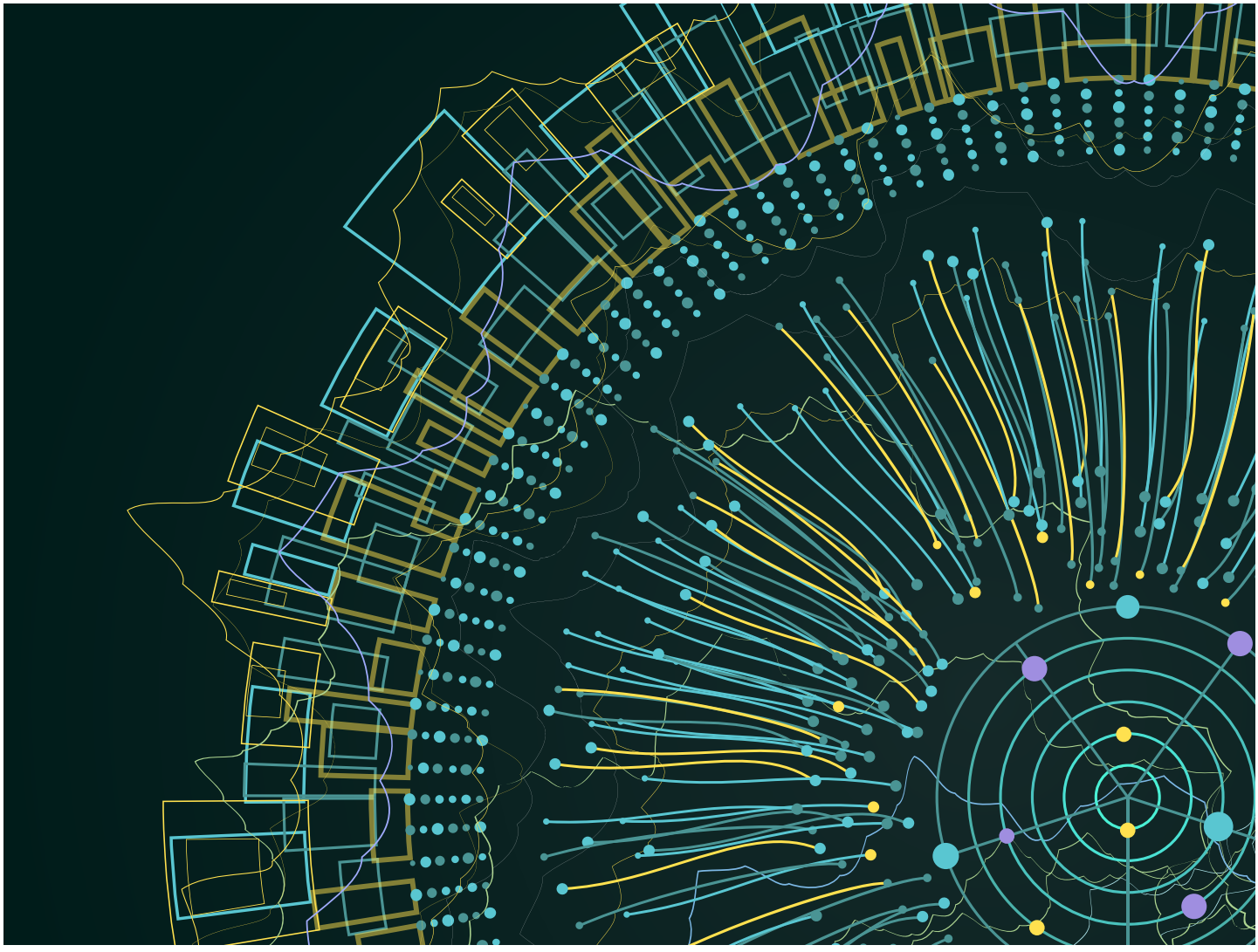
MIT Technology Review Insights

Produced in association with

dremio

Open data lakehouse architectures speed insights and deliver self-service analytics capabilities.

