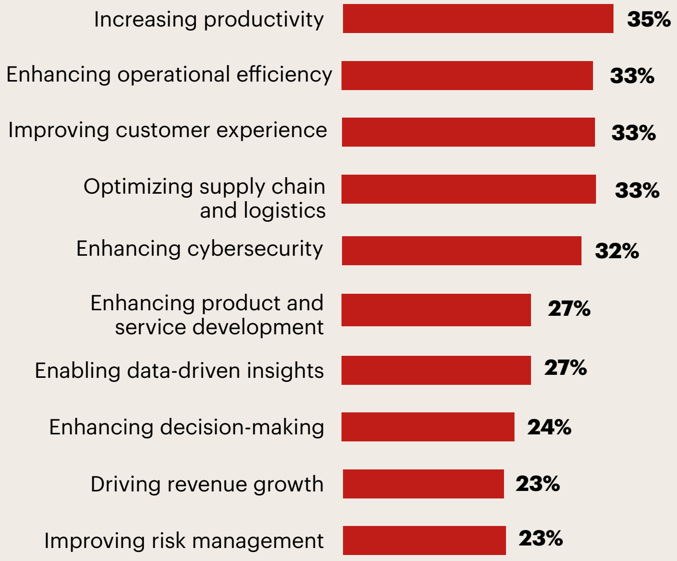
It has been said that data is the new oil, but data is arguably much more valuable. Unlike petroleum, information can be reused, stored, combined innovatively, shared, copied, and used as the foundation for critical decisions. As it is used, data also creates more data and additional value. The ability of organizations to find, classify, and expose data to all who need it, in a safe and compliant manner, will separate the winners from the others.

The value of data has been brought home in 2023 by the astonishing popularity of generative AI (GenAI). The ability of large language models (LLMs) to understand and respond to complex questions, generate original content, develop software, and synthesize millions of data sources into practical advice has captivated IT and business leaders. GenAI makes insights available to anyone who can hold a conversation and expands the population of users who can derive value from data-driven insights. Gartner predicts that over 80% of enterprises will adopt GenAI APIs and models or deploy GenAI-enabled applications in production environments by 2028, up from approximately 5% in early 2023.

Among the Foundry respondents, three out of five said they are at least in the early stages of adopting AI, with only 8% saying they have yet to make plans for AI-related projects.

Those early adopters expect various benefits, including increased productivity, improved operational efficiency, enhanced customer experience, supply chain benefits, robust security, and risk management (see Figure 1).

Figure 1: AI holds potential across the enterprise

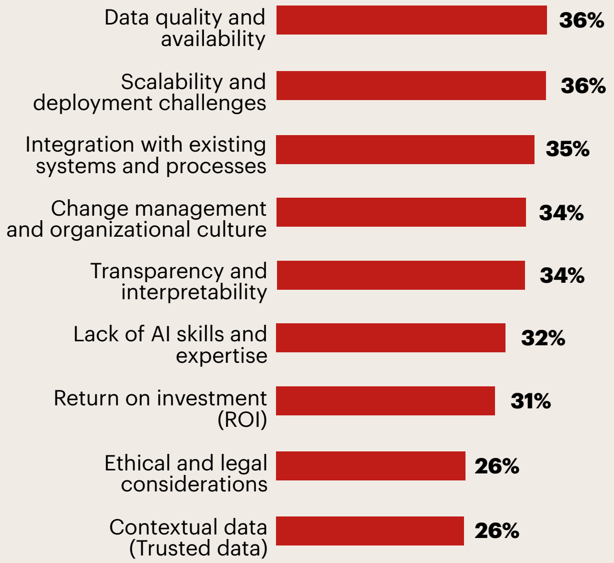


Source: Foundry

There are numerous challenges to achieving AI proficiency, however. About one-third of the respondents mentioned the quality and availability of data scalability, integration with existing systems, change management, and model transparency as hurdles. Respondents also pointed to skills shortages and questionable return on investment (ROI).

The volume, complexity of data, and security concerns also hamper the end-to-end data management needed for AI model development (see Figure 2).

Figure 2: The many challenges of AI at scale



The road to success

Organizations that want to achieve the promise of AI need three things: a modern data architecture; unified data management; and versatile, secure data platforms. The survey indicated that companies leading the way toward AI adoption are implementing the following structures.

