

2024

State of the Data Lakehouse

Survey Report from Dremio



The State of the Data Lakehouse

The data and AI technology landscape is in a constant state of rapid change. Businesses know they need to use data to compete, innovate, and succeed, but they struggle to deliver access to more of their data and to do so quickly, easily and cost-effectively. Central to meeting these challenges is a massive shift in foundational data architecture and management—**the rise of the data lakehouse**. Data lakehouses combine data warehouse functionality with the flexibility and scalability of data lakes, and if they are architected properly, they have the potential to improve data access, agility, and cost efficiency for analytics and AI workloads across industries. They have introduced new, streamlined data processes and have delivered value not previously available with warehouses and data lakes.

This survey of 500 full-time IT and data professionals from large enterprises offers fresh insights on data lakehouse adoption and associated cost savings, open table format trends, data mesh implementation for self-service, and use of the data lakehouse in building and improving AI models and applications. The survey also explores AI's impact on jobs and the most pressing issues of the day, reflecting the unique perspectives of this highly technical cohort. The majority of respondents were IT, data, and analytics managers and directors. Data scientists, software engineers, data analysts, and data engineers also contributed to the survey results.

Key Findings

The survey yielded key data in four major areas that help explain the current state of the data lakehouse. The body of this report dives into each of the summaries below, offering additional statistics and discussions about patterns the survey reveals.

01

With adoption surging, the data lakehouse is becoming the primary architecture for delivering analytics. Cost efficiency and ease of use are the top reasons.

Among enterprise IT professionals, the lakehouse concept is well understood. The adoption trend sees businesses moving away from cloud data warehouses and toward the data lakehouse. As the former's promise of scale has been compromised by high costs, businesses have sought the savings and additional benefits of the data lakehouse, including unified data access, reduced data movement and data copies, and query acceleration.

70% of respondents say more than half of all analytics will be on the data lakehouse within three years, and **86%** said their organization plans to unify analytics data

Over Half (56%) expect they are saving more than **50%** on analytics by moving to the data lakehouse

42% moved from a cloud data warehouse to the data lakehouse—more than from any other environment. Top reasons for the shift were cost efficiency and ease of use

02

Open table formats are transformative and Apache Iceberg adoption is rising.

Amid the generative AI frenzy, a quieter revolution has been taking place: bringing full SQL capabilities to big data. Open table formats enable data warehouse functionality directly on the data lake, while preserving freedom, flexibility, and full control over an organization's most important asset: their data. While customers have options in terms of table formats, Apache Iceberg is quickly growing due to its performance, interoperability with other tools and execution engines, and broad open source and commercial support.

Almost a third (31%) of survey respondents are using Apache Iceberg now, and over a third are on Delta Lake (**39%**). Among those planning to adopt a table format in the next three years, more are choosing Iceberg than any other table format, (**Iceberg 29% v. Delta Lake 23%**).

03

Data mesh is at the heart of digital transformation, with full or partial implementations happening at most enterprises and expansion expected by nearly everyone.

Data mesh is increasingly a business-driven strategy as more organizations see value in leveraging teams to build and deliver high-quality data products. As a key technology enabling the success of data mesh strategies, the data lakehouse is making self-service analytics, domain-driven data ownership, data as a product, and federated governance a reality for teams on the ground.

Data mesh initiatives are driven more by business leaders and business units (**52%**) than by centralized IT teams

Top objectives of implementing a data mesh are improved data quality (**64%**) and data governance (**58%**); almost half or just over half of respondents also named agility, scalability, improved data access, and improved decision-making

84% of respondents have fully or partially implemented data mesh, and **97%** expect data mesh implementation to continue to expand in the next year

04

The data lakehouse is critical in the AI era for improving AI-driven data management, governance and compliance, as well as the work lives of IT professionals.

Data self-service, which data lakehouses enable, is fundamental for AI development, as the vast majority of respondents say their enterprise is using a lakehouse to support data scientists building and improving AI models and applications. With respect to job-related issues, a majority cited manual repetitive processes and manual data merging and reconciliation as problems, speaking to the need for more automation and AI-assisted data management and governance.

