FORRESTER®

Use AI Via An End-To-End Data Lakehouse To Increase Data Lifecycle Efficiency From Ingestion To Prediction

Enable Employees To Generate And Execute On Valuable Insights With The Right Data Platform

A FORRESTER CONSULTING THOUGHT LEADERSHIP PAPER COMMISSIONED BY CLOUDER, INTEL AND HPE, AUGUST 2023

Table Of Contents

Project Team:
Madeline Harrell,
Market Impact Consultant

Kate Pesa,
Associate Market Impact Consultant

Contributing Research:
Forrester's technology architecture and delivery research group

ABOUT FORRESTER CONSULTING

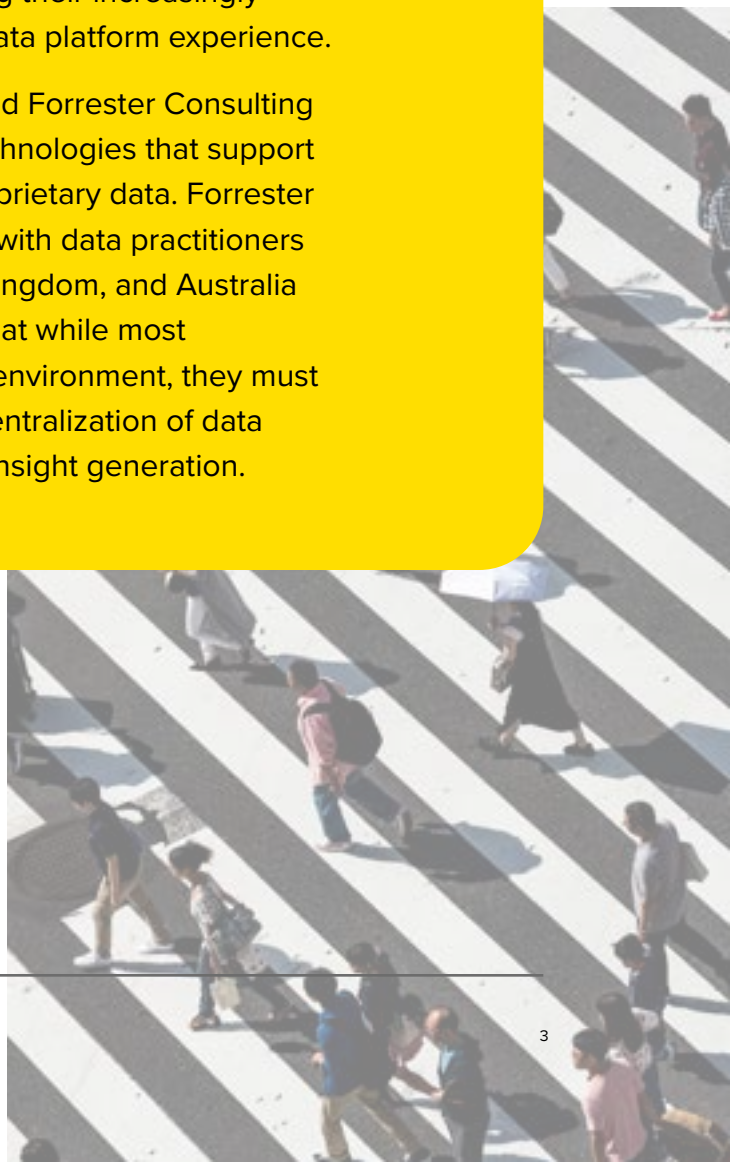
Forrester provides independent and objective research-based consulting to help leaders deliver key transformation outcomes. Fueled by our customer-obsessed research, Forrester's seasoned consultants partner with leaders to execute on their priorities using a unique engagement model that tailors to diverse needs and ensures lasting impact. For more information, visit forrester.com/consulting.

Executive Summary

Enterprise organizations are on a transformational journey to improve time to value in their data engineering, analytics, and machine learning processes. An increase in the amount of structured and unstructured data, and the number of tools used to derive value from that has reduced productivity and profitability. Data decision-makers and practitioners struggle to perform their core job functions while operating in environments that do not combine data management, analytics, data science functions and extract, transform, load (ETL).

