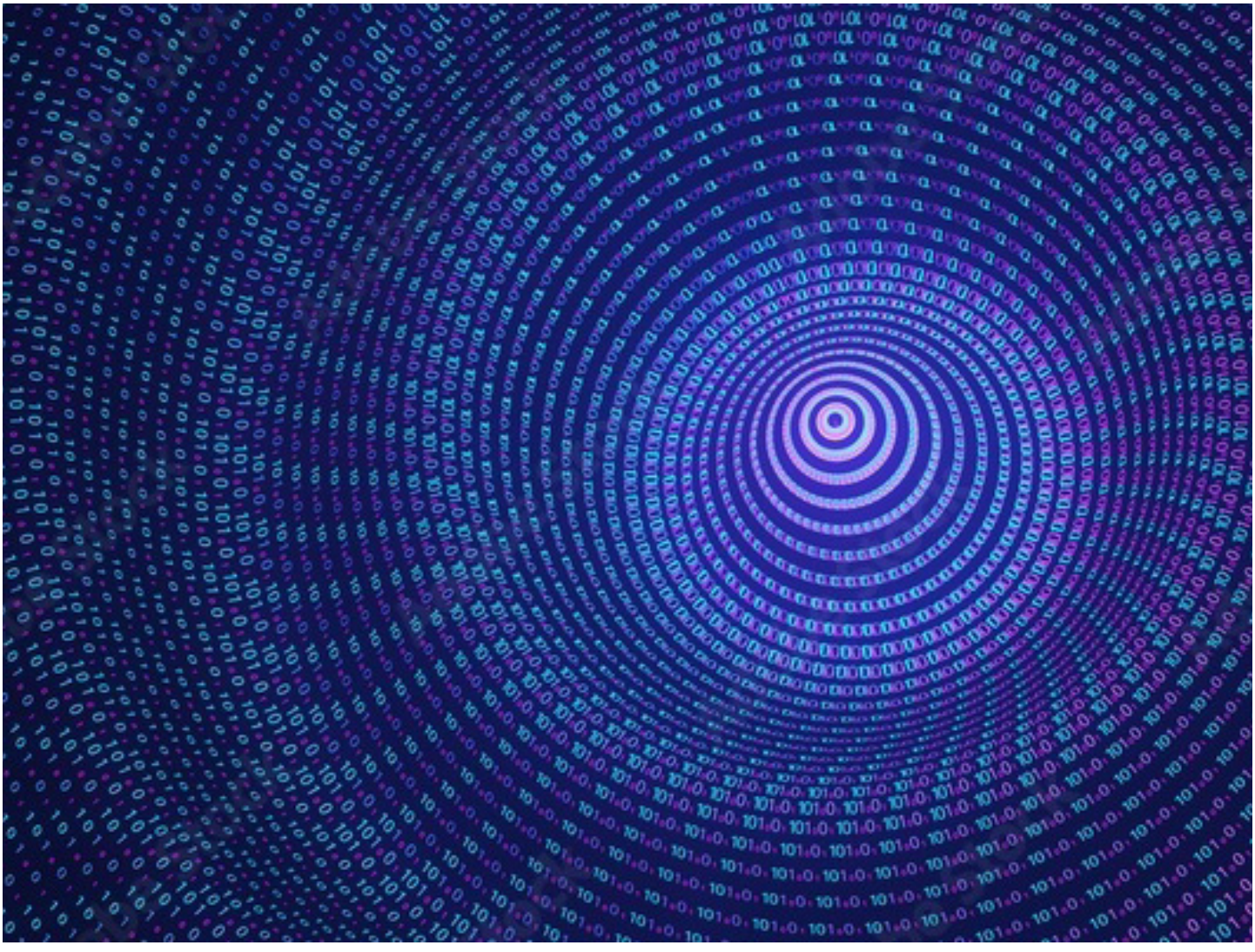
MIT Technology Review
Insights

Produced in partnership with

dremio

Effectively managing and integrating rich data sets can reap business benefits, such as increased productivity and decreased costs.