81% of respondents are using a data lakehouse to support data scientists building and improving AI models and applications

62% disliked manually merging and reconciling data from multiple sources, repetitive manual processes, and cleaning up raw data

Technical professionals overwhelmingly agree AI is a national security priority (**84%**), noteworthy in light of the recent [U.S. executive order on AI](#)

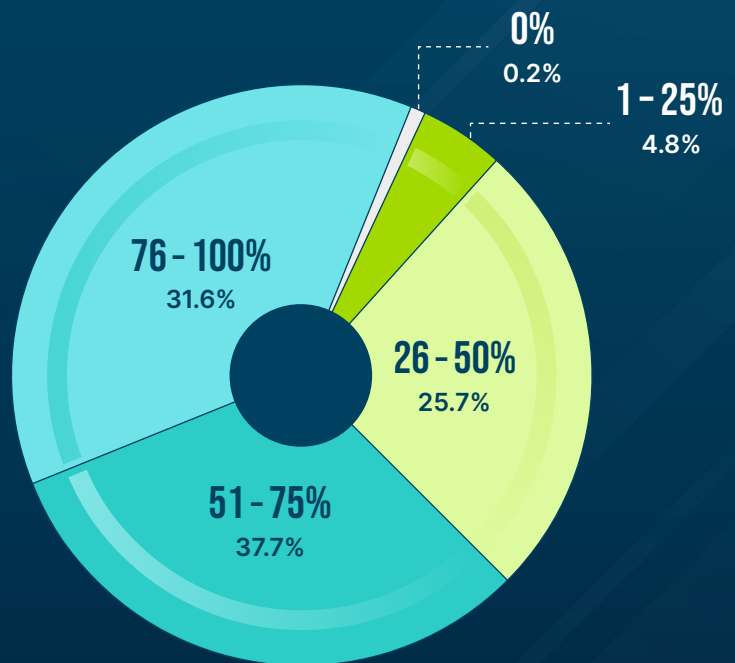
Data Lakehouse Adoption Trends

Almost 85% of enterprise IT and data professionals surveyed were very familiar with the concept of the data lakehouse. For them, ease of use and cost efficiency were the top reasons for adoption.

The survey data suggests that lakehouses are now the primary architecture for delivering analytics, surpassing cloud data warehouses, with 65% of respondents estimating more than half of their analytics are

running on a data lakehouse. The data showed a trend toward increasing volumes of analytics running on the lakehouse at organizations with more than 1,000 employees: 70% of IT and data professionals said more than half of their analytics will be on the lakehouse within three years. At enterprises with more than 10,000 employees, 78% said more than half in that same timeframe.

What percentage of your analytics do you predict will be running in a data lakehouse in the next three years?



Source: Dremio State of the Lakehouse Survey

Surging lakehouse adoption signals a fundamental shift in the way data is stored, organized, moved, processed, and consumed—happening right now in the enterprise. The reasons are many. In addition to delivering the capabilities of both traditional data warehouses and newer cloud data lakes, data lakehouses have empowered businesses with open data architectures that offer prime flexibility for different kinds of workloads and eliminate vendor-specific formats. Data lakehouses reduce or eliminate Extract, Transform, & Load (ETL) and ELT pipeline challenges and enable better data governance with fewer data copies.

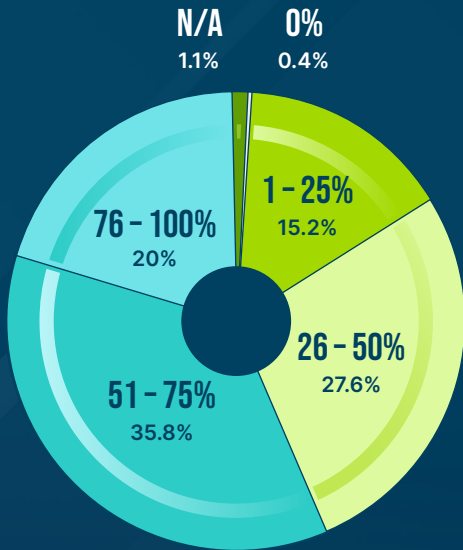
Specifically, data lakehouses can leverage git-inspired version control systems, which serve as a centralized repository that tracks and manages the history of changes made to tables and metadata.

Perhaps most importantly, lakehouses can dramatically reduce costs, profoundly accelerate queries and performance, and radically democratize data access through an architecture enabling federation—enabling data consumers to access and utilize data beyond a single storage system. This is crucial for the expansion of data analytics across business units.

Expected Cost Savings and Other Key Motivations for Adoption

Businesses not yet running analytics on a data lakehouse may be operating in much less efficient environments, like enterprise data warehouses, cloud data warehouses, and data lakes. Over half of respondents (56%) said their expected lakehouse savings are greater than 50%. Almost 30% of respondents from large enterprises with more than 10,000 employees expect their savings are greater than 75%.