To increase productivity and profitability, data scientists need access to a seamless data experience that combines management, analytics, and data science functions to promote interoperability, automation, security, and governance. These innovations will ease deployment so that data scientists can deliver end-to-end results quickly with less resources. Data decision-makers will need to be aware of where they are in their data journey and what they can achieve by enabling their increasingly distributed data science teams with a centralized data platform experience.

In May 2023, Cloudera, Intel and HPE commissioned Forrester Consulting to evaluate how organizations are choosing the technologies that support the storage, management, and analysis of their proprietary data. Forrester conducted an online survey with 840 respondents with data practitioners and decision-makers in the United States, United Kingdom, and Australia and New Zealand to explore this topic. We found that while most organizations have begun to modernize their data environment, they must prioritize the hybridization of their teams and the centralization of data lifecycle steps to reap benefits in productivity and insight generation.



Key Findings

Data science teams — and data — are becoming more distributed across organizations. Data science teams are becoming more decentralized, frequently by adopting hybrid models of reporting into the business. At the same time, data decision-makers are adopting point solutions that meet immediate needs instead of making larger, strategic purchases that centralize their data and data management functions.

Environments with too many tools are costing employees valuable productivity time in their work day. Employees are not frustrated with the performance of individual solutions, but rather the large number of tools, challenges in activating machine learning models, and the lack of effective integration.

End-to-end data lakehouses reduce the complexity of the data environment, and improve the employee experience (EX). Adopting an end-to-end lakehouse that consolidates data tool functionality eases the stress of managing the full data lifecycle. It also provides data ownership clarity, ultimately saving valuable company revenue.

Data And Data Science Teams Are Becoming More Distributed

Today's changing work environments are forcing organizations to evolve how they manage their data; this is thanks to an increase in distributed data across public cloud, private cloud, and hybrid cloud environments, constantly evolving compliance and governance standards, and new threats to data security. The ability to perform the steps of the data lifecycle within a single platform will become increasingly critical to the success of enterprises that want to provide excellent customer experience (CX). The consolidation of data tools into a single platform — or at least having less than eight tools — can reduce the time needed to perform core data functions, improve time to value, and also time to customer satisfaction. This demands streamlined integrations, reduced complex customization, and increased automation. Unfortunately, most respondents are still in an intermediate state of maturity, using eight or more tools to complete each step of the data lifecycle and losing hours of productive time in each workday. As data science teams become more hybridized across their organization, they will need a more access to more centralized functionality from their data tools. We found that:

- **Data science teams are more spread out across the organization.** Organizations are moving towards a hybrid model for their data science teams; this means the data science team is managed centrally, but members are assigned to work with specific business units rather than report into one data science team or into individual lines of business. Fifty-five percent of respondents said they report into a hybrid model, with another 17% reporting into a strictly decentralized model. With this move towards hybridized team structures, any move to streamline or consolidate their data-related functions will be a boon for productivity.
- **Data science teams have plenty of tools, but no long-term strategy behind them**. Respondents are fairly satisfied with the functionality and interoperability of their current data tools, but they seem to be settling for overlapping functionality and needless context-switching between tools or applications to manage each step in the data lifecycle. Respondents shared that each step in their organization's data lifecycle takes at least eight tools each, with publishing into the business taking at least 10 tools. (see Figure 1). Although 97% of respondents said their tools integrate on

some level with tools from other vendors, basic levels of integration are not going to cut it as teams — and data — become more decentralized. A hybridized team needs centralized data to reduce silos and increase productivity.