Optimizing the supply chain with a data lakehouse



Key takeaways

1. **Data generated by the increasingly digitalized supply chain can languish in data lakes, disconnected from the business. As the data deluge continues, this exacerbates complexity and data silos.**
2. **Organizations that effectively manage data operations and integrate these rich sets of data are poised to reap valuable business benefits, such as increased productivity and decreased costs.**
3. **A data lakehouse combines the scale of a data lake with the performance and functionality of a data warehouse. This can help companies unify supply chain data and increase access to data, including structured, semi-structured, and unstructured data.**

When a commercial ship travels from the port of Ras Tanura in Saudi Arabia to Tokyo Bay, it's not only carrying cargo; it's also transporting millions of data points across a wide array of partners and complex technology systems.

Consider, for example, Maersk. The global shipping container and logistics company has more than 100,000 employees, offices in 120 countries, and operates about 800 container ships that can each hold 18,000 tractor-trailer containers. From manufacture to delivery, the items within these containers carry hundreds or thousands of data points, highlighting the amount of supply chain data organizations manage on a daily basis.

Until recently, access to the bulk of an organizations' supply chain data has been limited to specialists, distributed across myriad data systems. Constrained by traditional data warehouse limitations, maintaining the data requires considerable engineering effort; heavy oversight, and substantial financial commitment. Today, a huge amount of data – generated by an increasingly digital supply chain – languishes in data lakes without ever being made available to the business.

A 2023 Boston Consulting Group survey notes that 56% of managers say although investment in modernizing data architectures continues, **managing data operating costs** remains a major pain point. The consultancy also

> "Every single data point is an opportunity for improvement – to improve profitability, knowledge, our ability to price correctly, our ability to staff correctly, and to satisfy the customer."
>
> Mark Sear, Director of AI, Data, and Integration, Maersk

expects data deluge issues are likely to worsen as the volume of data generated grows at a rate of 21% from 2021 to 2024, to 149 zettabytes globally.

"Data is everywhere," says Mark Sear, director of AI, data, and integration at Maersk. "Just consider the life of a product and what goes into transporting a computer mouse from China to the United Kingdom. You have to work out how you get it from the factory to the port, the port to the next port, the port to the warehouse, and the warehouse to the consumer. There are vast amounts of data points throughout that journey."

Sear says organizations that manage to integrate these rich sets of data are poised to reap valuable business benefits. "Every single data point is an opportunity for improvement – to improve profitability, knowledge, our ability to price correctly, our ability to staff correctly, and to satisfy the customer," he says.

Organizations like Maersk are increasingly turning to a data lakehouse architecture. By combining the cost-effective scale of a data lake with the capability and performance of a data warehouse, a data lakehouse promises to help companies unify disparate supply chain data and provide a larger group of users with access to data, including structured, semi-structured, and unstructured data. Building analytics on top of the lakehouse not only allows this new architectural approach to advance supply chain efficiency with better performance and governance, but it can also support easy and immediate data analysis and help reduce operational costs.

The best of both worlds

As data volume grows, so does the preponderance of unstructured data, differing data formats, and the number of data sources, all of which can exacerbate data management complexity and data silos. This is one of the factors driving interest in the data lakehouse. Many organizations are working with "isolated, cloud-based ecosystems that don't enable them to connect data to other parts of the business," says Nick Acheson, chief field data officer for Dremio, a data lakehouse platform provider.

Ideally, organizations need consistent, interconnected, and accurate data across the supply chain for real-time decision making. Retailers need to know what's in their warehouse and must be able to connect this data to

Data as a force multiplier at Maersk

Maersk is a Danish shipping and logistics company founded in 1904. It showed revenue of $51.1 billion in 2023.