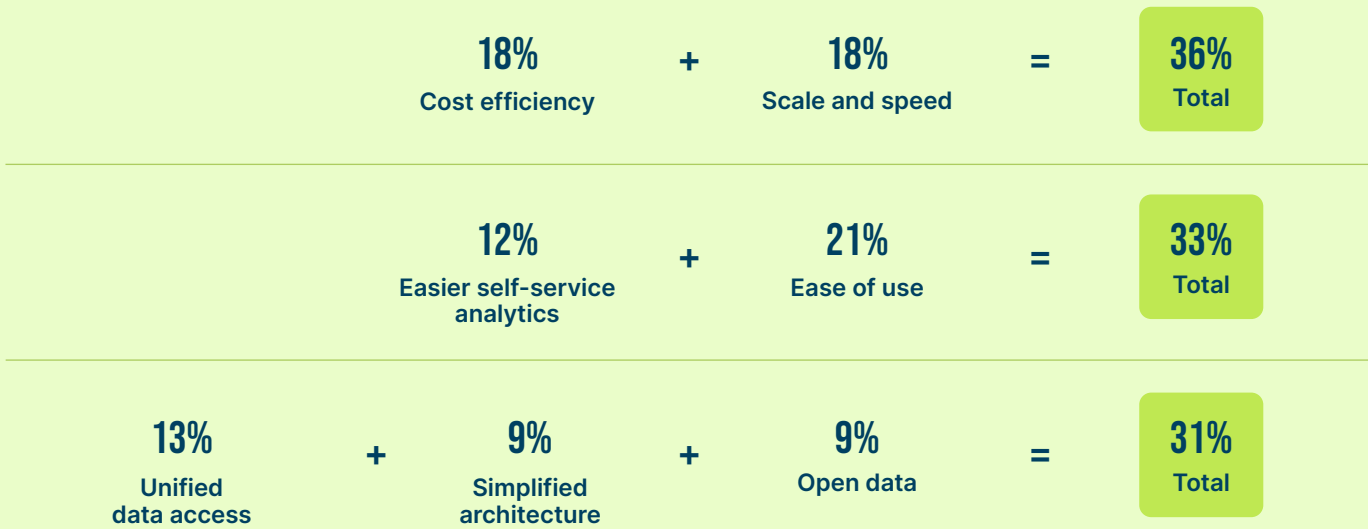
Expected savings on analytics costs by moving to a data lakehouse environment:



SAVINGS	RESPONDENTS
1-25%	15%
26-50%	28%
51-75%	36%
76-100%	20%*
NA	1%

*23% at enterprises with 5K employees, 28% at 10K+ employees

Overall motivations for adopting a data lakehouse included:

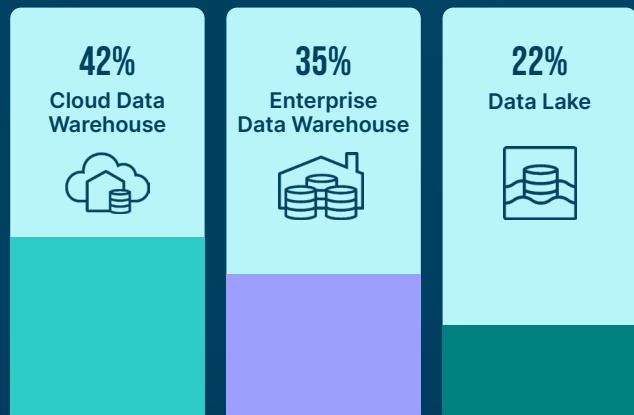


It's noteworthy that cost-efficiency and ease of use edged out other factors for organizations with 5,000 – 10,000 employees, and ease of use and speed were top priorities for organizations with 10K or more employees.