- **Modern data architecture**

Building a modern data architecture Is significantly accelerated by adopting a single data platform that works seamlessly across cloud and on-premises infrastructure. Whether deployed "privately" on-premises or in a public cloud, cloud-native computing is the preferred architecture for organizations seeking to unify their data platforms and set the stage for AI model training and inferencing.

A flexible approach such as utilizing data lakes or data lakehouses is ideally suited for managing the large volumes of unstructured and semistructured data needed for AI model training. IT decision-makers recognize this, as evidenced by the two-thirds who agreed that data lakehouses help reduce pipeline complexity. The fact that fewer than two in five enterprises currently use them serves to illustrate the difficulty of integrating significant new data management platforms.

The survey also revealed considerable interest in two emerging data management concepts: data mesh and data fabric. Although they sound similar, the principles are quite different.

A data mesh is a data management paradigm based on four principles: decentralization, data as a product,self-service analytics, and federated governance. It distributes data ownership and responsibility across various teams and domains, using the concept of "data products," which are domain-specific packages of data services. Governance is federated, but ownership is entrusted to the teams closest to the data, and they are encouraged to treat data just like any of the company's other products.

A data fabric is distributed and takes a more centralized approach to data management. The objective is to unlock data sources at scale in an automated manner that promotes context and business relevance.

Based on that insight, data can be made available in a safe, compliant, and self-service manner across the organization, using a single abstraction layer that hides the complexity of the underlying data sources.

Enthusiasm for the data mesh concept is pronounced. Even though the approach was defined only two years ago, 54% of the respondents said they expect to have a mesh in place by the end of 2024; nearly half (48%) also plan to implement a data fabric in that time frame. Interestingly, organizations with entirely on-premises infrastructure have already deployed a data mesh by a 45%-to-35% margin over those using public cloud. This may indicate that a mesh is more straightforward to implement when resources are confined to local infrastructure instead of being spread across multiple locations.

Data mesh vs. data fabric

Data mesh	Data fabric
• It uses “data products,” domain-specific packages of data services. • Ownership and governance are entrusted to the teams closest to the data. • Data owners are encouraged to treat data like any of the company’s other products.	• It takes a centralized approach to data management. • It creates a unified data architecture in which data is seamlessly connected and accessible, using a single abstraction layer • The abstraction layer hides the complexity of the underlying data sources.

Enterprise data strategy across industries and regions

About half of the organizations surveyed reported having a formal enterprise-wide data strategy, but the prevalence varies widely by industry. Those using the public cloud are likelier to have a strategy than those with primarily on-premises computing infrastructure.

A significant 85% also said their data strategy effectively enables AI/machine learning (ML). Respondents in the northern European region of Europe, the Middle East, and Africa (EMEA) were more likely than those in other regions to rate their strategy as “very effective.”

On an industry basis, manufacturing and financial services firms lead the way; about two-thirds have a data strategy. Healthcare, retail, and telecommunications firms are the least likely to have a plan.

EMEA respondents led their North American and APAC counterparts in strategy adoption by about a five-to-four margin.

Although not essential to AI development, federated data strategies enable data understanding at an enterprise scale and accelerate business decisions by putting them in the hands of those best equipped to use them in an agile and flexible manner unfettered by reliance on centralized IT.