Figure 1

Number Of Tools Used For Each Step In The Data Lifecycle

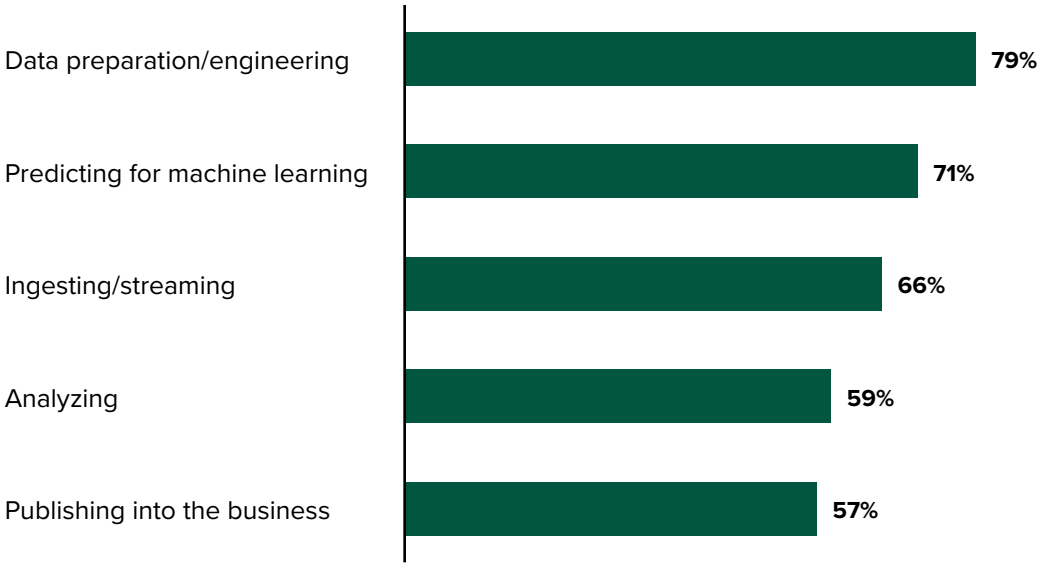
	MEAN
Ingesting/streaming	**8.3**
Data preparation/engineering	**9.2**
Analyzing	**9.2**
Predicting for machine learning	**9.2**
Publishing into the business	**10.4**

Base: 840 global practitioners and decision-makers in development and data science
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

- **Data decision-makers are purchasing data tools with current needs in mind, rather than future needs.** Not all data warehouses are created equal; while 68% of respondents currently have a data lakehouse and 83% have a data warehouse, the steps of their data lifecycle are not necessarily consolidated into the same lakehouse. When each step takes more than eight tools to complete, employees are spending valuable time switching between tools. Yet, 59% of respondents indicate that they select tools based on immediate needs. To effectively enable employees, data decision-makers will need to adopt a more holistic approach to their data lifecycle. That could mean zooming out to see the larger strategic picture and purchasing larger solutions that can both integrate their products on a deeper level and also enable employees to do more difficult data tasks, like creating machine learning models that can provide value to other departments and the business overall (see Figure 2).

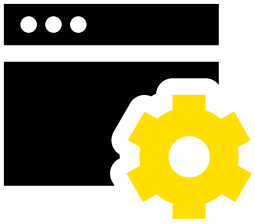
Figure 2

“Does your data platform perform any of the following functions?”



Base: 840 global practitioners and decision-makers in development and data science
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

Data scientists are using eight or more tools per step in the data lifecycle.



Too Many Tools Muddy The Waters

While data decision-makers continue to meet their immediate needs by packing their technology stack with new point solutions, employees are losing valuable time in their day to context switching between tools to perform each step in their data lifecycle. Employees are mostly satisfied with the interoperability and functionality of these tools, but they are not maximizing their productivity, especially when it comes to complex tasks like programming machine learning models to enable the rest of the business. On average, prediction for machine learning takes at least nine tools alone. Meanwhile, data decision-makers struggle to report the value of data to the business via their current data environment structure. Their challenges with the availability and quality of their data across different but somewhat solutions are leading to larger problems with machine learning, in driving revenue, and in reporting the overall value of data within the business. Respondents currently face:

