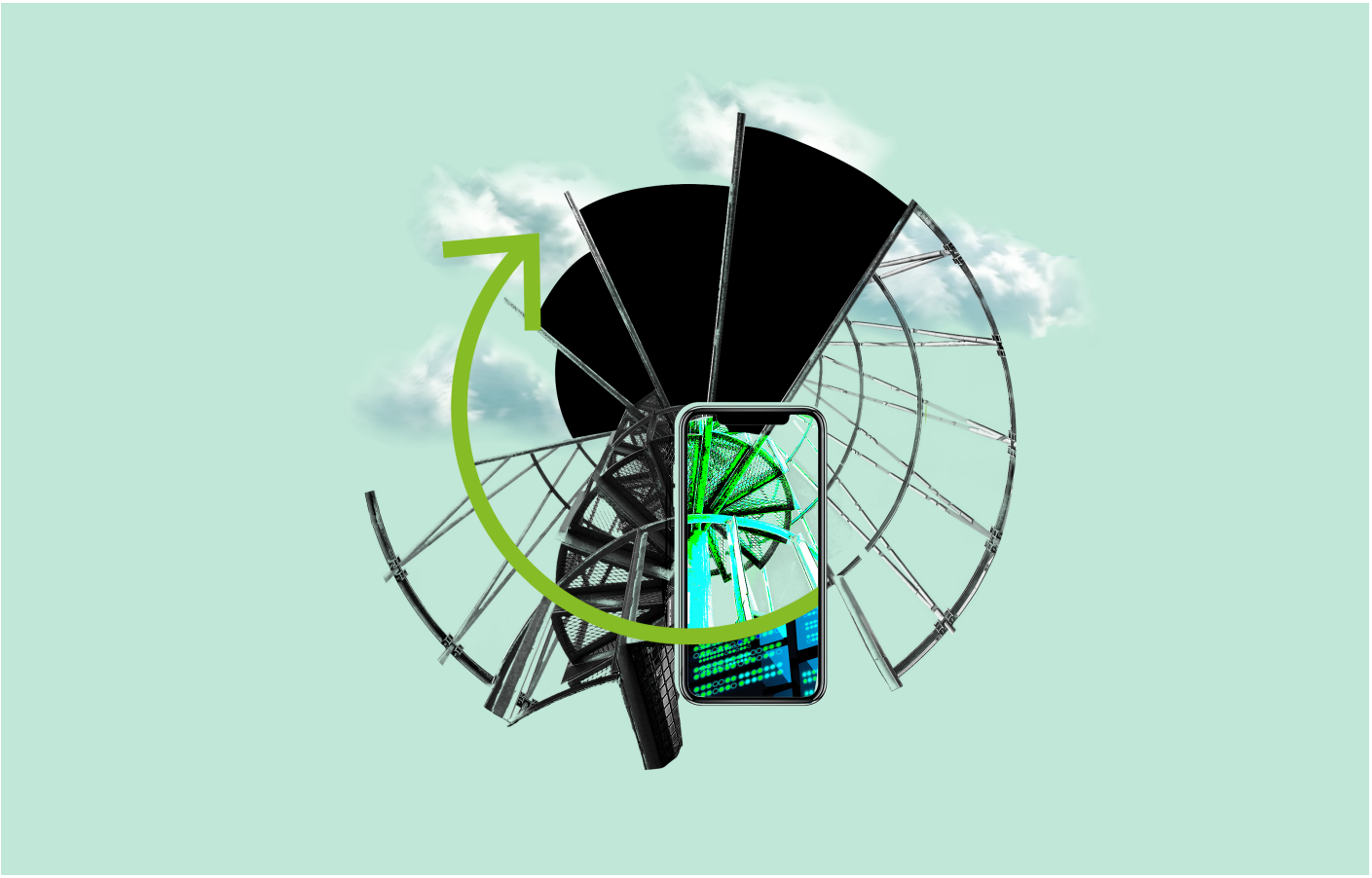
Deloitte. Insights



FEATURE

The edge of cloud

A discussion with Mahadev Satyanarayanan, the Carnegie Group University professor of computer science

Diana Kearns-Manolatos, Myke Miller, David Linthicum

THE DELOITTE CENTER FOR INTEGRATED RESEARCH

Five cloud-edge trends and how organizations can stay ahead of them

IN NOVEMBER 2020, the Deloitte Cloud Institute invited Mahadev Satyanarayanan, the Carnegie Group University Professor of Computer Science at Carnegie Mellon University, as the first Deloitte Cloud Institute Fellow.[1] Dr. Satyanarayanan (Satya as he's popularly known) has made pioneering contributions to advance edge computing, mobile computing, distributed systems, and the Internet of Things (IoT) research focused on performance, scalability, availability, and trust challenges in computing systems.