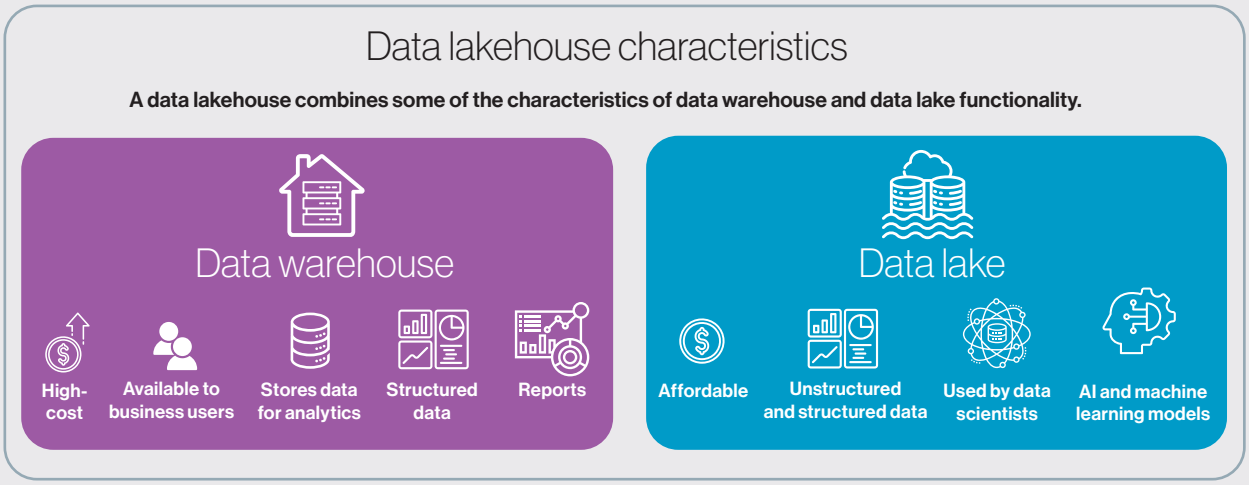
Maersk's largest vessel, a Triple E, of which it has four, can hold 8,000 40-foot shipping containers, some with environmental controls. One ship can move about 64 million boxes of shoes in one trip; typically, though, the containers belong to many different organizations and can carry almost any product. The container contents also have unique origins and destinations, shipping requirements, regulatory rules, and value: But the vessel contains more than just goods, it carries a mountain of data.

To Mark Sear, director of AI, data, and integration at Maersk, this data is an important asset, a force multiplier: a factor that dramatically increases effectiveness. "It's an asset that never depletes, never wears out, and can be used in multiple use cases at zero marginal costs," Sear says.

Making this data effective ensures flexible, fast access to reliable data and helps products arrive on time. Companies should actively avoid spending too much time wrangling data, Sear says. "Data scientists historically, and anybody using data, really, have spent up to 80% of their time cleaning data rather than creating insights," he says. "If you're only spending 20% of your time creating insights, that's a real problem."

For Maersk, Sear says, machine learning—facilitated by data lakehouse technology—is a force multiplier, bringing visibility to customers and supply chains, helping them deal with the complexities. "We also want to make sure that all of our partners are in complete control of their cargo, start to finish, so we can give them that integrated supply chain from the factory right the way through to the end user," Sear said.

Figure 1: What does a data lakehouse do?



Source: Compiled by MIT Technology Review Insights with data from Dremio, 2024

fluctuating product demands to provide supply chain partners with accurate demand forecasts while, at the same time, satisfying customer needs.

"The customer is truly the king of our business," says Sear. "They need to know, 'When is this [product] going to be available to me?' Because there's no point in listing an item on Amazon or saying it will be in my store, or put it on my website, if that ship turns up six weeks late, or if a container arrives at the port just as the ship is leaving. It's about visibility."

Economic realities are also piquing greater interest in lakehouse technology, because of its potential to reduce costs associated with data warehouse scaling and management. "In the world that we're in right now, the compelling economic case for lakehouse architecture is super important," says Acheson.

> "In the world that we're in right now, the compelling economic case for lakehouse architecture is super important."
>
> Nick Acheson, Chief Field Data Officer, Dremio

The benefits of embracing data lakehouse architecture

Organizations must be able to react to emerging issues in near real-time, be it a sudden supply chain disruption or an emerging market trend. Data lakehouse architecture can fuel faster decision-making and greater flexibility so that organizations can access and analyze all of their data from one place.

"The expectation of having real-time data insights is crucial in supply chain scenarios to manage disruptions and be more agile," says Amir Khan, a practice director at Everest Group, a research and advisory firm. "Data lakes consolidate all the data in one place, which is effective for gathering real-time insights."

Visibility means organizations can see where disruptions are arising and react accordingly. For example, in the event of an unanticipated geopolitical disruption, an organization can pinpoint the location of certain containers, and determine how their freezer temperatures must be adjusted to account for any delays.

Another advantage of a data lakehouse architecture is its ability to accelerate AI model development and training to drive new business outcomes. Hidden within today's volumes of data are insights ranging from a company's critical skills gaps to untapped market opportunities. The problem is that 60% to 80% of a

data scientists' time is spent identifying, aggregating, and preparing data for model development and training. A lakehouse can help accelerate access and data wrangling so data can be quickly prepared for AI model development and training.