Current status of enterprise data strategy

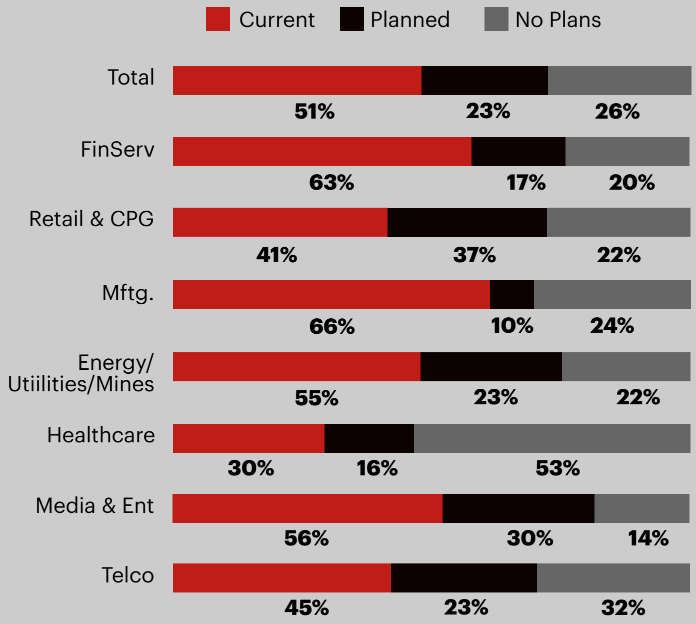
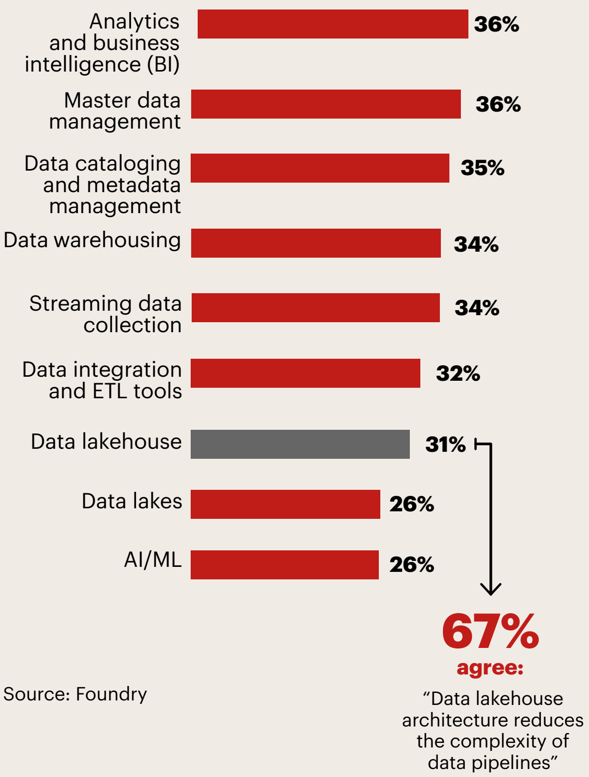


Figure 3: How do you manage data?

While fewer than two-in-five enterprises use data lakehouses, two-thirds of decision-makers agree they help reduce pipeline complexity



Source: Foundry

- **Unified data management**

An overwhelming 90% of all respondents agreed that unifying the data lifecycle on a single platform is critical for analytics and AI. Nearly all perform fundamental data tasks such as ingestion, monitoring, and data pipeline processing, and 97% use traditional business intelligence tools.

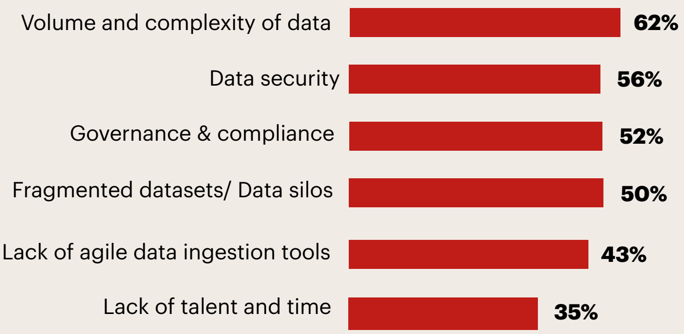
About 80% of organizations have reached the predict and/or publish phases of data and analytics, meaning that they have conducted at least some modeling and have deployed analytics models throughout the organization.

A slightly smaller but still significant proportion of the respondents, three-quarters, also conducts modeling, training, and data visualization.

Organizations see a wide variety of benefits to modern data architectures, including:

- Simplifying data/analytics processes (40%)
- Gaining flexibility in handling all types of data (38%)
- Enhancing data governance and security (37%)
- Enabling easier integration with new tools and models for AI (35%)
- Improved scalability (32%)

Figure 4: What's holding end-to-end data management back?



Source: Foundry

Almost half of the respondents (46%) reported that their organization interacts with all stages of the data lifecycle process. Complete control of and visibility into every aspect of data give them the capabilities required to drive AI-fueled innovation.

