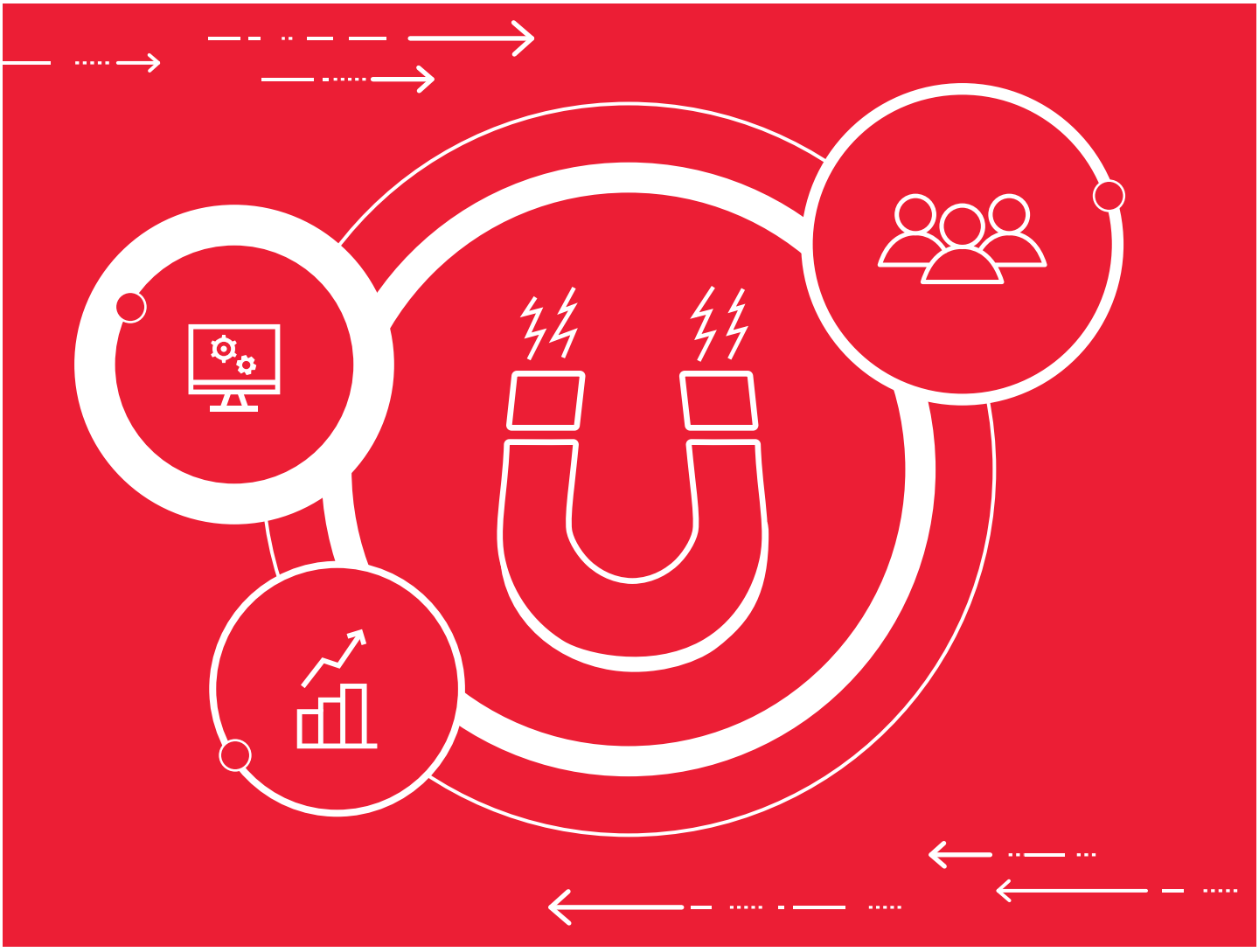
MIT Technology Review
Insights

Produced in association with

HITACHI
Inspire the Next

The practice focuses on collaboration and automation to speed delivery of analytics—and accelerate innovation.

DataOps and the future of data management



Businesses today are facing a mammoth digital challenge – they're striving to find ways to extract value from data. That is, they want to achieve specific business outcomes while the volume and variety of available data are rapidly increasing.

