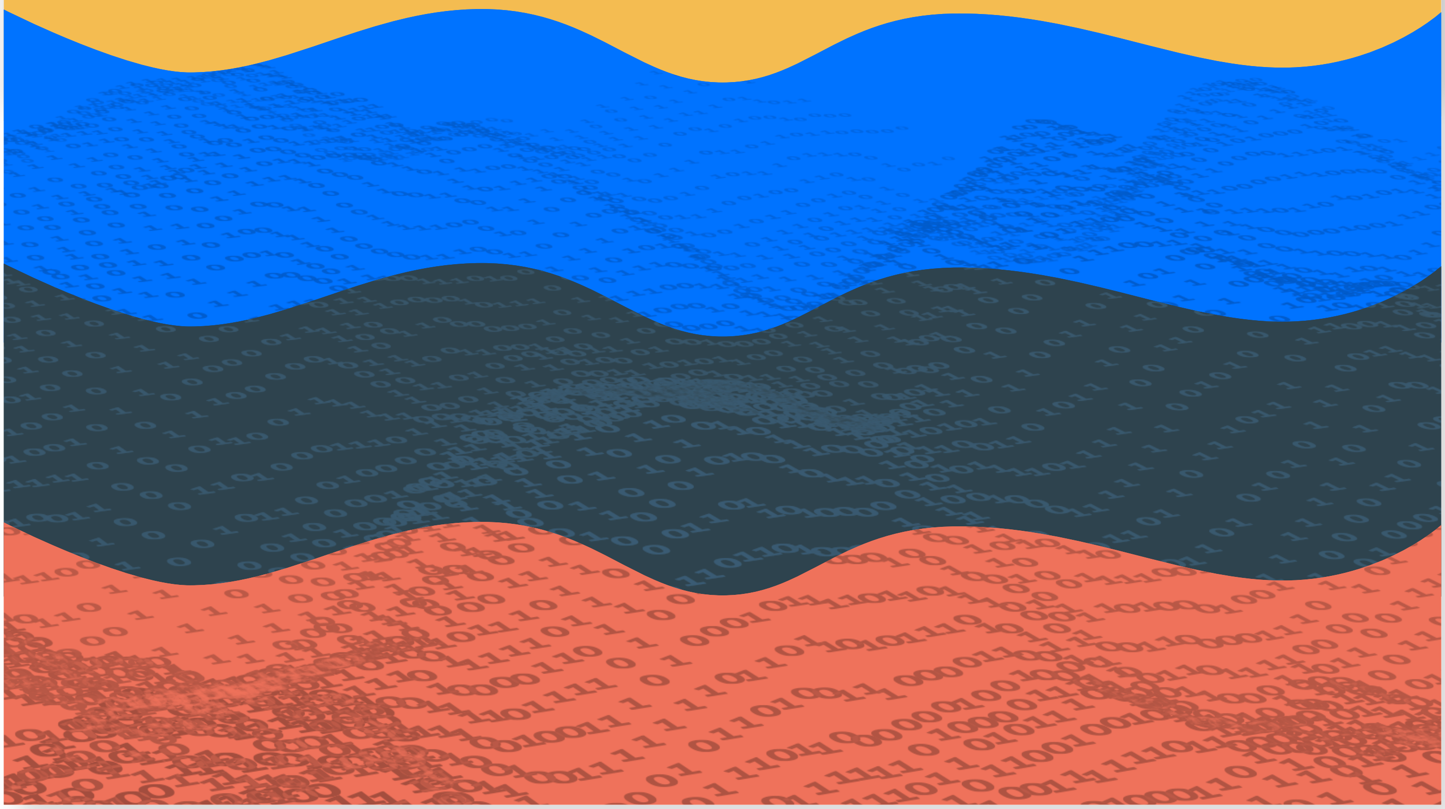
Market Pulse

Building a modern data stack of the future

POWER YOUR DATA MOVEMENT FOR ANALYTICS AND AI ON THE LAKEHOUSE



databricks

Fivetran

SPONSORED CONTENT

FOUNDRY
Formerly IDG Communications

Executive Summary:

Companies must be data-driven in order to remain competitive in today's "new normal." But collecting, managing, and analyzing data is a complex task. Many organizations spend way too much time and money getting data ready for analysis — before a single insight is gleaned. Additionally, learning how to innovate and differentiate with data science and machine learning (ML) is a big challenge because that capability needs to be integrated into the IT stack. The traditional approach of standing up a separate stack for just artificial intelligence (AI) does not work because it's too complex to manage data replication between multiple platforms.

In this study from Databricks and Fivetran, IT leaders provide exclusive insights into their data pain points, how they plan to address them, and what roles they expect cloud and data lakehouses to play in their data stack modernization strategies.

Building a Modern Data Stack of the Future

THE AMOUNT OF DATA COMPANIES PRODUCE DEFIES IMAGINATION — more than 150 zettabytes by some estimates. And that data is transformational: 78% of IT leaders say the collection and analysis of data has the potential to fundamentally change the way their company does business over the next 1-3 years, according to a 2022 Foundry survey.

More than 80% of organizations are using or exploring the use of AI to become a data-driven business. "AI remains a foundational investment in digital transformation projects and programs," says Carl W. Olofson, research vice president with IDC. "The WW Artificial Intelligence spend will exceed $221B by 2025."[1]

But though AI by itself is having a big impact in some areas— enough to account for a significant amount of earnings — AI/ML and analytics insights are not pervasive across the entire organization. Truly data-driven companies are rare.

"Building trust is foundational in how organizations approach creating enterprise intelligence, and trust comes through transparency, provided by intelligence about data," says Stewart Bond, research director, Data Integration and Intelligence Software research at IDC. "Growth in the data integration and intelligence software market is continuing to be driven by the need to activate intelligence about data to gain control of the modern data environment and help organizations compete in a digital-first world."

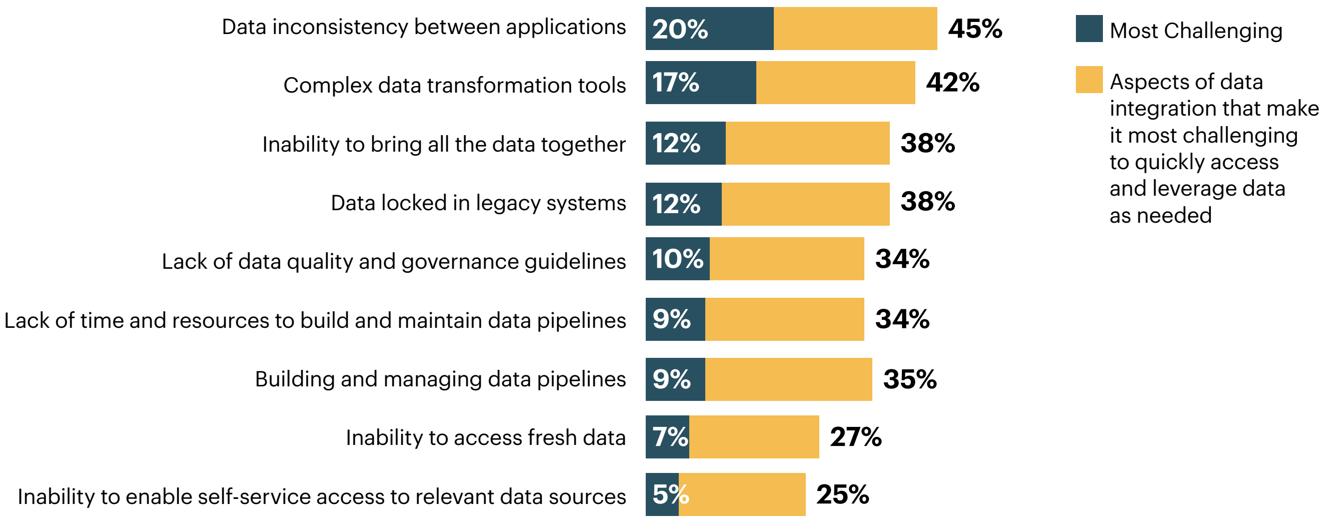
To better understand the state of the data stack in companies around the world, Databricks, Fivetran, and Foundry surveyed 401 senior IT decision-makers and leaders in data analytics/AI roles at large companies around the world. The average number of employees at these respondents' companies was 12,148, and they were located in the United States, the UK and Ireland, Australia and New Zealand, Singapore, and India.