AI models built using data in the lakehouse can also be used to improve demand forecasting. "Demand forecasting is one of those areas where organizations are using AI/ML technologies on top of data to help them forecast better," says Khan.

Acheson says one example of forecasting could be predicting, from the point of purchase, the time needed to obtain a drum of oil to using that oil to power a ship. A data model can factor-in lead time and lag time, as well as cost volatility, "if the data is consistent and trusted," says Acheson. As a unified platform for data storage and processing, a data lakehouse can help enable data accuracy and consistency across all the nodes of a supply chain.

Enhanced visibility is another advantage of using a data lakehouse to manage supply chain data. Data silos can impede data visibility, making it challenging to see the flow of goods throughout the product lifecycle. A data lakehouse can unify all data to improve visibility so organizations can take preemptive action, from rerouting shipments to avoid delays to identifying signals that might indicate a shift in customer product preference.

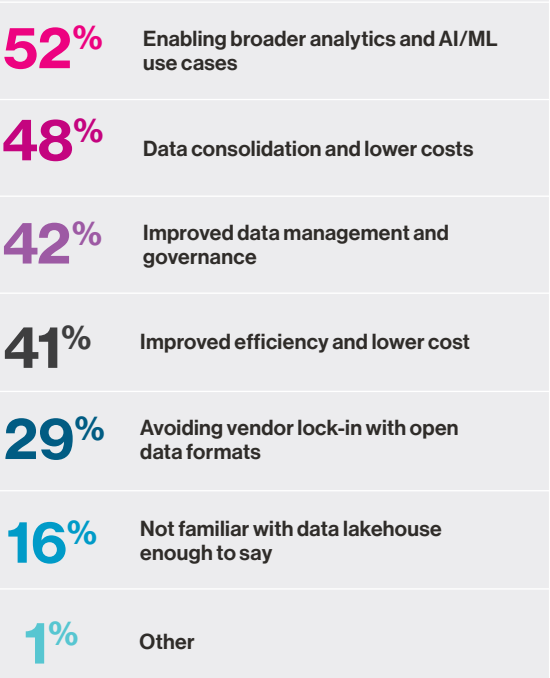
The data lakehouse solution promises to enable the aggregation of data into a single repository with data warehouse capabilities so it can be more easily used for analytics. "When it comes to managing data from different resources, it can be very cost intensive," says Khan. "But whenever you're using a consolidated form of data such as a data lake, it's highly scalable and existing technologies can be integrated into it."

Speed, agility, and cost savings comprise a trifecta of competitive advantages. But business success hinges on more than how quickly a company can respond to change and how effectively it manages costs. Meeting environmental, social, and governance goals, such as decarbonization and net-zero emissions, can have a lasting impact on an organization's ability to retain talent, attract investors, and foster customer loyalty. A growing number of organizations are discovering the unique "opportunity to utilize data lakehouse technology and its capabilities to focus on good," says Acheson.

For example, a transportation company might leverage internal data sets to create a model that shows which routes are most likely to reduce carbon emissions. Similarly, AI models can be used to predict an organization's energy use by tracking the number of truckloads needed to deliver a particular product. Organizations can integrate supply chain partner data with their organizational data into the lakehouse to ensure all parties remain in compliance with global emissions regulations.

"If you're using disparate systems," Khan says, "you have to combine data from all these Tier 1, Tier 2, and Tier 3 suppliers whereas a data lakehouse makes it easier to comprehend all of the data coming in."

Figure 2: Data lakehouse benefits
A survey of 217 data architecture professionals asked them to select which of the following benefits they expected to see from implementing a data lakehouse.



Source: Compiled by MIT Technology Review Insights with data from Market Study: 2023 Modern Data Architecture Trends, Unisphere Research and Radiant Advisors, 2024

Steps to success

A huge amount of data is being generated by an increasingly digitized supply chain. Hidden within this data are insights that can drive favorable business outcomes, from increased employee productivity to reduced costs. But reaching beyond operational gains to realize data's long-term value requires a new architectural approach. A data lakehouse architecture aims to satisfy this demand by providing the functionality and agility of a data warehouse with the scalability of data lakes.

Realizing the business benefits of a data lakehouse requires more than simply consolidating an organization's data centrally, Acheson says. Rather, organizations should consider following some best practices to ensure success.