From the cloud data warehouse to the lakehouse

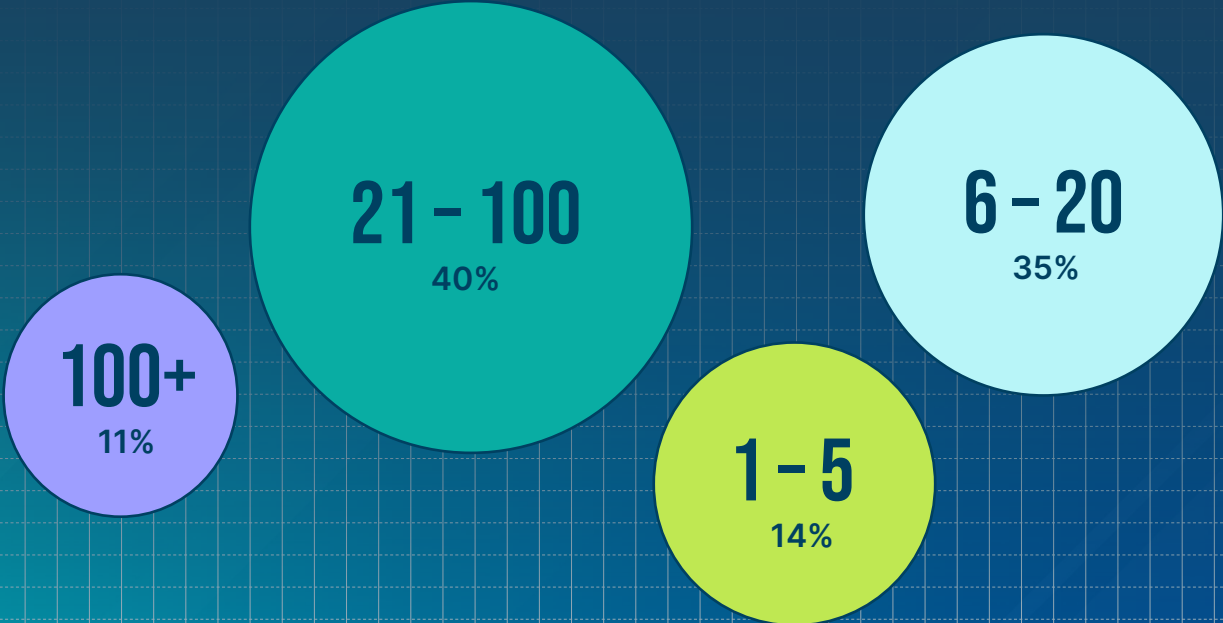
As adoption has intensified, more IT and data professionals report moving to data lakehouses from cloud data warehouses (42%) than from any other environment. We expect this trend to continue as more users experience the benefits of the data lakehouse. Across IT for more than a decade, there was a massive push into cloud data warehouses for the promises of scale and flexibility, but the high costs associated with architectural inefficiencies, including proliferating ETL pipelines and data copies, have outweighed the benefits for many enterprises. Development of data lakehouse architectures and technologies incorporate lessons learned during the advent and maturity of the cloud, including the preservation of interoperability and the reduction of data movement and data redundancy.

When adopting a data lakehouse, where did you move the data from?



As enterprises shift to data lakehouse adoption, they are looking to consolidate their analytics data—a powerful majority (86%) plan to unify this data in one place at some point. They are also juggling high numbers of data sources, with more than half (51%) reporting they have 20 to 100 or more data sources.

How many sources of data do you have today?



SOURCES

RESPONDENTS

1-5

14%

6-20

35%

21-100

40%

100+

11%

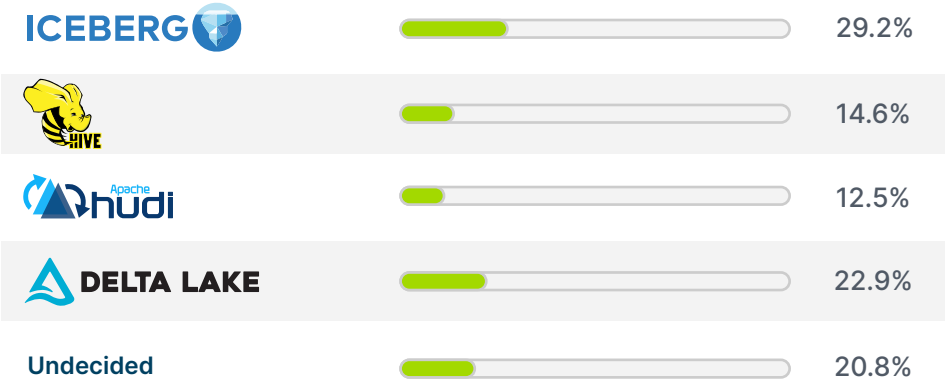
Open table formats are transformative and Apache Iceberg adoption is rising

While generative AI is the big data story making headlines, another significant technology shift is taking place for enterprise data teams: the adoption of open table formats. Open table formats bring full SQL functionality directly to the data lake, enabling organizations to finally move away from decades-old data warehouse architectures and their associated inefficiencies. Ideally, open table formats are vendor-agnostic, meaning no single company exerts outsized influence over the project, and the format enjoys broad commercial and Open-Source Software (OSS) support, so data teams can use any tool or engine to work with their data.

Open table formats are a core component of the data lakehouse. Users have a choice of modern table formats, including Apache Iceberg, Delta Lake, and Apache Hudi. Among those options, Iceberg is gaining significant momentum, with nearly 40M downloads. Iceberg has the largest community of contributors, as well as the broadest commercial and OSS support.