The picture that emerged shows that the enterprise is struggling with data integration and data quality. Data integration challenges are inhibiting progress on AI initiatives at 84% of respondents' organizations. Nearly half (45%) cite data inconsistency between applications as a challenge, and one in five (20%) cite it as their top challenge. See chart on next page.

84%

of respondents' organizations say data integration challenges are inhibiting progress on AI initiatives.

Challenging aspects of data integration



Complexity of the Traditional Data Stack

THIS STATE OF AFFAIRS REGARDING DATA AND DATA ANALYSIS ISN'T SURPRISING. After all, Business intelligence (BI) and AI/ML are often managed in separate data stacks, which introduces an enormous amount of unnecessary complexity. But that's not the only barrier. Issues with data quality and integration are holding back data-driven decisions, the development of a data-driven culture, and the adoption of AI and ML. It's no exaggeration to say that this dependence on siloed, legacy infrastructure is preventing the transition to a fully digital organization.

A modern data infrastructure capable of supporting both BI and AI/ML across the organization must address data integration and quality issues, take advantage of open-source technology when and where appropriate, and place data governance as a top priority.

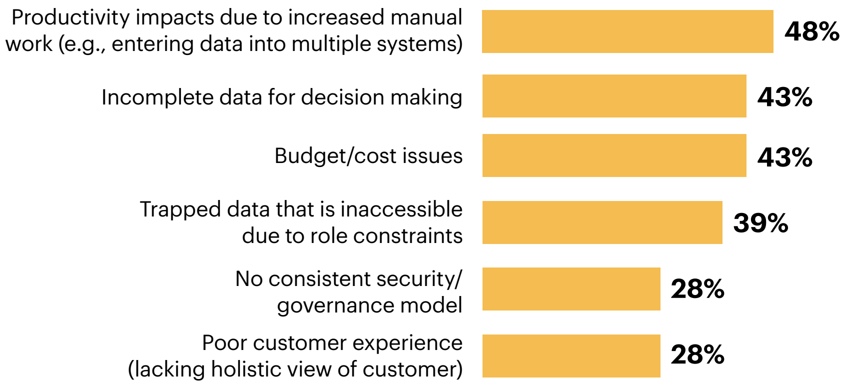
"AI is becoming ubiquitous across all the functional areas of a business," says Ritu Jyoti, program vice president for AI Research at IDC. "Advancements in Machine Learning, Conversational AI, and Computer Vision AI are at the forefront of AI software innovations, architecting converged business and IT process optimizations, predictions and recommendations, and enabling transformative customer and employee experiences."

Indeed, poor data quality isn't just a problem because it's a barrier to AI/ML. It's a serious issue that has negative effects across the company. Data quality is the number one factor impeding data-driven initiatives. Data quality refers to a measure of the accuracy, reliability, relevance, completeness, and consistency of the data to ensure it's fit to serve its purpose. Poor data quality also increases the risk of being out of compliance with government and industry regulations, of course, but there are even greater long-term risks of business damage, including disengaged customers, missed opportunities, brand value erosion, and bad business decisions.

96%

of respondents reported negative business effects as a result of integration challenges.