A new paradigm for managing data



Key takeaways

1. Business leaders recognize the imperative to build a data-driven culture, but they are challenged by enormous amounts and varying types of data, as well as their legacy data management systems.
2. Functioning as a single environment to capture all types of data while also enabling business intelligence and analytics, a data lakehouse can be a "best-of-both-worlds" data architecture solution.
3. A data lakehouse unites disparate data types and use cases, providing simple, self-service data access across the organization – while also simplifying IT workloads.

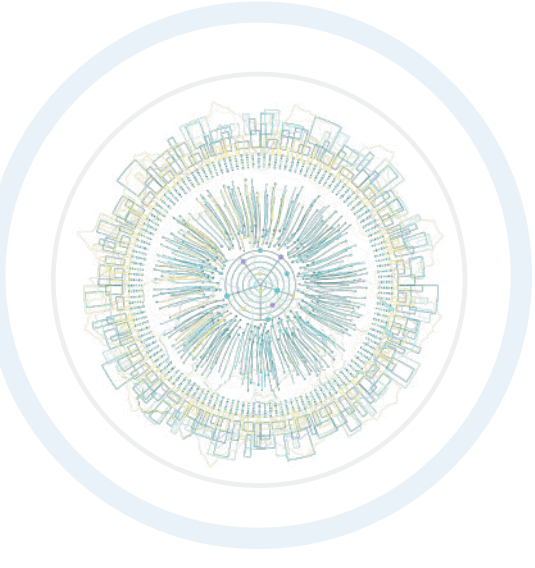
Regeneron Pharmaceuticals, a biotechnology company that develops life-transforming medicines, found itself inundated with vast volumes of data during the peak of the covid-19 pandemic. In order to derive actionable information from these disparate data sets, which ranged from clinical trial data to real-time supply chain information, the company needed new ways to join and relate them, regardless of what format they were in or where they came from.

Shah Nawaz, chief technology officer and vice president of digital technology and engineering at Regeneron, says, "At the time, everybody in the world was reporting on their covid-19 findings from different countries and in different languages." The challenge was how to make sense of these massive data sets in a timely manner, assisting researchers and clinicians, and ultimately getting the best treatments to patients faster. After all, he says, "when you're dealing with large-scale data sets in hundreds, if not thousands, of locations, connecting the dots can be a complex problem."

Regeneron isn't the only company eager to derive more value from its data. Despite the enormous amounts of data they collect and the amount of capital they invest in data management solutions, business leaders are still not benefitting from their data. According to **IDC research**, 83% of CEOs want their organizations to be more data driven, but they struggle with the cultural and technological changes needed to execute an effective data strategy.

In response, many organizations, including Regeneron, are turning to a new form of data architecture as a modern approach to data management. In fact, by 2024, more than three-quarters of current data lake users will be investing in this type of hybrid "data lakehouse" architecture to enhance the value generated from their accumulated data, **according to Matt Aslett**, a research director with Ventana Research.

"Data lakehouse" is the term for a modern, open data architecture that combines the performance and optimization of a data warehouse with the flexibility of a data lake. But achieving the speed, performance, agility, optimization, and governance promised by this technology also requires embracing best practices that prioritize corporate goals and support enterprise-wide collaboration.



"When you're dealing with large-scale data sets in hundreds, if not thousands, of locations, connecting the dots can be a complex problem."

Shah Nawaz, chief technology officer, Regeneron

The evolution of data management

To appreciate the capabilities of the data lakehouse, it's helpful to understand the history of data management. Data warehouses were developed in the 1980s to effectively store structured data from business systems in the data center. But over the last fifteen years, as the volume and variety of data collected by businesses grew, data lakes emerged as a more flexible, scalable, and cost-effective alternative.

Data lakes eliminate the need to store different data formats in different environments, and they can house large volumes of semi-structured and unstructured data. With the right tools in place, data lakes can be used with a variety of analytics products, enabling organizations to glean insights from all types of data. "If the data warehouse is a dedicated environment, the data lake, in theory at least, is more enterprise-wide, with data coming in from multiple sources and used by multiple parts of the organization," says Aslett.

But just like data warehouses, data lakes also have their shortcomings. Designed to accommodate all data formats, data lakes can become disorganized, making it challenging to use them with analytics tools and hard to properly secure and govern them. Additionally, data

A look inside data lakehouse capabilities

At the core of data lakehouse technology are five key features.