And don't be fooled: Searching for those specks of data gold in a roiling sea of information is no easy feat. It takes the focus and effort of data scientists to do the mining, development teams to craft analytics applications that support smarter decision-making and fuel innovation, data stewards to maintain data quality and security, and IT folks to ensure availability and performance. Different teams and different goals for data means analytics applications often take too long to develop and reach the users who need them. Or they might not do what they're intended to, which is help grow the business.

To make analytics more effective, organizations are replacing traditional data management with an emerging set of practices focused on collaboration and automation. It's called data operations, or DataOps, a confluence of advanced data governance and analytics delivery practices that encompasses the entire data life cycle, from data retrieval and preparation to analysis and reporting. Like DevOps, which aims to speed up software development, DataOps incorporates agile and continuous-delivery development methods supported by on-demand IT resources. DataOps promises to help organizations optimize their data management; drive initiatives involving data-intensive technologies such as artificial intelligence (AI), machine learning (ML), and deep learning (DL); and consistently produce desired business outcomes.

"Finding high-quality, relevant data, in a timely manner, with time to contemplate the meaning of that data will be the challenge," says Michael Daw, law library associate dean and director and professor of law at Golden Gate University School of Law. "Big data vendors, such as LexisNexis and Westlaw, have designed their natural-language searching – their 'smart searching' – to retrieve as much information as possible from as many of their

Key takeaways

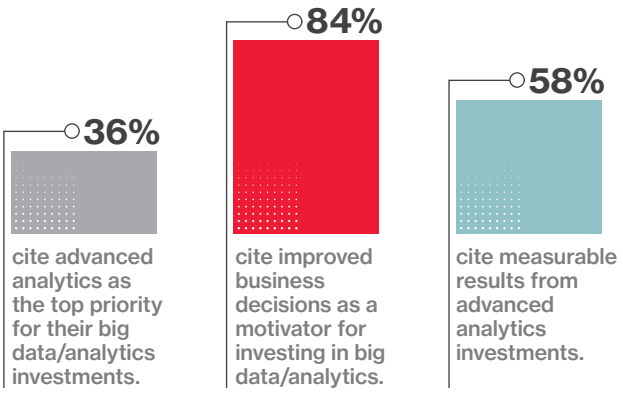
1. Companies are looking to analytics to help them make smarter business decisions, but the ever-increasing volume and variety of data are gumming up the process that delivers applications.

2. DataOps, a set of practices focused on collaboration and automation, is emerging to replace traditional data management and help drive initiatives involving data-intensive technologies such as artificial intelligence.

3. For DataOps to be successful, collaboration is key: Groups that may not interact regularly – for example, business leaders, data scientists, analysts, and IT – must work together.

Setting their sights on analytics

Developing advanced analytics to support improved business decision-making is now a focus of the C-suite.



Source: New Vantage Big Data Executive Survey 2018

thousands of databases as possible, assuming that more data from more sources leads to better choices." But this is not always the case – decisions are often made on highly refined bits of "small data" that is a result of big data analyses.

In fact, established practices no longer do the trick. "Traditional data management practices don't work well in AI," says Lothar Schubert, senior director of technical marketing at Hitachi Vantara. "You need to be able to apply more data with new data feeds, much more frequent updates, and new data pipelines. A more agile approach is needed."

"You need to be able to apply more data with new data feeds, much more frequent updates, and new data pipelines. A more agile approach is needed."

Lothar Schubert, Senior Director of Technical Marketing, Hitachi Vantara

The roots of DataOps

DevOps redefined and optimized the way organizations develop and deliver applications by having development and engineering teams collaborate more closely to accelerate delivery. In the same way, DataOps promises to streamline data management processes to maximize the value of massive corporate data stores. "The principles we've seen with DevOps are also applicable to data management with DataOps," says Schubert. "But they're even more critical since not every organization does software development, but every organization uses data."