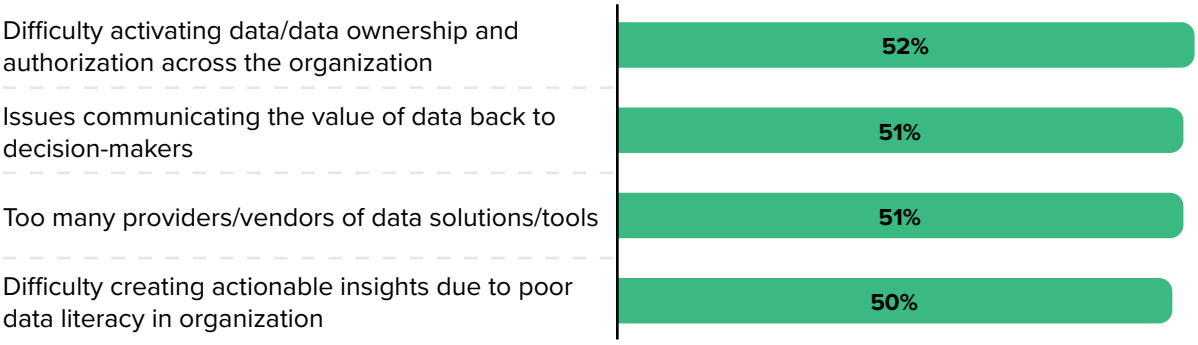
- **Technical challenges.** Namely, machine learning enablement and those related to insights. The most common technical challenges were related to enabling other teams with machine learning models (58%), difficulty drawing useful insights from their data (58%), difficulty integrating multiple products (58%), difficulty putting machine learning models into production (58%), and a lack of confidence in data security (57%). Using data to draw insights and build machine learning models to enable the business are core functions of data science workers, so their technical challenges with their current data environment are direct challenges to their value as employees. The difficulty in integrating multiple products is likely due to the number and manners of adoption of their data products, which also increases technical debt. As long as leadership continues to prioritize purchasing point solutions to meet immediate needs, their stack will keep growing — employees will continue with context switching between tools and applications, and they will have to hunt for the data they need to build the machine learning models they're already struggling with.

- **Organizational challenges.** More than half of respondents (51%) struggle with activating their data and understanding data ownership across their organizations. This can especially challenge hybrid or decentralized teams. Additionally, 60% of respondents also noted the lack of availability of all data to employees across the business. Naturally, this challenge hinders productivity and data literacy across the business. Where data practitioners could potentially enable cross-team collaboration, their data issues keep them boxed in. Unsurprisingly, this translates to issues with communicating the value of data back to decision-makers (51%). Furthermore, their habit of adopting point solutions to meet immediate needs has created a data environment with too many data solution providers or vendors (51%) (see Figure 3). Without the ability to demonstrate and communicate the value of data back to the business and decision-makers, data workers are stuck with their current data environment, unable to push for a larger, more strategic investment.

Figure 3

Technical Challenges Relating To Data Projects

Organizational Challenges Relating To Data Projects



Base: 840 global practitioners and decision-makers in development and data science
Note: Showing only top 4 challenges
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

- **Business consequences.** When employees cannot easily access the data necessary for them to perform their core job functions or enable teams across the business, this has larger implications for the business. Their data availability or ownership issues, and a large number of tools with varying levels of integration creates difficulty in meeting compliance and governance standards (63%), and increases vulnerability to cyberthreats (61%). These challenges also naturally lead to revenue implications: Respondents noted increased time to value (56%) and increased technical debt (54%) eating into their profits (see Figure 4).

Figure 4

Business Consequences Of Waiting To Invest In An End-to-End Data Lakehouse

Base: 840 global practitioners and decision-makers in development and data science
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

- **Challenges with switching to the new solution.** Investing in a platform they know can address these challenges is oftentimes easier said than done. An end-to-end lakehouse can address many of these challenges and enable employees to better support the business, but several things stand in their way of adopting a larger tool that can meet these needs. Beyond struggling with integration, 56% of respondents also have difficulty integrating with their cloud or on-prem infrastructure. Fifty-five percent of them are still adopting solutions that meet their immediate needs, and not strategic ones. Even if they found an appropriate platform to address their issues, they are dealing with the sunk cost of their existing vendor relationships (54%). Fearing an inability to integrate with existing infrastructure that they have already paid for with their current vendors, data decision-makers may feel like they are being limited.