Ensure data governance: It's essential to make the right data be available to the right users. Balance data access and control to drive data driven decisions while ensuring compliance and data security.

Make managing data projects a shared responsibility: IT and supply chain teams aren't always in agreement about how the lakehouse should serve the business. "The IT team is often more focused on the business metrics and quality of the data whereas the objectives of the supply chain team are more about finding the best business output," says Khan. In this case, dividing responsibility for the success of a data or analytics project equally among parties can promote collaboration and harmony.

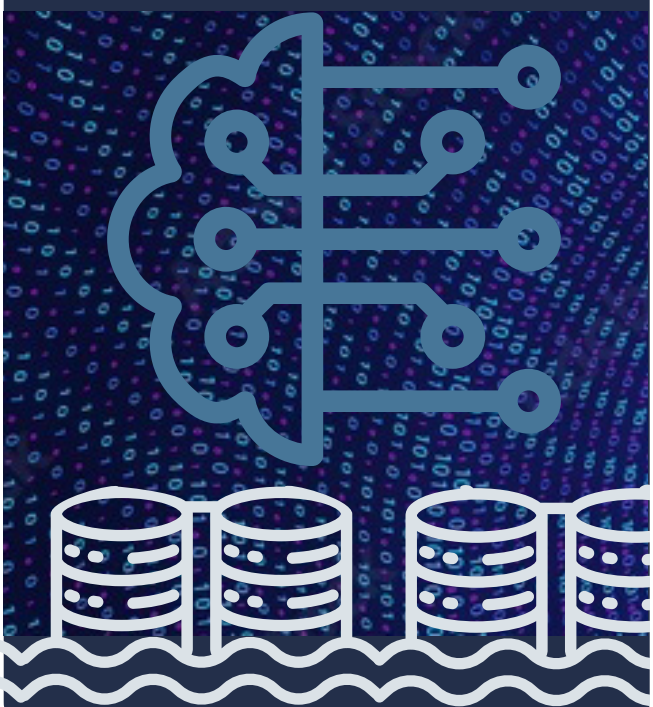
Ensure analytics on the data lakehouse are easily accessible and user friendly: "There's no point in having a data lake if a worker has to be a technical genius to use it, or if you have to know the ins and outs of security to use it," says Sear. "If so, it's not a data lake. It's just a walled garden. You can talk about it, but you're not going to get a lot of use out of it." Interoperability with other systems and self-service analytics and generative AI can prevent such a scenario and drive more consistent and collaborative user experiences.

Generative AI can make lakehouse analytics accessible

"Analysts spend 70% to 80% of their time answering basic questions in the data," says Nick Acheson, chief field data officer at Dremio. The analytics powers that a data lakehouse makes possible has big implications for employee productivity.

"Generative AI allows you to bypass the analyst, bypass the engineer, and quickly get answers to basic questions about the business," Acheson says. In turn, engineers are freed to focus on more critical tasks than responding to data and analytics requests.

Mark Sear, director of AI, data, and integration at Maersk, says the company has ventured into generative AI. "We've built our own generative AI–type product that allows employees to type natural English questions against Dremio," says Sear. "[These] models can predict who is the best salesperson to take on a certain type of account, when are vessels most likely to arrive, what have been the five most profitable routes in the last six months, and how many bookings there were by an operator in 2022."