Transaction support: Though data lakes work well with unstructured data, they lack data warehouses' support for ACID transactions. ACID refers to the four properties that define a database transaction (atomicity, consistency, isolation, and durability) and ensure data consistency and reliability across the enterprise. Data lakehouses, however, are designed to support ACID transactions, as well as other key features of traditional data warehouse workloads.

Business intelligence and analytics support: It's not enough to simply gather and store data. Organizations must be able to parse that data for meaningful and impactful insights. For this reason, data lakehouse technology can connect directly to business intelligence (BI) dashboards and interactive analytics on data in the data lake, which are difficult to access with traditional data architectures.

Open data architecture: Leveraging open standards, technologies, and formats within the open data lakehouse enables data teams to choose the best tools and execution engines for each analytic workload. This makes data teams more flexible and efficient – and more able to easily adopt the next wave of innovative technologies.

Decoupled storage and compute: The concept of separating compute and storage has been around for years, but next-generation cloud data storage enables the separation of compute and data in a lakehouse architecture, with data being its own tier. Managing storage and compute using separate clusters provides the ability to better manage costs and scale as necessary. It also enables workload isolation, so data consumers don't have to compete for resources.

Governance and security: Given the damaging impact of a data breach, organizations can't afford to risk exposure to bad actors. By serving as a single repository for all of an organization's data, a data lakehouse can simplify security measures and improve data governance. Flexible security controls ensure that data can be safely accessed from data sources across the enterprise, while seamless integration with an IT environment's existing security controls allows organizations to employ their existing security elements, such as authentication and authorization.

lakes struggle with query performance at the scale and concurrency required for enterprise-grade business intelligence (BI) and reporting.

Many organizations have tried to handle these challenges by building out multiple data management systems – one or many data lakes alongside several data warehouses, as well as additional specialized systems, such as image databases. But this approach introduces multiple layers of complexity to the data architecture and creates unnecessary work for data teams.

To leverage data in a data lake for BI and reporting in this type of mixed environment, data teams must build custom pipelines to move and transform data into proprietary data warehouses and properly prepare data sets for analysis. These pipelines are often manual and ad hoc, resulting in "a slow process for companies to work through," according to Tomer Shiran, founder and chief product officer at Dremio, an open data lakehouse provider. Consequently, data engineers become focused on responding to data access requests rather than on more long-term organizational priorities.

Enormous amounts and disparate types of data

Source: The Evolving World of Analytics and Data: Market Insights from Benchmark Research, Ventana Research, 2022

> "It's about having the flexibility and scale of a data lake but also the necessary transactional guarantees and performance of a data warehouse."
>
> Tomer Shiran, founder and chief product officer, Dremio

Enter the data lakehouse – a big data storage architecture that serves as a single repository for all types of data while also enabling business intelligence and analytics capabilities. Rather than integrate or replace data warehouses and data lakes, data lakehouse technology offers "the best of both worlds in one system," says Shiran. "It's about having the flexibility and scale of a data lake but also, in use cases where the data should be structured, the necessary transactional guarantees and performance of a data warehouse."

The result, says Aslett, is a modern and open architecture that "adds structured data management and data processing capabilities, such as analytics, acceleration, and table formats to support consistency. What's more, a data lakehouse enables environments to better support data from multiple sources and to be used for multiple use cases and applications."

Nawaz attests to the value of bringing data together at Regeneron. He explains, "By joining cross-functional domain data, regardless of where it sits, data lakehouse technology enables us to create an entire value chain story from early discovery all the way to commercialization of new products."

From simplification to greater collaboration

This data management architecture can benefit business leaders and IT teams alike. Chief among the advantages is a data lakehouse's ability to deliver secure, self-service data access, liberating data with live, interactive queries directly on Amazon S3, Microsoft Azure Data Lake Storage, HDFS, or another S3 storage solution. The

Top data cultures reap benefits

Organizations with the strongest data culture (top quartile) are

4.6x more likely to use data in major decisions
6.3x more likely to use data in daily meetings
8.1x more likely to use data in approach to work
10.7x more likely to use data to support proposals

Source: "How Data Culture Fuels Business Value in Data-Driven Organizations," IDC Thought Leadership White Paper, 2021

result is greater self-sufficiency and faster insights for data consumers at a time when organizations can't afford to waste time preparing data for analysis.