Business impacts due to data integration challenges



Specifically, nearly all survey respondents (96%) reported negative business effects as a result of integration challenges. More than 40% reported data integration challenges have hurt productivity, data quality, or budgets. See chart above.

Large organizations are increasingly adopting more and more software-as-a-service (SaaS) platforms that generate enormous amounts of valuable data. But modern data teams face daunting integration challenges as the volume and granularity of data continue to grow. With legacy tools and methods, organizations cannot scale with the growth in data sources, nor do they have the resources, time, or budget to spend months developing data pipelines that demand constant maintenance and attention.

On average, respondents are spending 41% of their total time on data analytics projects dedicated to data integration and preparation. But, even with so much effort spent on data integration, more than half still find it challenging to leverage data to provide new products and services to gain a competitive edge (57%), adapt products and services to meet market needs (55%), and deliver data-driven decisions for businesses (52%).

The survey responses show that data integration and data quality significantly affect how much value companies derive from the data they collect. Modernizing analytics environments with a managed data pipeline solution reduces operational risk, ensures high performance, and simplifies ongoing management of data integration.

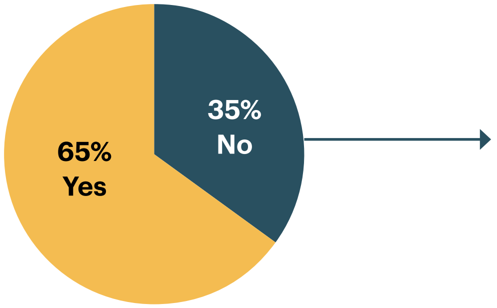
Searching for Modern Data Stack Solutions

Enterprises are taking action to address these data integration and data quality challenges. When asked what are the most important technology changes that will help modernize their data stacks over the next one to two years, the most common answer by far (59%) was data quality tools. Additionally, more than one-third cite open-source technologies (38%), data governance tools (38%), and self-service tools (38%) as critical to their modernization efforts.

If data pipelines are part of the core infrastructure, be sure to keep the following in mind when evaluating tools:

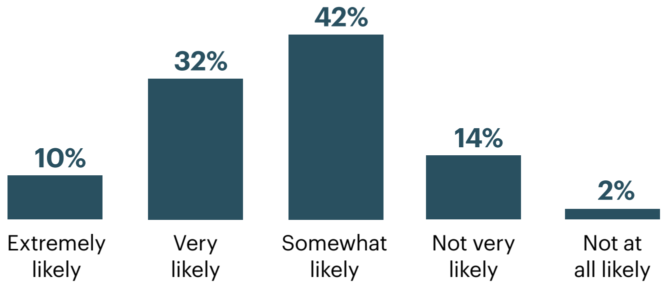
- Whether you want a fully managed or self-managed service
- Product reliability to ensure adequate uptime
- Connector quality to ensure usability of delivered data and access to needed tables
- Customer service and support that provides a timely response and rapid issue resolution
- Product performance capable of handling large volumes of data

Organizations currently using a data lakehouse



Likelihood of an organization considering use of a data lakehouse over the next 12-24 months

(among those not using a data lakehouse)