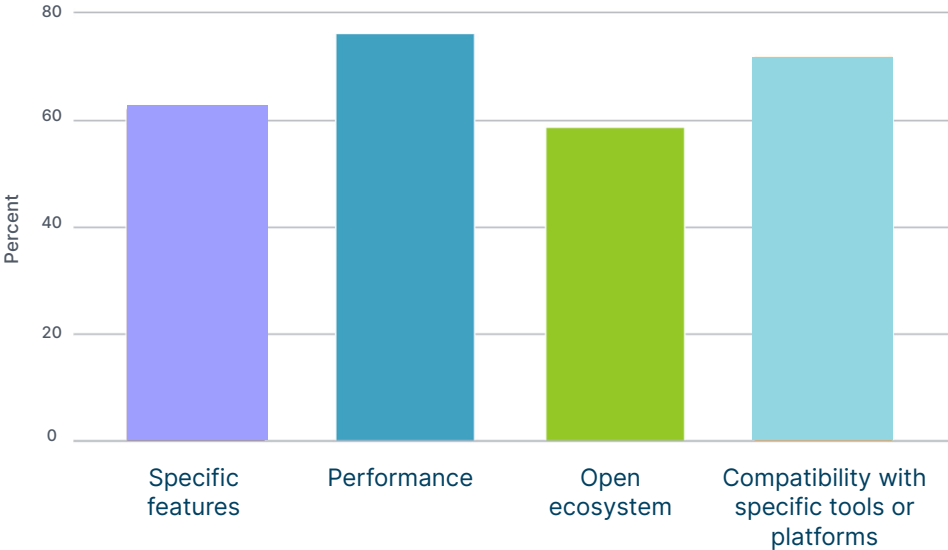
Iceberg and Delta Lake are clearly the leaders, and the survey confirms Iceberg's growing popularity. While 39% of respondents are currently using Delta Lake, compared to 31% who are using Iceberg, 29% adopting an open table format in the next three years plan to choose Iceberg, compared to 23% for Delta Lake.

No, but planning to adopt in the next 3 years



Respondents cited multiple factors that influenced their choice of a particular table format including: performance (77%), compatibility with specific tools or platforms (72%), specific features (62%), and an open ecosystem (59%).

These answers point to continued growth of Iceberg adoption, which was purpose-built for high-performance analytics on massive tables, scaling to tens of petabytes of data. Regardless of dataset size, Iceberg enables sub-second analytics performance while reducing costs with efficient query planning, fewer full-table scans, and easy data operations.



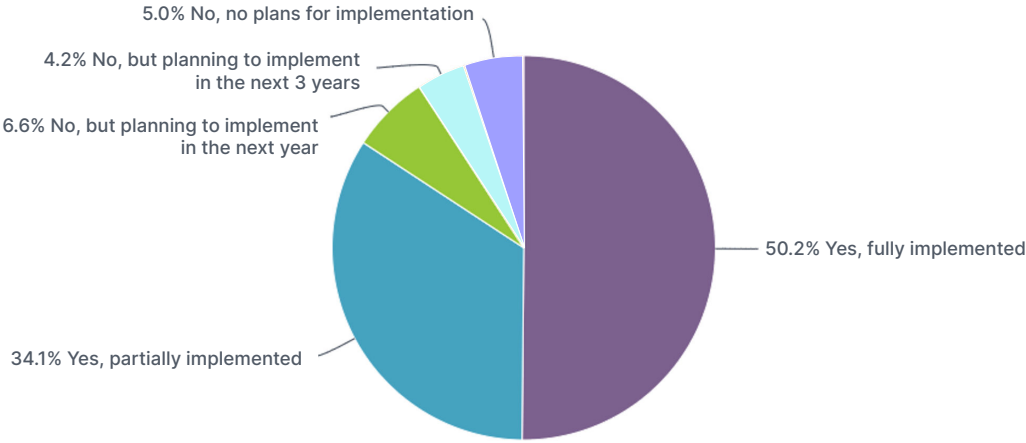
Data Mesh Strategy and a Self-Service Future

What is a data-first organization? A data-driven approach? Data democratization? Businesses know that innovation and market advantage require better, more widespread use of reliable data across an organization. All employees in business units need to be able to safely access the rich, abundant, untapped data in a company’s data lakes, query it and gain insights, then put those insights to work to solve challenges and create new value. This is what digital transformation means today in the data and AI era—and data mesh is the strategy at its heart. A data mesh strategy incorporates self-service tools for analytics, domain-driven data ownership,

the practice of treating data as a dynamic and reusable product, and federated governance that ensures data controls, compliance, and security. The data lakehouse is a primary tool for making the strategy concrete, enabling the shift from theory to practical implementation of data mesh for enterprise organizations. The lakehouse offers self-service tools that help domains build, test, and manage their data products, an intuitive semantic layer that makes it easy to share and access data products across domains, and centralized governance and access controls to ensure data security and safety.