This architecture provides users with easy access to data for a wide variety of tasks. A marketing manager, for example, may wish to reduce customer churn while a data analytics team leverages data to predict factory maintenance issues. In the case of Regeneron, the company improves the lives of patients by combining data – both structured and unstructured – in a single, centralized repository. "If we're trying to address our patients' needs, we strongly feel that there has to be a connected data ecosystem so that we can respond much quicker," says Nawaz. Whatever the scenario, employees at Regeneron are empowered to discover, curate, analyze, and share datasets with a distinctly self-service mindset.

IT teams also benefit from data lakehouse technology. By simplifying infrastructure, a data lakehouse can significantly ease the burden on time-strapped IT teams. "There are advantages in only having one environment to manage," says Aslett. In fact, he says, not only does data lakehouse technology "consolidate multiple different data spread across the organization," but it can also "consolidate the numerous platforms companies have, reduce data silos, improve knowledge sharing, and enhance information flows."

Another advantage of a data lakehouse is its power to encourage enterprise-wide collaboration. "When all this technology was separate, people and processes were separate," says Shiran. "There was a separate warehouse team, a separate BI team, a separate data science team, and a separate data lake team. However, once you merge technology in a way that works for all these use cases, it can change a company's culture. It's really an opportunity to bring together people and processes."

In fact, "by bringing data together" while encouraging "knowledge sharing across the organization," Aslett says data lakehouse technology can drive innovation, enabling teams "to develop new projects, new initiatives, and new ideas" for a distinctly competitive edge.

Strategies for data management success

To fully realize the benefits of modern data architecture, organizations must establish best practices. One such practice is viewing the modernization of IT infrastructure as not only a technological feat but also as a critical cultural shift.

"It's a people, process, and technology issue," says Shiran. As a result, he says, "Organizations have to embrace cultural changes, especially well-established companies that have legacy systems and IT processes and architecture." Nawaz agrees. Creating a data lakehouse "is a paradigm shift," he says. "We're helping to shape how our organization thinks about analytics as a whole."

For many organizations, this means adopting a data-centric approach to all aspects of the business. "Organizations that are more successful are making

cultural changes to make data the focal point of the organization, in terms of driving the development of new products and initiatives," says Aslett.

But facilitating any type of cultural shift requires C-suite commitment. "Cultural change has to come from the top and it has to be driven by leaders," says Aslett. "A lack of leadership buy-in can be a real impediment to success with data analytics."