DataOps is a process-oriented methodology that improves cross-functional communication, integration, and automation of data flow between various stakeholders, from the network edge to multi-cloud platforms. Andy Palmer, the co-founder and CEO of data management developer Tamr, coined the term in 2015, defining it as a meeting of several aspects of data management: "The framework of tools and culture that allow data engineering organizations to deliver rapid, comprehensive, and curated data to their users ... [it] is the intersection of data engineering, data integration, data quality, and data security."[1]

Gartner defines DataOps similarly as "a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and data consumers across an organization."[2]

DataOps connects data creators with data consumers – both human and machine – to accelerate collaboration and digital innovation, which makes it particularly effective for the enormous volume of high-quality data required for AI processing. "It's critical to ensure data is available at the right quality, to the right person, at the right time, and in the right location," says Schubert. And it needs to be readily usable. Data scientists work with different tools – for example, services such as Google's Cloud Machine Learning Engine or Amazon Web Services' EMR, or locally

Ensuring a successful DataOps practice

To deliver data that will meet the needs of the business, companies must focus on people, process, and technology.

hosted Jupyter notebooks. "Whatever they run, the notion of provisioning data in any platform is something the infrastructure must support."

Forrester Research outlines what it sees as some of the key priorities of a DataOps practice:

- Quickly spin up labs to support insights from enhanced data analysis.
- Develop pipelines and APIs for operational real-time services.
- Develop and deploy architecture from data strategies.
- Federate data governance.
- Tackle data regulatory requirements.[3]

A recently released report from 451 Research states that most enterprises are planning significant DataOps investments in the coming year: 86% of survey respondents planned to increase DataOps investments, and 92% anticipated a positive impact.[4]

Another report, from Experian Data Quality, states that 53% of survey respondents report the lack of data access is the most significant barrier to improved decision-making. That same report also reveals 96% of CDOs believe business leaders demand increased access to data.[5]

Automate and collaborate

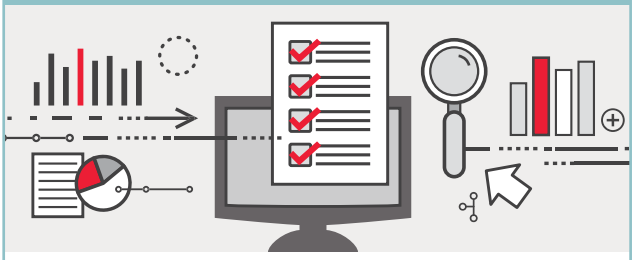
While DataOps is a relatively new discipline, best practices and proven technologies are already emerging. The practice has broad applicability for data analytics, data governance, and operational agility but is particularly well-suited to drive AI, ML, and DL. These data-hungry technologies require incremental improvements to remain effective.

For example, the data models need revisions, and the algorithms need to be retrained with new data – functions for which traditional data management practices are insufficient. Because it streamlines the creation, governance, and operations of vast data supply chains, DataOps is a better method for making quality data available for these processes. "Good data is the foundation for AI," says Schubert. "How these models learn is by data, so having rich data is critical."

As organizations prepare to adopt DataOps, or move deeper into an existing practice, data management and analysis processes must be automated. The data pipelines and ML models need to be continuously updated to ensure successful outcomes. "For automation to work, a focus on metadata is essential – data assets need to be tagged, indexed, and cataloged. A strong data governance function is also necessary to ensure data quality and the right retention practices," says Schubert. "Automation is critical for DataOps."

DataOps connects data creators with data consumers – both human and machine – to accelerate collaboration and digital innovation.

DataOps tech: A to-do list



Tools that help organizations integrate DataOps must do a number of things for adoption of the practice to be successful over the long term.

- ☐ Provide access to multiple data sources.
- ☐ Facilitate collaboration among teams.
- ☐ Support rapid and multiple changes.
- ☐ Allow for benchmarking AI, ML, and DL models.
- ☐ Facilitate monitoring and testing of data management processes.
- ☐ Generate data model and data set versioning.
- ☐ Provide bulk, streaming, and virtual data provisioning.
- ☐ Facilitate data governance to ensure quality, security, privacy, and retention.
- ☐ Help manage and deploy data infrastructure.