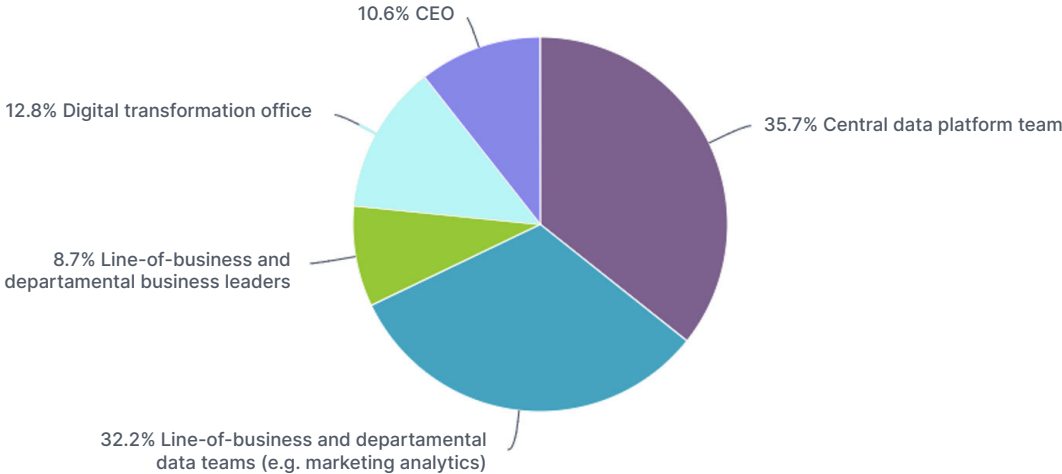
The survey confirmed that data mesh has taken hold in the enterprise: 84% of respondents said they or their organization have fully or partially implemented data mesh. And, 97% expected data mesh implementation to continue to expand in the next year.

Have you or your organization implemented a Data Mesh strategy?



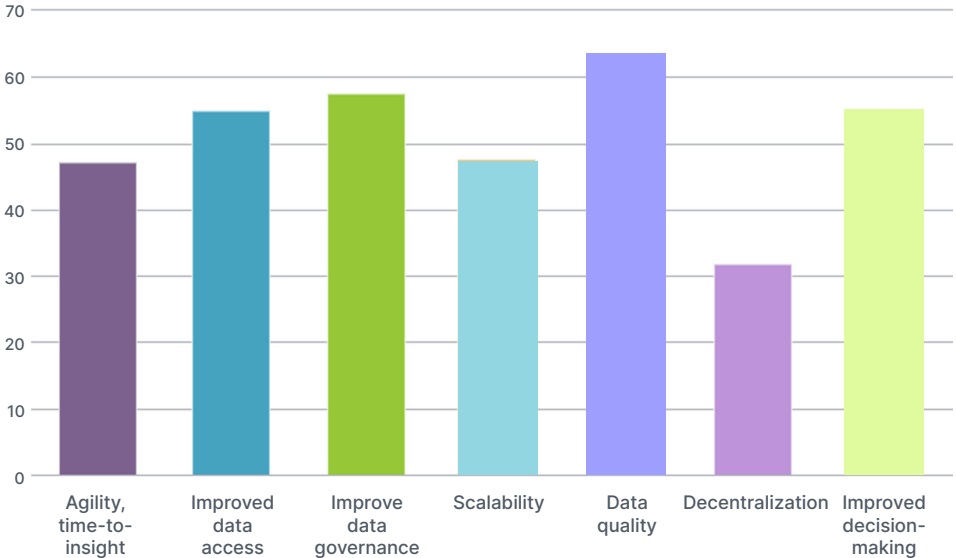
Data mesh is increasingly a business strategy that spurs agility and speed in problem-solving and innovation. The survey found data mesh initiatives are driven more by line-of-business units and business leaders (52%) than by central IT teams.

Who is driving the data mesh initiative in your organization?



Respondents cited a wide range of objectives for implementing data mesh, with the top ones being improved data quality (64%) and data governance (58%). But other factors mattered, too, with nearly half or just over half of respondents also citing improved data access, improved decision-making, scalability and agility:

What is your top objective of implementing a Data Mesh strategy? Please select all that apply.