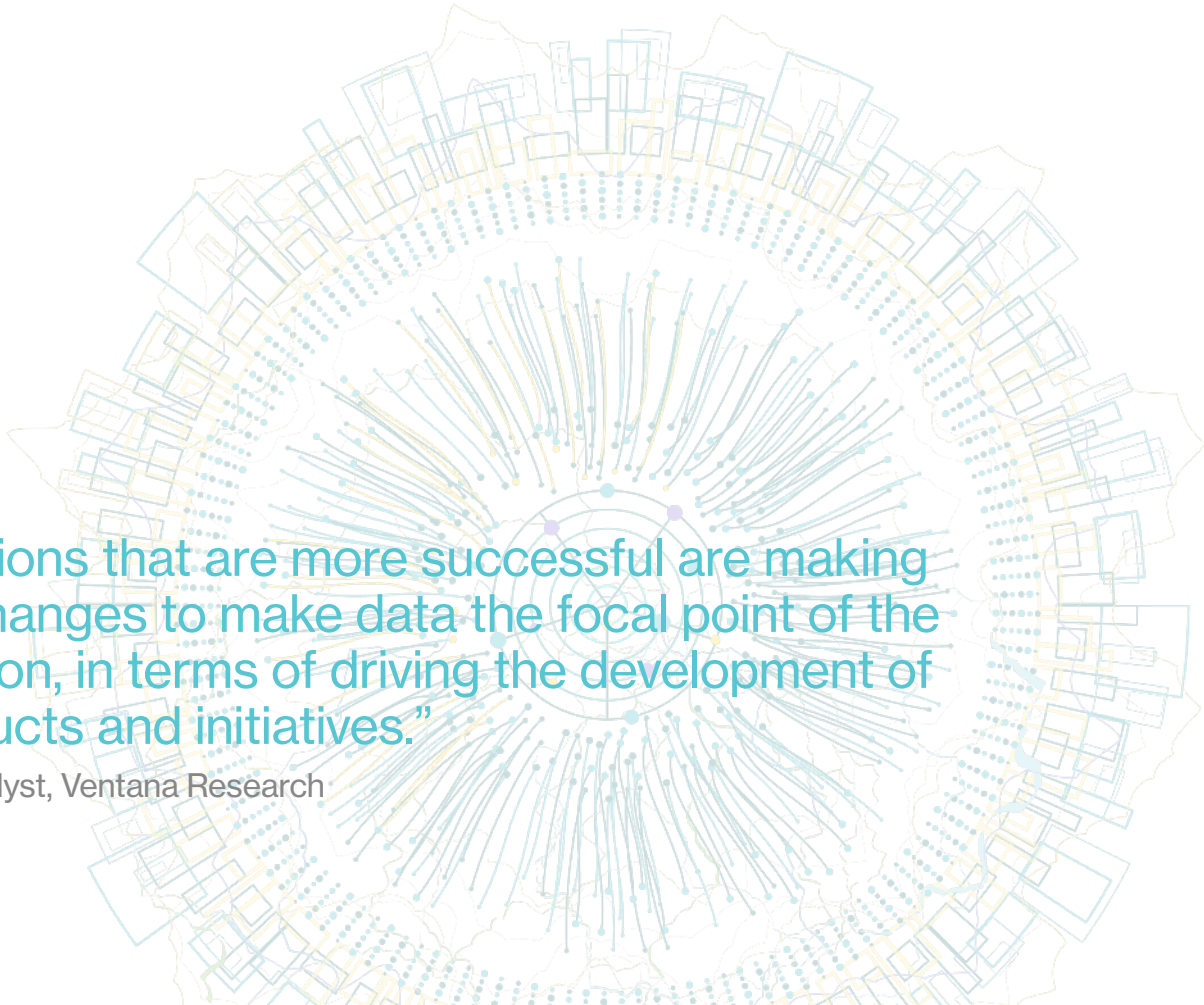
Another essential is developing a deep understanding of how your data architecture will serve your business needs. "People talk about data and it sounds great, but at the end of the day, what's in it for the business? Is it really making an impact?" asks Nawaz.

To answer these important questions, companies must consider what they hope to achieve from their data management efforts. Nawaz says Regeneron invested heavily in the "thought process and design thinking" around building its modern data architecture.

"From the laboratory system to the shop floor system, nearly twenty different systems contain supply chain data," he says. "But how do you streamline supply chain analytics? In order to run the supply chain business, we need to join all these data streams together. That's one area where data lakehouse technology can be applied."

Not only are use cases growing, so too are the potential beneficiaries of this technology. In the future, Aslett says, data lakehouse capabilities will be "more suitable for supporting self-service analytics where business leaders and senior executives will access data themselves rather than have reports and dashboards created for them."

For now, though, data lakehouse technology is helping companies like Regeneron unite disparate data types and a wide variety of workloads in a single, big-data storage solution. After all, says Nawaz, "To respond to patient needs effectively, we believe that all these dots need to be connected."



"Organizations that are more successful are making cultural changes to make data the focal point of the organization, in terms of driving the development of new products and initiatives."

Matt Aslett, analyst, Ventana Research

"A new paradigm for managing data" is an executive briefing paper by MIT Technology Review Insights. We would like to thank all participants as well as the sponsor, Dremio. MIT Technology Review Insights has collected and reported on all findings contained in this paper independently, regardless of participation or sponsorship. Laurel Ruma and Teresa Elsey were the editors of this report, and Nicola Crepaldi was the publisher.