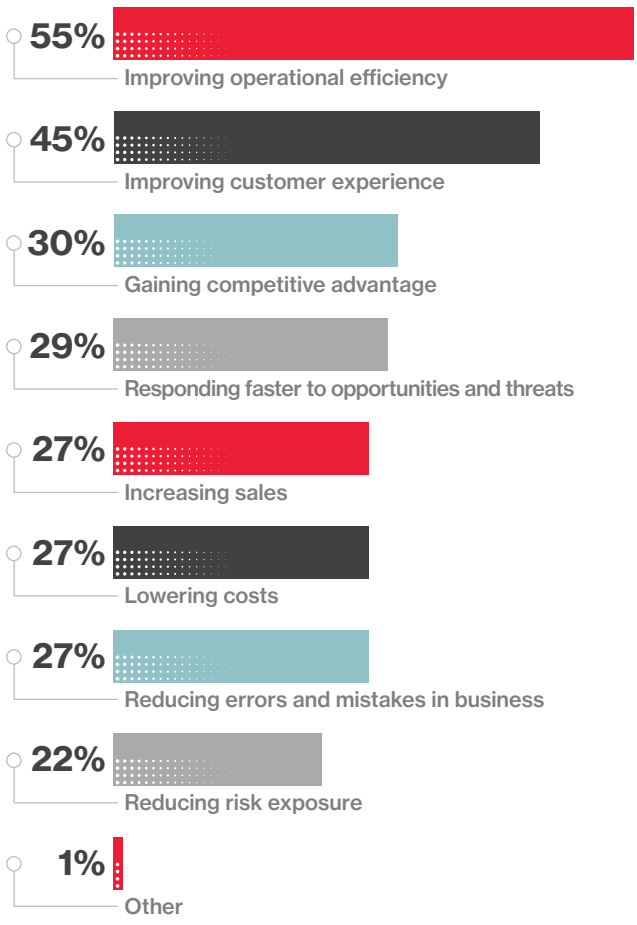
An equally critical part of successful DataOps practices is developing a culture of collaboration. This means bringing groups together that may not interact on a regular basis under traditional data management frameworks. A successful DataOps practice hinges on the teams involved. "It's not just a technology exercise," advises Schubert. "Involve cross-functional teams that include business leaders, data scientists and analysts, data engineers, and IT."

Evolving technical skills and infrastructure capabilities help support this enhanced level of collaboration. "Data engineers and architects must work with the coding and flow management tools to build and deploy data pipelines," Schubert says. "IT and infrastructure management must ensure the infrastructure is available and support automation to spin up clusters for analytic processing."

According to Gartner, growth in the number of data and analytics experts in business units will surpass that of experts in IT departments by 2020, and that will force companies to rethink their organizational models. Also, by 2020, 80% of organizations will educate their workforces in data literacy, "acknowledging their extreme deficiency."[6]

Slicing and dicing data for good

Organizations are seeing a number of business gains from the use of data analytics.



Source: 451 Research's Voice of the Enterprise: Data & Analytics, Workloads & Adoption Patterns 2019

DataOps in action

Some organizations are already employing their data to help increase efficiency and drive advanced functions. For instance, DataOps can help power AI-driven, data-intensive applications such as the internet of things, advanced research efforts, and complex financial analysis.

"Algorithm experts are dreaming up new ideas every day to leverage this new scale [of data]," says William Mayo, CIO at the Broad Institute of MIT and Harvard, which does genomic research on the treatment of human disease. "On one hand, change is simple – more of the same. The problem is that 'the same' now means everything will continue to change, just faster."

The sentiment is echoed at Caterpillar Marine, a division of construction equipment manufacturer Caterpillar that produces vessel propulsion and auxiliary marine engines. There, "more of the same" refers to processing and analyzing an extraordinary amount of data using modern data management practices to help automate shipping fleet maintenance. To do this, the company used predictive analytics on "trillions of data points" to estimate repairs and breakdowns, according to Jim Stascavage, marine asset intelligence technology manager at

> "On one hand, change is simple – more of the same. The problem is that 'the same' now means everything will continue to change, just faster."
>
> William Mayo, CIO, Broad Institute of MIT and Harvard

Caterpillar Marine, in a previously published article.[7] The focus on monetizing data helps reduce shipping fleet downtime, saves time and costs, and most important, keeps the fleet operational.