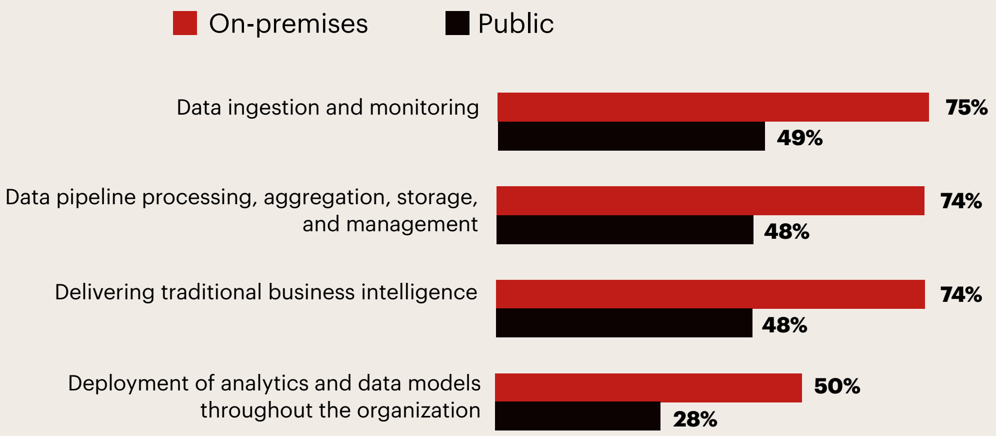
- **Versatile and secure processing platforms**

An overwhelming 89% of the IT decision-makers use the public cloud to manage data and ML analytics, whereas just 35% still rely on on-premises or private cloud processing (some use both, causing the total to exceed 100%), according to the Foundry study. A hybrid data management approach comprising both on-premises and public cloud infrastructure is the preferred data management strategy for the long term. Although only one-third of the respondents currently deploy multicloud/hybrid data architectures, almost all agree that they offer significant benefits, with 93% agreeing that "multicloud/hybrid capabilities for data and analytics are key for an organization to adapt to change."

Companies with at least some operations in the cloud reported higher overall use of the advanced data management systems needed for AI than those entirely on-premises. This is probably because this leading-edge software can be quickly deployed in the cloud without incurring the expense and long lead times of local installation.

However, companies whose infrastructure is primarily on-premises reported much higher overall use of technologies and processes in preparing data for analysis than those using a single public cloud (see Figure 5). This may be due to differences in data sources, compliance/security regulations, or industry standards.

Figure 5: Where is data prep done: On-premises vs. public cloud



***Some companies have deployed across hybrid and clouds, causing the total to exceed 100%.**

Source: Foundry

Moving to the cloud isn't without its challenges, however. Difficulties migrating data across different platforms and environments were cited by 35% of the respondents, followed by 28% who mentioned disaster recovery/business continuity, integration and interoperability, data visibility and data analytics, and orchestration for AI use cases.

Nearly one-fifth of the organizations also reported planning to repatriate data from the public cloud back to on-premises infrastructure over the next 18 to 36 months. A hybrid platform provides a choice of where to deploy analytics and AI, based on the needs of different workloads, teams, and departments as well as business and regulatory environment changes.

Data security, cloud cost, and compliance governance were also noted by at least one-quarter of the respondents, indicating difficulties in implementing hybrid and multicloud environments, depending on the organization and the use case.

Migrating data: differences by region

There were some notable regional variances regarding migrating data:

- EMEA respondents cited data security as a significantly lesser issue than did their North American counterparts.
- More than twice as many EMEA as Asia-Pacific (APAC) respondents mentioned cloud costs as an impediment.
- Just half as many North American as EMEA executives said balancing the needs of different stakeholders within the organization is a problem.
- North American respondents were much less likely to say data silos are an issue than did their EMEA or APAC counterparts.

Orchestrating multiple sources

Delivering business outcomes and building high-quality AI models typically involve integrating information from multiple sources but also introduce complexity. Again, the survey showed that organizations with one or more clouds can orchestrate data more smoothly from various sources.

Enterprises currently focus on critical data sources holding customer/prospect data, supply chain data, and customer sentiment data. More than 70% of the respondents reported using these types of data sources within their organization.

More than half said they also use other data sources, led by economic data, sensor data, market data, voice/images/text, and publicly available data.

Volume and complexity challenges grow as more sources are employed. Companies that use multiple clouds reported slightly higher overall usage of all data sources than those that deploy only a single public cloud or exclusively on-premises infrastructure. For example, 93% of the organizations that use multiple clouds incorporate customer and prospect data into their analytical models, compared to 75% of the on-premises users. APAC respondents reported higher overall use of every data source than those in other geographies.