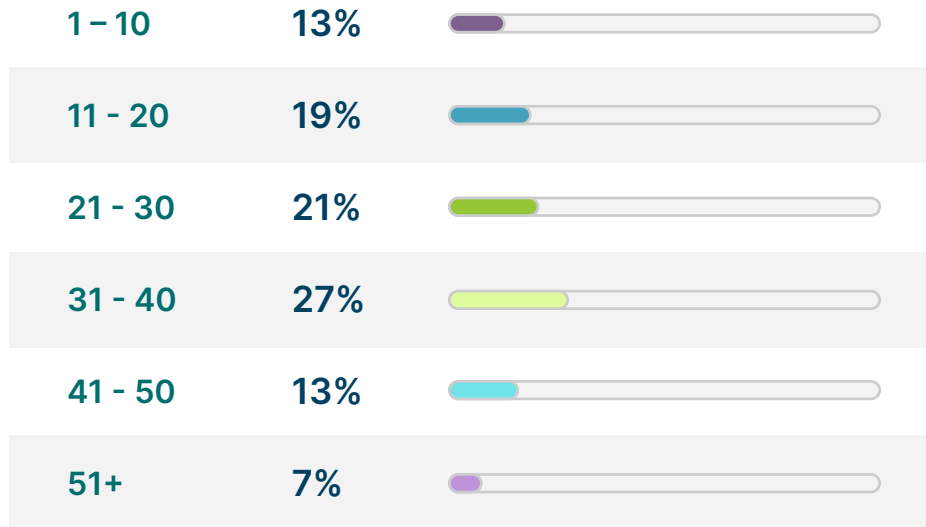
Data Lakehouse for the AI Era

The data mesh and self-service that data lakehouses enable are key to creating value in the AI era. They empower an enterprise’s people, across teams and regions, to work smarter, whether that’s using a natural language-to-SQL generative interface to more easily interact with data, eliminating manual repetitive tasks, or taking action to build a domain-specific AI application. Consider parallels between the rise of widespread AI and the rise of cloud computing. The whole idea with cloud was to counter the hold-ups around application development. We embraced microservices in place of monoliths. We embraced DevOps, not silos. Now

we should ask: what’s being held up around AI creation and decision-making in the AI era?

The good news is that the majority of survey respondents (81%) said their enterprise is already using a data lakehouse to support data scientists in building and improving AI models and applications. And things are moving quickly. Aligned with the AI/ML innovation ongoing across global enterprises and the ramp up of generative AI, 68% of respondents reported that they have more than 20 AI models and AI applications built on top of those models currently in production.

Number of AI models and applications built on top of those models in production at enterprises (all with 1K to 20K+ employees):



AI-driven data management and governance

Data quality is and will continue to be crucial in the AI era, and data governance will be critical in managing AI models and model drift. Recall that survey respondents cited improved data quality (64%) and data governance (58%) as the top priorities for data mesh. The data lakehouse provides a semantic layer that delivers broad access to data while maintaining security and governance, ensuring the good stewardship of customer data.

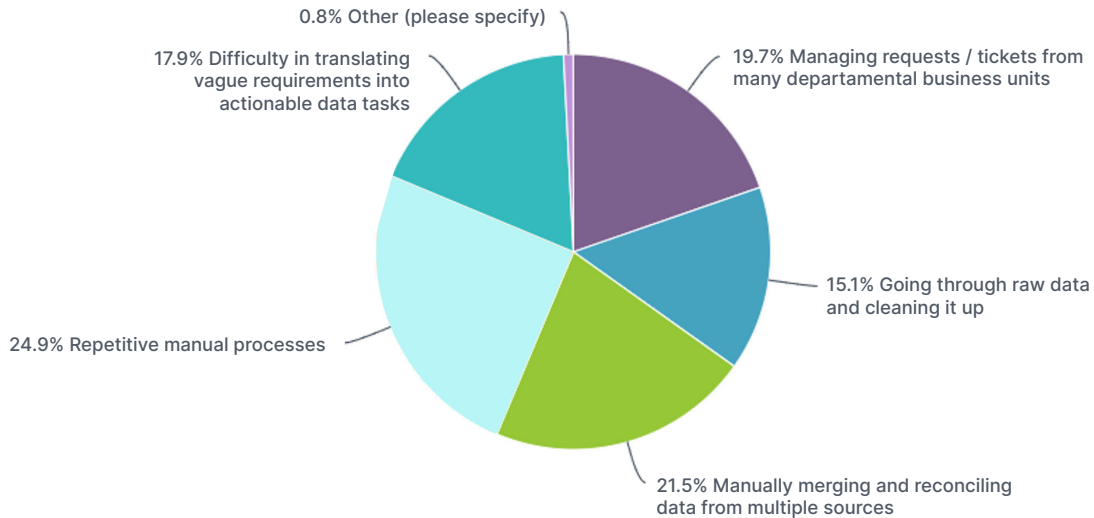
Using a data lakehouse can also improve AI-driven data management and the flexibility and speed of the AI data stack. For example, versioning data without maintaining copies transforms data

management to a new level of efficiency that can improve data security, too.