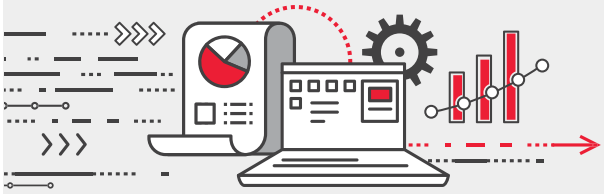
"We identify trends toward failure before they become alerts," says Stascavage. "We're mashing all this data together and trying to figure out what it means for the performance of the ship."

Autonomous vehicle operation is another area poised to benefit from DataOps. All the major auto companies are investing in autonomous driving, and there are petabytes of data that need to be managed. There is telemetry data, image data, and lidar data, all captured from those self-driving cars and traffic infrastructure. "You have to protect and manage that data. The data also needs to be cleansed and enriched to help build a solid AI foundation," says Schubert. "You need to bring all that data together to enrich it with context. Then when data scientists come to use data, they need to be able to find it and apply it effectively to build and test and train models."

DataOps reflects a true evolution of process and philosophy. "To a very real extent, we have inverted the very basis of the scientific method, which was essentially to form a hypothesis, design an experiment and collect data, and analyze, refine, and repeat," says Mayo. "Today, it is more like collect data, ask an algorithm what might be interesting, then react and repeat."

Creating a data-driven culture

Executives expressed a commitment to forging a data-driven culture to avoid disruption and displacement at the hands of more agile competitors.



In 2018,
98.6% of firms
aspired to a data-driven culture –
a 15.3% increase from 2017.

32.4% reported their organizations are **driven by data.**

67.6% said it is **too early** to determine if they will be **successful** at creating a **data-driven culture.**

Source: New Vantage Big Data Executive Survey 2018

Footnotes

1. Kobbi Gal, "DataOps and the Data Platform," Software Engineering Daily, Oct. 18, 2018, https://softwareengineeringdaily.com/2018/10/18/dataops-and-the-data-platform/.
2. Nick Heudecker, "Your Data Culture Is Changing – Do You Need DataOps?" (presented at the Gartner Data & Analytics Summit, Orlando, Florida, March 2019).
3. Goetz, Michele, "A New Age for Data." Forrester Research, January 2019.
4. "DataOps Lays the Foundation for Agility, Security and Transformational Change," 451 Research, January 2019, https://www.delphix.com/white-paper/dataops-451.
5. "The Chief Data Officer: Powering business opportunities with data," Experian Data Quality, June 2017.
6. Heudecker, "Your Data Culture Is Changing – Do You Need DataOps?"
7. "Making the ships run on time with predictive analytics," SiliconAngle, Dec. 8, 2016, https://siliconangle.com/2016/12/08/making-ships-run-time-predictive-analytics/.

DataOps and the future of data management is an executive briefing paper by MIT Technology Review Insights. It is based on research and interviews conducted in June and July 2019. We would like to thank all participants as well as the sponsor, Hitachi Vantara. MIT Technology Review Insights has collected and reported on all findings contained in this paper independently, regardless of participation or sponsorship. Jason Sparapani was the editor of this report, and Nico 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 US and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. And through its growing MIT Technology Review Global Panel, Insights has unparalleled access to senior-level executives, innovators, and thought leaders worldwide for surveys and in-depth interviews.