Managing data complexity

AI model training and fine-tuning require huge amounts of data. Enterprises use a variety of solutions to manage data, such as:

- Analytics/business intelligence (62%)
- Master data management (58%)
- Data cataloging (57%)

Machine learning and data lakes ranked lowest, with a 34% usage rate, but those technologies are less well established than others. APAC residents indicated above-average use of all but one of the nine suggested data management categories: data warehousing.

Not quite ready for real time

Many AI use cases — fraud detection, online shopping recommendations, advertising, healthcare monitoring — require real-time streaming data. The survey indicated that many organizations don't see real-time capability as critical, at least not yet. Almost half (45%) rely primarily on historical data to make business-critical decisions, and 81% use batch data collection. About one-quarter process near-real-time data, and only one-quarter are equipped to work with streaming data. Half of the IT and data leaders said their organization's data management architecture doesn't meet the needs of real-time use cases.

They may not think it necessary. Of the 51% of respondents whose organization doesn't currently process streaming data, just 4% said they plan to add that capability in the next 18 months; that percentage is probably so low because business conditions don't require it.

Conditions are changing, however. IDC expects the stream processing market to grow at a compound annual growth rate (CAGR) of more than 21% through 2028, driven by data volumes, the need for real-time analytics, and the growing adoption of intelligent internet of things (IoT) devices.

Among the top roadblocks to moving and collecting data at high speed, according to respondents, are difficulty managing pipelines, security/governance challenges, and the need for extensive customization. Security and governance are seen as more significant challenges in EMEA and APAC than in North America.

Metrics, value and impact

The metrics that organizations use to measure the success of a data strategy relate primarily to bottom-line results, with revenue growth and cost savings at the top of the list. However, many other factors received numerous mentions, including achieving environmental, social, and corporate governance goals; reducing business risk; increasing customer satisfaction; achieving shorter time to market; and improving business agility.

Companies that employ primarily an on-premises or hybrid architecture pointed to revenue growth and cost savings as critical metrics much more than those using cloud resources. Conversely, just 9% of the respondents with a hybrid cloud strategy said business agility is a crucial success metric, compared to 48% of those who use public cloud only. This illustrates how much infrastructure choices are influenced by business strategy. Public cloud appeals to organizations that value speed and flexibility, whereas owned or leased data centers appeal more to those that value security and control.

> **"The research shows that treating data as a critical asset is an essential organizational skill, and creating a data architecture grounded in business strategy is the foundation."**

The bottom line

The research shows that treating data as a critical asset is an essential organizational skill. A modern data architecture grounded in business strategy is the foundation. Flexible and scalable cloud management technologies provide the tools to turn information into insights and facilitate AI model training and inferencing. A single data platform that spans local and cloud infrastructure gives organizations the power to process data wherever it's needed and to share it seamlessly with stakeholders and business partners. It provides for unified governance, consistent data quality, and the scalability needed to adopt AI models of increasing size and sophistication.

Trusted data is the foundation for trusted analytics and AI models. That factor is significant at this time of transition, when concerns about transparency, privacy, and respect for intellectual property are major points of concern. A global study of 17,000 people by KPMG and the University of Queensland found that over 60% are wary about trusting AI systems and only half believe that the benefits of AI outweigh the risks.

And yet Cloudera's research shows a huge shift in attitudes toward AI, ML, and data analytics and "great cause for optimism in how it can be used to shape the future through creating wider opportunities and supporting communities." The organizations that achieve the full potential of AI will be the ones that capitalize on these opportunities by demonstrating high levels of confidence in training data, model integrity, and respect for security and privacy. They will be in the best position to respond to change and drive innovation.

> "There is great cause for optimism on how AI can be used to shape the future through creating wider opportunities and supporting communities."

For more information, visit cloudera.com/solutions/data-leaders.

About the study

The Foundry research surveyed more than 600 IT decision-makers in North America, the northern Europe region of EMEA, and APAC.

The target company size was organizations with annual revenue of more than $500 million or more than 1,000 employees globally. All participants were data leaders and IT decision-makers with titles of director and above (or equivalent). They had to have a prominent role in the selection of data-related products and services, including, but not limited to, infrastructure.

CIO

SPONSORED BY

CLOUDERA