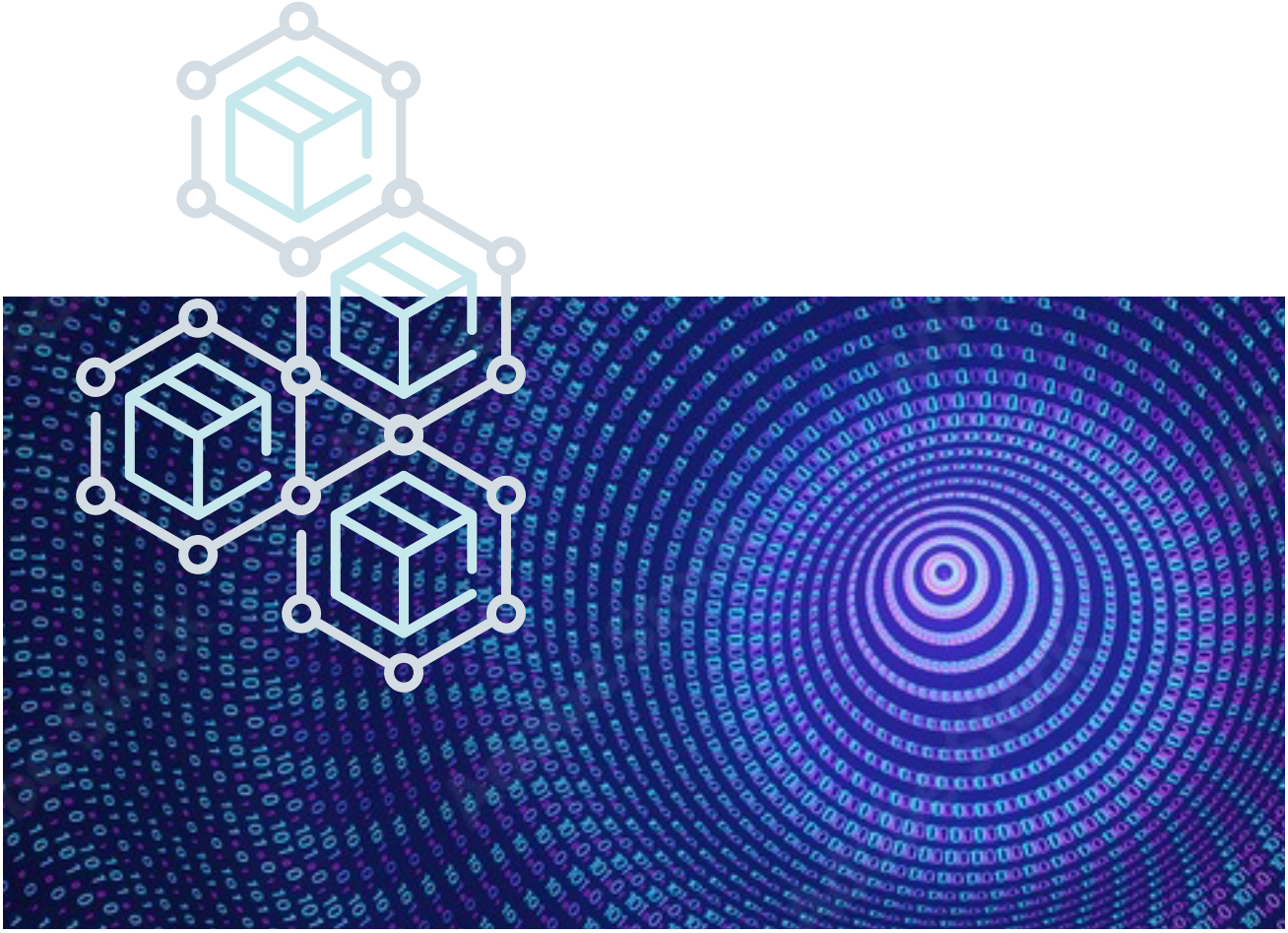
"The expectation of having real-time data insights is crucial in supply chain scenarios to manage disruptions and be more agile."

Amir Khan, Practice Director, Everest Group

Engage change management: Deploying a data lakehouse at a large scale involves change management and raising awareness across the organization. To help drive adoption, Khan recommends that organizations "align teams around a similar objective or outcomes so that they're working towards a common goal."

Carefully select use cases: A solid strategy is critical for data lakehouse success. Organizations must outline the objectives they hope to achieve by consolidating data and aligning data lakehouse capabilities with business goals. "Like any architectural practice, you have to start with some of the key use cases in the business and find a maturity path to be able to bring those assets in and get the benefits of it," says Acheson.

Use AI responsibly: Before companies can begin identifying AI/ML use cases enabled by the data lakehouse, Acheson advises them to first establish policies around responsible AI. "Responsible AI needs to be top of mind for most companies," says Acheson. Only then can companies ensure that their AI deployments empower employees and benefit society and the environment.



"Optimizing supply chain logistics with a data lakehouse" 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. Michelle Brosnahan was the editor of this report, and Nicola Crepaldi was the publisher.