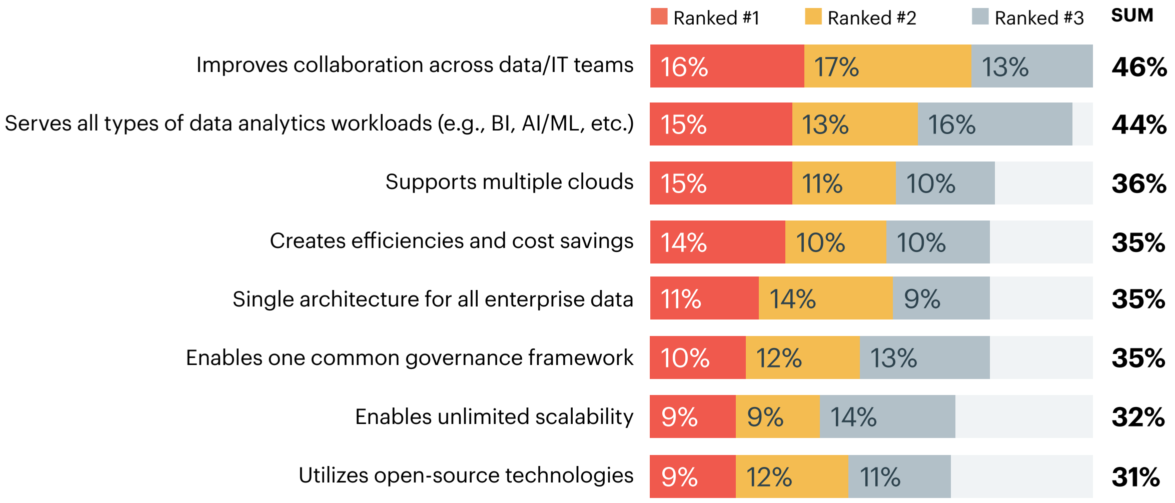
Cloud is also playing a critical role in data stack modernization; the vast majority (71%) of respondents have already adopted it across at least half their data infrastructure. Nearly two-thirds (66%) are also using a data lakehouse, and 84% of those who aren't are likely to consider doing so. See chart above. As for why enterprises are using or are interested in a data lakehouse, the respondents' answers were clear: improved data quality (50%), productivity gains (37%), collaboration (36%) and the elimination of data silos (33%).

84% of respondents who aren't using a data lakehouse are likely to consider doing so.

The Data Lakehouse — A Paradigm Shift

DATA LAKEHOUSES ARE BUILT WITH AN OPEN SYSTEM DESIGN that provides features that are similar to those used in a data warehouse, but they do so using the low-cost, commodity storage that is typically employed for data lakes. Data warehouses have long been used in BI applications, but their ability to support unstructured data or large volumes of data was limited, due to cost and technological constraints. As a result, data lakes emerged to handle many different formats for raw data on inexpensive storage, making it cost-effective for AI/ML use cases. But data lakes have limitations as well — they don't enforce data quality, they cannot support transactions, and it's extremely difficult, if not impossible, for them to support batch and streaming workloads.

Critical capabilities of a modern data technology stack



To overcome these limitations, teams often create complex integrations between the data lake and data warehouses. Stitching the two together can enable data to flow between them. But that often results in greater operational complexity and cost, duplicate data sets, additional infrastructure to manage and maintain, and a much greater attack surface to secure.

A unified data lakehouse provides the strengths of both data lakes and data warehouses without their weaknesses. Data teams can access data without having to work with multiple systems, which increases agility. Additionally, a data lakehouse makes it far simpler to ensure that everyone is working off the complete, most up-to-date data.

Data Lakehouses Come of Age

A NUMBER OF KEY TECHNOLOGICAL ADVANCEMENTS ENABLED THE DEVELOPMENT OF DATA LAKEHOUSES. Metadata layers that sit on top of open file formats increase data quality. Advances in query engine technologies mean that slow performance is no longer an issue. Data can be hot cached in faster storage, such as solid state drives (SSDs). Optimization of data layout to cluster data that's co-accessed and execution can be vectorized with advanced CPUs. The end result is that data lakehouses can provide performance benchmarks that are competitive with data warehouses.

The Databricks and Fivetran Solution

"DATABRICKS INTRODUCED A LAKEHOUSE PLATFORM PRODUCT CALLED DELTA LAKE that provides integration and governance services for data across the data lake and data warehouse, which is meant to ensure not only better data consistency across those systems but improved data processing support with fine-grained access control lists," says IDC's Olofson. "This has set the pace for others seeking to offer data lakehouse services in the cloud."