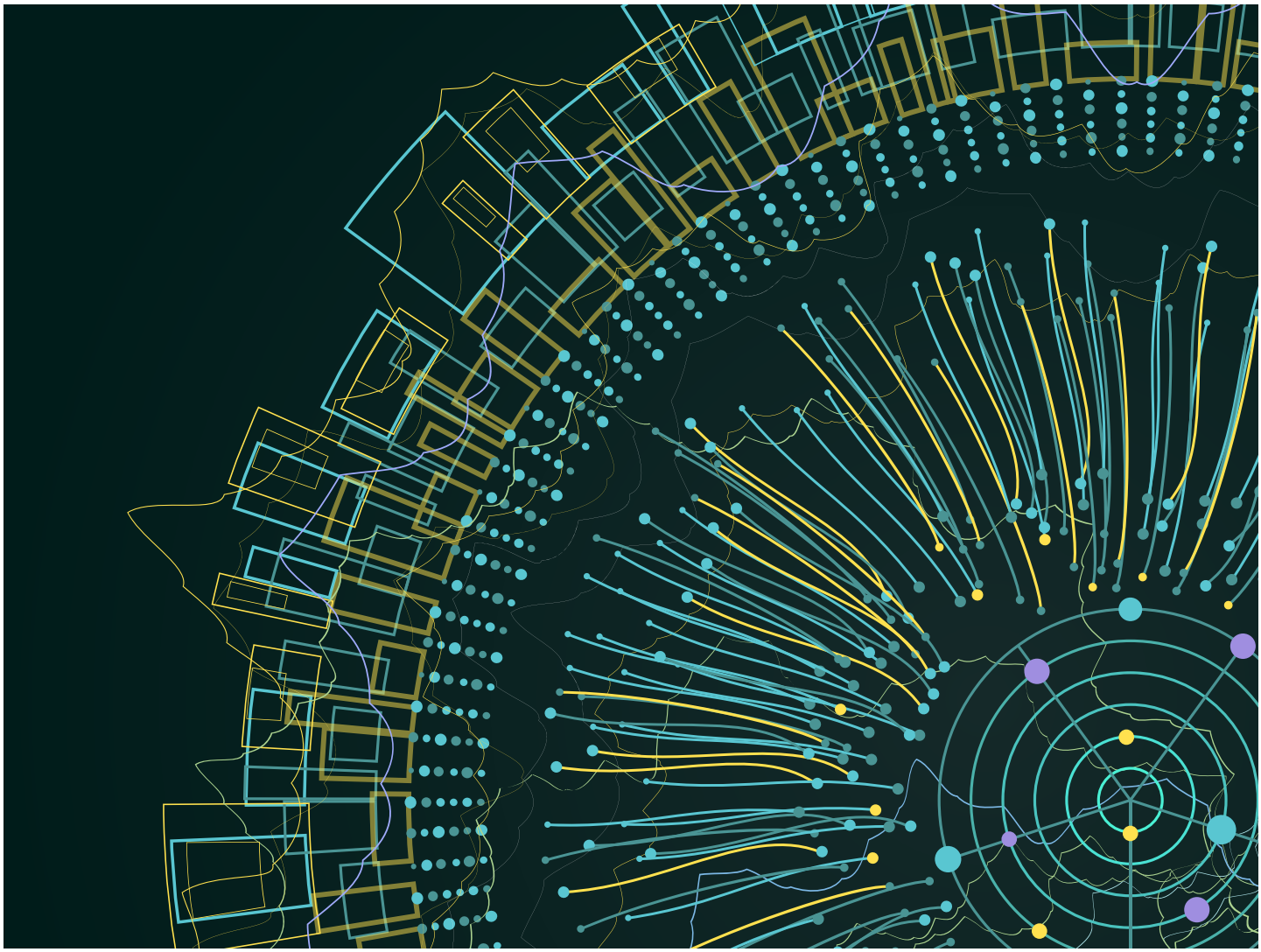
About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of MIT Technology Review, the world's longest-running technology magazine, backed by the world's foremost technology institution – producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the U.S. and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. And through its growing MIT Technology Review Global Insights Panel, Insights has unparalleled access to senior-level executives, innovators, and entrepreneurs worldwide for surveys and in-depth interviews.

From the sponsor

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. 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**.

Illustrations
Cover art by Adobe Stock and spot illustrations created by Chandra Tallman with icons by The Noun Project and Adobe Stock.



MIT Technology Review Insights

www.technologyreview.com

@techreview @mit_insights

insights@technologyreview.com