We examine the present and future of distributed edge-cloud architectures in an exclusive discussion with **Satya**; **David Linthicum**, chief cloud strategy officer, Deloitte Consulting LLP; and **Myke Miller**, dean, Deloitte Cloud Institute. **Diana Kearns-Manolatos** from the Deloitte Center for Integrated Research moderated this discussion. But first, a quick foundation on edge computing and five key insights from the discussion.

On the edge

In 2009, Satya's research introduced the seminal concept of "cloudlets"[2] (a "cloud" *close by* the device). His 2017 article, "The emergence of edge computing,"[3] expounded on the idea and argued that mobile and proximity computing devices (e.g., IoT) demand dispersed computing and a new "edge computing" paradigm—cloudlets could be placed at the internet's edge for high-speed computing, scalable edge analytics, privacy mediation, and failover in the case of a cloud outage. Over the past decade, his research, industry product developments, and business use case implementations[4] have adopted and adapted the cloudlet into what we have come to understand as modern-day "edge computing."

More recently, Satya authored "The computing landscape of the 21st century,"[5] in which he introduced a model to organize the distributed computing universe into four segments: (1) cloud computing, (2) mobility and sensing devices, (3) small, dispersed data centers (i.e., cloudlets), and (4) continuously reporting batteryless sensing platforms.

Five cloud-edge knowledge bytes

Our discussion with Satya delivers insights on five cloud-edge trends and how to stay ahead of them.

- **Evolution**: Boundaries are blurring across the four technical computing architecture tiers (mobility and sensing devices, edge, cloud, and continuously reporting sensing platforms) as cloud and edge technologies evolve. Organizations may need *configuration management and orchestration systems* to manage and move across tiers automatically and dynamically.
- **The intelligent edge**: Companies are increasingly integrating artificial intelligence (AI) and machine learning (ML) into edge use cases, creating new capital and data orchestration requirements. A well-architected four-tier computing architecture model (e.g., mobile, edge, cloud, platforms) can unlock the potential for enhanced data security/privacy/trust, improved latency, and expanded bandwidth.
- **Implementation challenges and solutions**: Smart factories, utilities management, and connected vehicles demonstrate the potential and complexity of mobile, edge, cloud, and continuously reporting sensing platforms. Organizations can partition data and workloads across architecture tiers. This should be done based on current

functional data/workload needs with the flexibility to adjust to future functional requirements and dynamic resource allocation.

- **Cross-tier optimization**: A strong edge business case considers ecosystem computing needs to determine what belongs on premise, on mobile, on the edge, in the cloud, and across sensing platforms. As such, organizations will pay a premium for proximity, latency, bandwidth, performance, scalability, and other technical benefits of edge computing. However, one must factor in human costs to manage the solution while calculating the break-even point.
- **Innovation**: Telecommunications and technology organizations continue to introduce new cloud, edge, and networking solutions while businesses explore edge-native applications. This innovation *further* blurs the boundaries across the four computing architecture tiers. Harness the innovation potential with business cases that require proximity, bandwidth, and latency around a mobile/platform sensing ecosystem.

The following is an edited and condensed transcript of the discussion with Satya, David, and Myke on January 19, 2021, that was aligned to the five cloud-edge themes.

Evolution: To the edge and back

Diana Kearns-Manolatos: As a technology pioneer and innovator who has been at the forefront of distributed computing for the last three decades, what are you seeing next for the edge of cloud?

Satya: We're roughly a decade from when cloud computing started. Then, the cloud was about driving economies of scale to their limit, and we largely achieved that vision with gigantic data centers concentrated in remote places. However, fundamental challenges involving the speed of transmission, round-trip times, and bandwidth remain.

The next chapter in computing is going to be about the creation of this edge computing infrastructure worldwide to enable brand-new, edge-native applications.

The time is ripe for the next inflection point—could we get the kind of compute that cloud computing is able to give but deliver it without those limitations of latency and bandwidth? That's what edge computing is about. Once you think about it this way, it's clear why it's the next step in the evolution.

The next chapter in computing is going to be about the creation of this edge computing infrastructure worldwide to enable brand-new, edge-native applications you simply couldn't create if you only had to rely on the cloud. What is going to be very exciting is how new edge-native applications are going to bring productivity enhancements to industries that have not yet been transformed. For example, there's a whole class of industrial troubleshooting applications, which we can talk more about later.