Moreover, AI and machine learning can be used to lower costs, optimize ETL/ELT, and improve the work lives of IT and data professionals. A majority of survey respondents (62%) cited repetitive manual processes, cleaning up raw data, and manual data merging and reconciliation as problems, speaking to the need for greater automation and AI-assisted data management and governance that the lakehouse offers. Implementing AI not only reduces frustration, but it reduces costs.

The breakdown of least-enjoyable tasks related to data management looked like this:

Which of the following do you LEAST enjoy at your job?



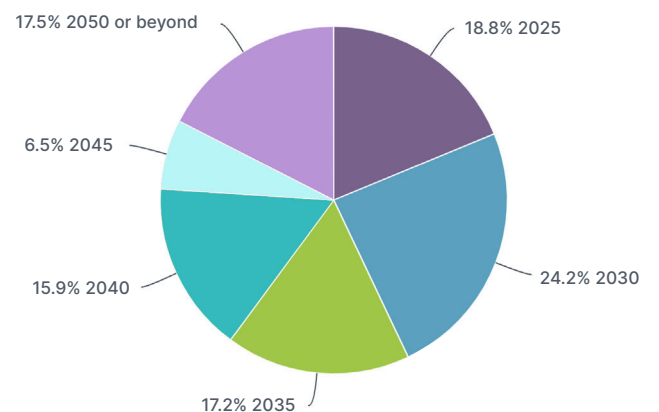
Respondents shared the kinds of data-specific work they enjoyed, too. That included: successfully streamlining and automating complex data workflows (35%), helping other teams access and utilize data seamlessly (17.5%), discovering meaningful patterns and trends in data (17%), predictive modeling that helps anticipate future trends (15.5%), and seeing raw and messy data transformed into clean, usable information (15%).

IT and data pros weigh in on larger AI trends

With the groundbreaking [U.S. executive order on AI](#) soon to impact most aspects of AI development and proliferation, it's useful to understand the perspective of technical professionals. Survey respondents overwhelmingly agreed that AI is a national security priority (84%). Pondering AI's potential use in addressing climate threats and supporting disease research, among other use cases, the vast majority (89%) also believed that AI will ultimately be beneficial for humanity.

Whether and when AI will rival human intelligence were of interest as well: 70% believed AI programs can someday rival the intelligence of every human on earth, and 43% believed that could happen by the end of this decade.

By what year do you believe AI programs can someday rival the intelligence of every human on earth?



About the Survey

A nationwide survey of 500 full-time IT and data technology professionals distributed across industries was conducted by Propeller Insights, sponsored by Dremio, between August and September, 2023. All worked at enterprises with 1,000 or more employees (59% 1K – 5K, 21% 5K – 10K, 7% 10 – 20K, 13% 20K+). Industries represented included: banking, financial services and insurance; health technology; manufacturing; science and technology; high tech; telecommunications and media; retail; education; construction; and other industries. Roles included

IT directors (64%), as well as data and analytics managers and directors, data scientists, software engineers, data analysts, and data engineers.

Propeller Insights is a full-service market research firm based in Los Angeles. Using quantitative and qualitative methodologies to measure and analyze marketplace and consumer opinions, they work extensively across industries such as travel, brand intelligence, entertainment/media, retail, and consumer packaged goods.



About Dremio

Dremio is the easy and open data lakehouse, providing self-service analytics with data warehouse functionality and data lake flexibility across all of your data. Use Dremio's lightning-fast SQL query service and any other processing engine on the same data. Dremio increases agility with a revolutionary data-as-code approach that enables Git-like data experimentation, version control, and governance. In addition, Dremio eliminates data silos by enabling queries across data lakes, databases, and data warehouses, and by simplifying ingestion into the lakehouse. Dremio's fully managed service helps organizations get started with analytics in minutes, and automatically optimizes data for every workload. As the original creator of Apache Arrow and committed to Arrow and Iceberg's community-driven standards, Dremio is on a mission to reinvent SQL for data lakes and meet customers where they are on their lakehouse journey.

Hundreds of global enterprises like JPMorgan Chase, Microsoft, Regeneron, Maersk, and Allianz Global Investors use Dremio to deliver self-service analytics on the data lakehouse. Founded in 2015, Dremio is headquartered in Santa Clara. CNBC recognized Dremio as a [Top Startup for the Enterprise](#) and Deloitte named Dremio to its [2022 Technology Fast 500](#). To learn more, follow the company on [GitHub](#), [LinkedIn](#), [Twitter](#), and [Facebook](#), or visit www.dremio.com.