75%

of respondents have acknowledged that they can save more than 4 hours each day, if the stages in the data lifecycle are integrated into a single platform.

End-to-End Lakehouses Enable End-to-End Machine Learning

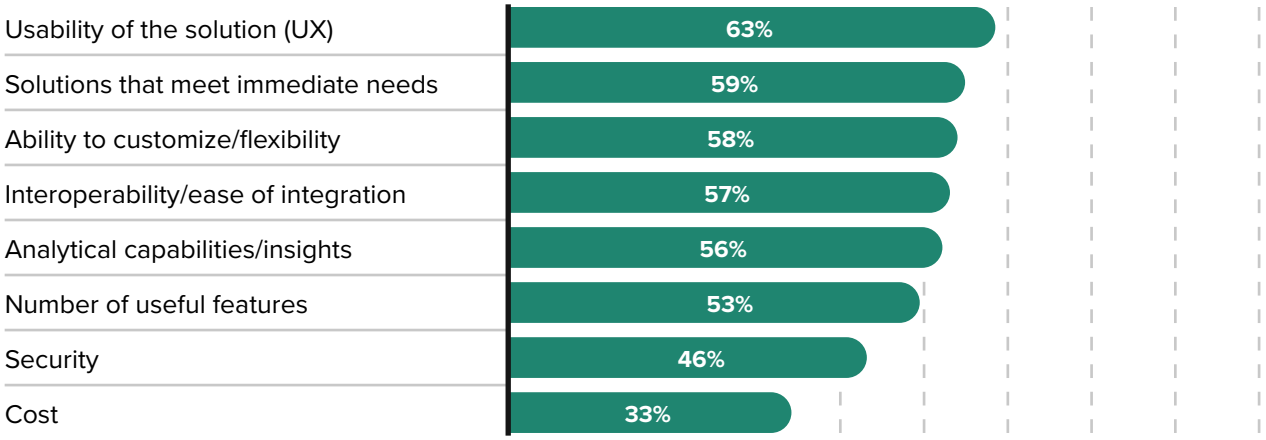
Adopting more point solutions to meet immediate business needs can only subsist for so long. As data practitioners are bogged down from switching between platforms to perform individual tasks, they will continue to be less productive over time. Consolidating the data lifecycle into a single end-to-end lakehouse — especially when it comes to machine-learning tasks — will increase productivity, reduce the time spent on context switching, and increase employee satisfaction. In fact, 94% of respondents say having an end-to-end lakehouse will positively impact their company's success, as:

- **Buyers know what to look for in a holistic data platform, especially for machine learning.** At a certain point, the need to keep data secure and usable should outweigh their vendor lock-in and the need to justify sunk cost on their current technology stack. When they do select a new data platform, data decision-makers expect a shared data experience or data fabric (48%), best-in-class integration with their existing infrastructure (47%), the ability to keep metadata synced (46%), end-to-end machine learning (46%), and improved security (44%). These desired capabilities and functions reflect their top challenges faced with their current data structure, such as issues with machine learning models, integrations with existing technology, and keeping their data secure.
    - **Machine learning enablement is a key priority.** When it comes to selecting a new data storage or management tool for machine learning, data decision-makers have several key capabilities in mind (see Figure 5). They are prioritizing the tool's usability (63%) before its ability to customize (58%) and its interoperability or ease of integration (57%). Decision-makers want to enable their practitioners to build their machine learning models immediately.
- **Organizational benefits.** An end-to-end lakehouse solution does more than just provide additional capabilities: It has the ability to level-up data quality to improve collaboration across teams (57%), streamline the ability to communicate machine learning results to the business (58%), enhance the ability to communicate the value of data back to decision-makers (56%), and clear data ownership across the organization (54%).

This addresses data activation issues and makes machine learning work for the business. Respondents also indicated that if all the stages of their data lifecycle were integrated into a single platform, nearly half of respondents (49%) could save at least 5 hours in a single workday, and 9% of them expected to save 7 or more hours (see Figure 6).