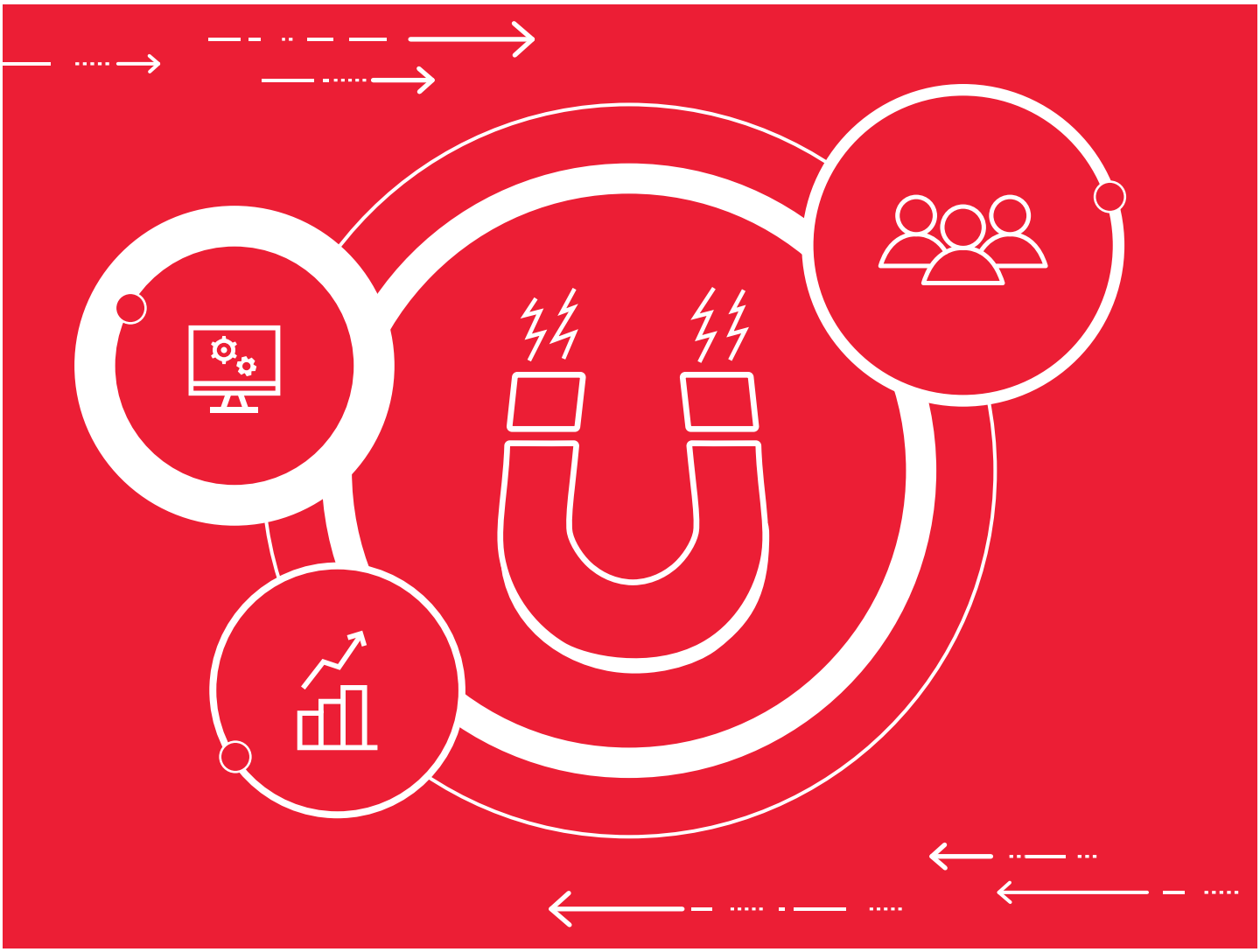
From the sponsor

DataOps is data management for the AI era. It enables organizations to gain frictionless access to their data, allowing them to automate data governance and enhance data pipelines for faster AI. DataOps can make automating data migration from edge to core to cloud easier and can help organizations apply metadata intelligence so they can track, locate, manage and protect it. Hitachi delivers technologies and solutions that accelerate data-driven innovation by getting the right data in the right place at the right time.

Hitachi Vantara, a wholly owned subsidiary of Hitachi, Ltd., helps data-driven leaders find and use the value in their data to innovate intelligently and reach outcomes that matter for business and society – what we call a double bottom line. Only Hitachi Vantara combines over 100 years of experience in operational technology (OT) and more than 60 years in IT to unlock the power of data from your business, your people and your machines. We help enterprises store, enrich, activate and monetize their data to improve their customers' experiences, develop new revenue streams and lower their business costs. Over 80% of the Fortune 100 trust Hitachi Vantara for data solutions.

HITACHI
Inspire the Next

Illustrations
Pages 4 and 6 illustrations by Michele Singh, with icons from Molnia at Adobe Stock.



MIT Technology Review Insights

insights.techreview.com

@techreview @mittr_insights

insights@technologyreview.com