Diana: In some of your prior research, you've talked about blurring the boundaries between tiers. How do you see the mobile edge, intelligent edge, and cloud tiers functioning to create innovative computing architectures?

Satya: Mobility is a big driver of modern computing and, by its very nature, imposes demands on computing. First, the phone has to fit in your pocket, the virtual reality (VR) glasses on your face, or the wristwatch on your wrist. The device has to be small and light; it cannot get too

hot; and it has to have decent battery life. Let's call these fundamental challenges to mobility the "mobility penalty." It is the price you pay for making things mobile.

My next challenge is: How do I get cloud computing's power and capabilities at the extreme mobile edge to overcome latency (distance) and bandwidth limitations? Edge computing emerges as a solution. It uses what we've learned in cloud computing—managed infrastructure, multitenancy, virtualization, virtual machines, and containers—but instead of putting the data into an enormous, exascale data center, it puts it closer to the mobile devices in a much smaller data center (a data center in a box!), a "cloudlet" that's really a small cloud close by.

The intelligent edge

Diana: What advice do you have for organizations looking to bring AI to the edge of cloud?

Satya (laughs): I'm laughing because I just finished writing a paper[6] on exactly this topic. So, a few tips:

First, the sheer speed with which things are changing in AI is dramatic. You have to pay attention to the algorithms, but by the time you use them, newer algorithms appear. You may have to slightly change how those models work in your system.

Second, because of this rapidly changing frontier, the optimal hardware and software combination is very hard to pin down. For example, the iPhone® 12 has neural engine chips for deep neural network inferencing.[7] If your application is written to use this advantage, your phone can compute in a way not previously possible.

Companies should think about what should run on the mobile device, what should run on a cloudlet, and maybe what should run on the cloud. Nonetheless, these decisions should not be viewed as rigid decisions that are architected permanently into the system, but as decisions that you're making for here and now. The next generation of IT hardware and IT technology might actually require you to rethink that partition.

Dave Linthicum: The big thing that I see evolving is the intelligent edge and the ability to put AI engines in edge-based architectures. A couple of years ago, we worked on a connected bike concept where a hardware device served as a knowledge engine communicating to the cloud.[8] These tiered systems can support big data analysis on the back-end cloud systems and more transactional knowledge, such as the ability to shut the bike off if it was on fire, at the edge. We are partitioning the "brain," but each section has access to other sections. Things are done faster, but each section is still able to operate in an independent mode.

The big thing that I see evolving is the intelligent edge and the ability to put AI engines in edge-based architectures.

This was a more advanced organization. Most companies are still figuring out what an AI engine is, how to leverage the training data, how to be effective with AI meeting data analytics, and how to leverage data virtualization and the operations around it. They're learning how to get their computing infrastructure up, running, and operationalized. The potential is just going to be amazing.

Implementation challenges and solutions

Diana: How are organizations implementing their infrastructure across edge and cloud tiers?

Dave: The challenge that we see organizations running into is architectural. Essentially, they need help with dividing the architecture into tiers. As technologists, we hear questions, such as *Where do you keep the data? Where do you keep the knowledge? What's the most efficient way to do it?* Moving forward, whether it's for robots or edge-based private cloud services, organizations will need to create a configuration management system that can dynamically manage information flow across architecture tiers.

The biggest fear I hear about when I talk to executives moving their organizations to edge computing is not that edge computing can't be done but whether it's able to operate the infrastructure or tiers in a way that allows for dynamic changes. I had a mining company client, and they put out sensors that were satellite connected to gather information around working mines across the Northwest United States. They had the sensors, but they needed to be able to manage the data across the edge and the cloud on demand with the same ease as on a mobile device.

The biggest fear I hear about when I talk to executives moving their organizations to edge computing is not that edge computing can't be done but whether it's able to operate the infrastructure or tiers in a way that allows for dynamic changes.

Diana: How are organizations starting to address these challenges?

Dave: That ability to push software changes remains a challenge, as does tiering of the system. Suddenly, there are 10,000 devices, and managing a large number of dispersed devices can be incredibly complex and exhausting. It requires the ability to have dynamic tiering that leverages different technologies in different ways, with operational configuration management. More vendors will get into this space and make purpose-built tools with AI ops, configuration management, identity access management distributed encryption, and key management systems that work with edge-based systems.

Optimization

Myke: I spend a lot of my time working in energy and resources with organizations like Wichita State University,[9] which has a 60,000 square foot smart building on a smart grid with robotics, AI/ML, and 3D printers. Are there manufacturing use cases where the four-tier architecture is well-suited? Are there economic considerations around smart metering in power and utilities?