Here is a snapshot of just a few of the myriad use cases for the combination of Databricks and Fivetran.

- **Marketing analytics:** Allows organizations to integrate data from across the marketing stack to create a holistic view of all marketing channels. As a result, marketers can optimize their spend by comparing campaign performance between platforms and tracking the success of every campaign to gain insights that enable the next campaign to be even more effective.
- **Finance analytics:** Makes it possible to effortlessly transfer data from hundreds of apps into the data lakehouse, enabling the finance team to run rapid queries to create insights that are easily visualized and shared.
- **Sales and customer success analytics:** In just minutes, organizations can centralize all sales and customer data from SaaS and other data sources within Delta Lake to monitor the sales pipeline, predict future needs, and create accurate forecasts using historical revenue, customer feedback, and product data.
- **Database replication:** Move massive amounts of data from Oracle, SQL Server, and other databases, as well as from SAP ECC and S/4HANA with low-impact change data capture (CDC) for real-time data delivery on the lakehouse.
- **Data products:** Software development teams can build analytic applications and portals that provide customers with insights about their data.

Delta Lake replaces data silos with a single home for structured, semi-structured, and unstructured data, providing the foundation for a cost-effective, highly scalable modern data stack.

As a result, organizations can leverage a platform that creates a single source for high-quality, comprehensive, current data. Delta Lake forms the open foundation of the lakehouse by providing reliability and performance directly on data in the data lake. With its support for ACID transactions and schema enforcement, you're able to build a modern data stack that runs AI/ML and BI/analytics workloads directly on the data lake.

Additionally, Delta Lake uses the industry's first open protocol for secure data sharing, which enables different organizations to securely share data, regardless of where the data is stored. As a result, organizations can safely and securely collaborate with partners, suppliers, and other stakeholders on data assets regardless of where the data lives. Additionally, it can all be centrally managed and audited.

The platform also enables control and management across clouds with a fine-grained security model based on the ANSI SQL open standard, making data governance far more effective and efficient.

And because Delta Lake works with existing data, storage, and catalogs, organizations don't need to rip and replace, but can build on current resources while benefiting from a future-proof governance model.

Fivetran completes the picture with modern extraction, loading and transformation (ELT) creating automated, zero-maintenance data pipelines for the data lakehouse. With hundreds of fully managed SaaS and database connectors, Fivetran enables organizations to reliably move and centralize data in near real-time.

Fivetran provides fully managed, prebuilt connectors that automatically adapt as schemas and APIs change, ensuring consistent and reliable access to data. As a key component of the modern data stack, Fivetran continuously synchronizes data from source systems to Delta Lake — eliminating data silos, increasing data reliability and integrity, and improving data team impact.

Fivetran's Business Critical solution provides the highest level of protection for sensitive data, so that enterprise teams can move data securely between the source(s) and Delta Lake — while ensuring internal and regulatory requirements.

AI and ML are must-haves for any modern enterprise looking to compete in today's markets. A modern data stack built to support both traditional analytics/BI and AI/ML use cases is required to leverage these technologies at scale. With a modern data integration solution that securely and rapidly moves data from any source, coupled with a data lakehouse, enterprises can set themselves up for AI/ML success.

For more information, visit databricks.com and fivetran.com.

About Databricks
Databricks is the data and AI company. More than 7,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics, and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn, and Facebook.

About Fivetran
Fivetran is the global leader in modern data integration. Our mission is to make access to data as simple and reliable as electricity. Built for the cloud, Fivetran enables data teams to effortlessly centralize and transform data from hundreds of SaaS and on-premises sources into high-performance cloud destinations. Fast-moving startups to the world's largest companies use Fivetran to accelerate modern analytics and operational efficiency, fueling data-driven business growth. For more information, visit fivetran.com.

[1] Source: "IDC's Worldwide Artificial Intelligence Spending Guide, Feb V1 2022."

[2] 2021 Foundry Data & Analytics Study