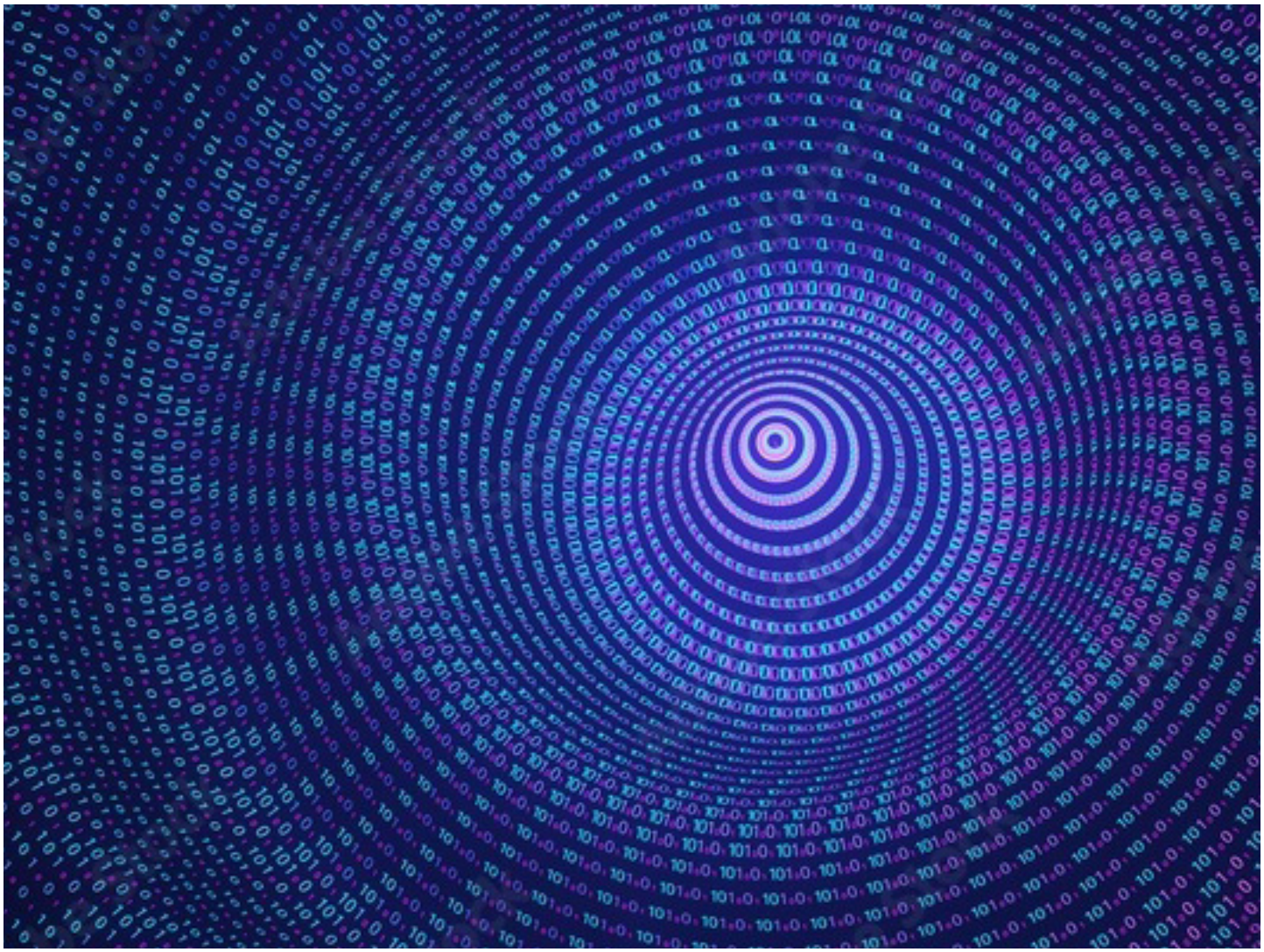
About MIT Technology Review Insights

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 unified lakehouse platform for self-service analytics and AI, serving hundreds of global enterprises, including Maersk, Amazon, Regeneron, NetApp, and S&P Global. Customers rely on Dremio for cloud, hybrid, and on-prem lakehouses to power their data mesh, data warehouse migration, data virtualization, and unified data access use cases. Based on open source technologies, including Apache Iceberg and Apache Arrow, Dremio provides an open lakehouse architecture enabling the fastest time to insight and platform flexibility at a fraction of the cost. Learn more at **www.dremio.com**.

Illustrations
Cover art by Adobe Stock, spot illustrations created by Chandra Tallman Design LLC, compiled from The Noun Project and Adobe Stock.



MIT Technology Review Insights

www.technologyreview.com

insights@technologyreview.com