Figure 5

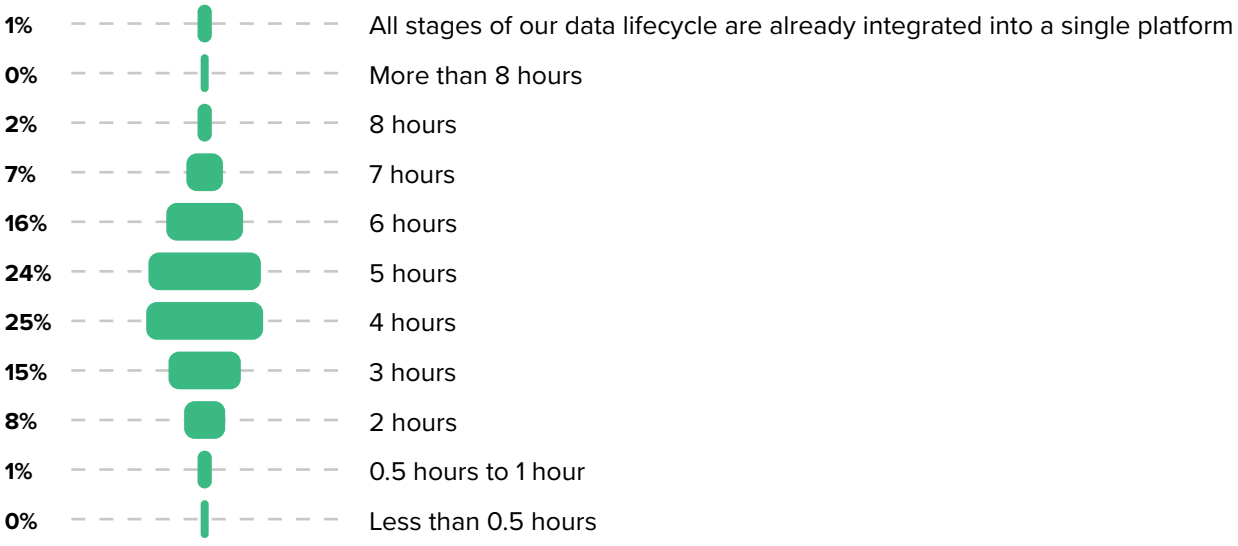
Biggest Drivers In Deciding Data Storage/Management Solutions To Purchase For Machine Learning



Base: 840 global practitioners and decision-makers in development and data science
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

Figure 6

"How much time in a single work day do you think you would save if all the stages of your data lifecycle were integrated into a single platform?"



Base: 840 global practitioners and decision-makers in development and data science
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

- **Positive business impacts.** The business will experience positive benefits after investing in an end-to-end lakehouse, including scaled data management and insight practices alongside business growth (53%), improved attribution of success to data efforts (50%), enhanced predictive analytics capabilities via machine learning (49%), improved collaboration across the organization (48%), and unified security across applications or data sources (48%) (see Figure 7). An end-to-end lakehouse improves employee productivity, collaboration across the organization, and the ability to attribute credit for success where it is due, all while keeping their data safe — these capabilities are lacking in most traditional lakehouses. While data decision-makers are currently focused on meeting the immediate needs of the business, sometimes a strategic approach can meet both current and future needs.

Figure 7

Expected Business Effects Of Having An End-to-End Data Lakehouse

Base: 840 global practitioners and decision-makers in development and data science
Source: A commissioned study conducted by Forrester Consulting on behalf of Clouder, Intel and HPE, May 2023

94%

of respondents say having an end-to-end lakehouse will positively impact their company's success.

Key Recommendations

Our research has uncovered the benefits that organizations can gain adopting an end-to-end data lakehouse, including reduced efforts in data integration, accelerated time to value for advanced analytics, and improved governance and productivity through enhanced collaborations. However, it is critical for IT decision-makers to consider how can they justify investing in another data product when their technology stack already contains numerous tools.

Align end-to-end data lakehouse adoption with technology upgrades and business objectives.

Leverage technology upgrades, digital transformations, or data migrations as an opportunity to introduce an end-to-end data lakehouse. While planning for infrastructure or software upgrades, emphasize the compatibility and integration benefits of a data lakehouse architecture. Showcase how it can leverage the latest technologies, such as cloud-based storage, distributed computing, and real-time processing, to optimize data management and analysis capabilities.

It is important to emphasize that solving technological challenges is not the sole value proposition. Businesses are seeking to foster a data-driven culture, enhance operational efficiency, improve CX, and ultimately drive revenue growth. By aligning the adoption of an end-to-end data lakehouse with both technology upgrades and overarching business objectives, organizations can realize the full potential of their data assets.

Associate end-to-end lakehouses with helping to break down data silos and improving productivity.

Organizations face the challenge of fragmented data across different systems and applications. Data scientists spend countless hours integrating these multiple data sources, products, and platforms to gain a holistic view of operations, customer behavior, or market trends. A key benefit of an end-to-end lakehouse is its seamless data integration and how it promotes interoperability, effectively breaking down data silos. It not only addresses

the complex nature of data integration but also boosts productivity. In fact, a staggering 75% of respondents have acknowledged that they could save more than 4 hours each day if the various stages of the data lifecycle were integrated into a single platform. An end-to-end data lakehouse enhances productivity by providing easy data access, self-service analytics, and fostering collaboration across teams.

Consider a phased approach and demonstrate ROI.

Our research indicates that obstacles in lakehouse adoption are namely: 1) sunk cost in existing technologies (54%), 2) focus on immediate needs (55%), and 3) lack of budget (43%). A phased approach to implement a data lakehouse along with a cost-benefit analysis to emphasize ROI could address challenges with end-to-end lakehouse adoption. Acknowledge that a full-scale implementation may not be feasible or desirable for every organization. Instead, prioritize specific use cases or departments where the benefits are most significant. Quantitatively calculate ROI through reduced time to insights and improved decision-making, and qualitatively emphasize enhanced data intelligence, which leads to revenue growth, cost optimization, and streamlined collaboration. This incremental approach demonstrates the value of the lakehouse concept, encouraging broader adoption across the organization.

Supercharge your machine learning and AI use cases.

AI-powered analytics, predictive modeling, and machine learning algorithms require access to large volumes of data for training models or analyzing complex patterns. Additionally, robust data governance is crucial to ensure compliance and maintain high data quality, which is essential for validating the accuracy of results. A data lakehouse enables data governance by offering data lineage tracking, data access controls, and a centralized data storage and management for structured, semi-structured, and unstructured data. It consolidates data from various sources, simplifying data quality issue identification and resolution. With greater scalability and flexibility, the data lakehouse supports advanced analytics — empowering data scientists and analysts to focus on analyzing data instead of looking for it.

Reduce tool proliferation in different stages of data lifecycle.

By implementing a comprehensive data lakehouse, organizations can streamline their data lifecycle processes, consolidate tools (i.e., eliminating the need for separate data warehouses and data marts), and ensure a unified approach to data handling throughout its lifecycle. This simplifies operations, enhances data consistency, and reduces the risk of errors and inconsistencies across various stages of the data lifecycle.

Appendix A: Methodology

In this study, Forrester conducted an online survey of 840 practitioners and decision-makers at organizations in the United States, United Kingdom, Australia, and New Zealand. Survey participants included decision-makers in development and data science roles. Questions provided to the participants asked about data storage and management solutions at their organization. Respondents were offered a small incentive as a thank you for time spent on the survey. The study began in April 2023 and was completed in May 2023.

Appendix B: Demographics

COUNTRY	
Australia	**34%**
United States	**28%**
United Kingdom	**22%**
New Zealand	**16%**

COMPANY SIZE	
500 to 999 employees	**22%**
1,000 to 4,999 employees	**56%**
5,000 to 19,999 employees	**19%**
20,000 or more employees	**2%**

TOP 5 INDUSTRIES	
Technology and/or technology services	**14%**
Telecommunications services	**13%**
Financial services	**12%**
Government	**12%**
Retail	**12%**

RESPONDENT LEVEL	
C-level executive	**24%**
Vice president	**28%**
Director	**23%**
Manager	**25%**

Note: Percentages may not total 100 due to rounding.


FORRESTER®