Satya: I have different thoughts for the manufacturing and metering use cases. To me, the potential opportunity for the edge-computing architecture is different in different cases.

The robotic factory you described is very interesting, but it is a capital-intensive investment. The robots are expensive, and each production change requires reprogramming and debugging. Therefore, each specification change comes with a certain fixed capital cost for the hardware. For each product change, there is a *nontrivial* cost for the software, testing, and troubleshooting. Today, the break-even point for robotics hardware may well be in hundreds of

thousands, or millions, of units. Regardless, manufacturing is moving toward shorter, faster production runs—such as, producing 1,000 units quickly. At these levels, making a business case for large fixed costs gets harder.

The whole point of the robots was to reduce the people costs, but if I produce 100 widgets, it's going to be faster and cheaper for a human with cognitive guidance to produce 100 units than to program and test those robots for such a small run. The truly unique opportunity for four-tiered edge computing (cloudlets in factories) are those situations where there is a human in the loop and the production is not multimillion-unit volumes but shorter runs, where some kind of cognitive assistance, guidance, automation to detect human error is needed.

Below a certain volume, having the *sensing* done by the system but leaving the *actuation* to the human is more cost-effective, and it aligns with the emerging desire to have short-run, highly agile manufacturing.

For the utilities and data collection, the opportunity for cloudlets and edge computing really is in the privacy space. The reason is increasing consumer concerns about how much data they're exposing and all the *inferencing* that can be done with it. For example, the data collected every day by a smart water meter can reveal when your dishwasher runs. From there, it is a short step to inferring your activity, which is concerning. Edge computing, however, can alleviate some of these concerns by keeping the data in escrow at the edge, giving the consumer the opportunity to receive, for example, a more personalized experience, in exchange for detailed consumption data.

Dave: It is going to be an evolving and learning process on how to convert this to make money. The ability to keep track of the devices out there and their operations, and then to sync with software, security, and governance issues is a problem still to be solved.

To give one example, one of our major retail clients had a recommendation engine that used customers' browser data and behavior to manage demographics targeting and determining logistics to recommend products that have a higher chance of interest. It raised sales by 25%. If they could integrate their mobile, IoT, and edge data from stores and their supply chain, they could slice and dice that data in new and different ways. The challenge right now to bringing in that data is privacy, security, and orchestration management.

Myke: We're evaluating options, such as deploying AWS/Outposts or Azure/Stack for clients and, at this point, they may be cost prohibitive. We're seeing some of the same challenges around 5G. I think economics shouldn't be underestimated when you talk about edge computing.[10] You can't penalize the first few use cases with the whole cost of the infrastructure because it can scale with much smaller incremental costs to accommodate future use cases.

Dave: There has to be a business case. My client, a government contractor, is deploying AWS Outposts because their data standards forbid them from leaving top secret data at their data centers. They would pay a premium for it. Ultimately, with the right business case, edge computing can make businesses more innovative and creative and can be a disruptor. Many businesses just look at the bottom line, and they're missing the core point and end up falling by the wayside.

Innovation: Across technology and industry

TECHNOLOGY INNOVATIONS

Diana: What developments are you seeing with the telecommunications and cloud providers that

are shaping what is possible for this four-tier architecture model?

Satya: I think the big players, the cloud computing players, have all realized that edge computing is not just vapor; it's real. But none of them quite knows yet how to translate that abstract concept into their edge strategies or to determine how similar their edge approach should be to the cloud.

In 2018, Microsoft announced that the intelligent edge and the intelligent cloud both would be equally important.[11] They have followed through on this with the introduction of Azure Edge Zones.[12] They created a brand-new profit and loss group called Azure for Operators, which is basically edge computing. So, not just Microsoft, but many corporate giants, are seeing the significance of edge computing—and they are embracing it.

I think the big players, the cloud computing players, have all realized that edge computing is not just vapor; it's real.

Initially, Amazon started with pure serverless compute for sensing. They now have an edge and fiber play to reduce latency and improve connectivity speeds[13] and continue to expand their large-scale software distribution ecosystem across the Alexa devices and skills network.

Vodafone and AWS are partnering on an innovation program for companies in the United Kingdom and Europe to incentivize the use of edge computing.[14] Additionally, Amazon has announced additional edge innovations with the AWS Snow Family.[15]

From the viewpoint of the developer and the customer, the best scenario would be if they don't know where the computing is taking place. If they can run an application and the edge looks exactly like the cloud, no one needs to change the workflow or software. That should be an enormous source of opportunity.

INDUSTRY INNOVATIONS

Diana: Looking forward, Satya, given your work with the Living Edge Lab and the Open Edge Computing Initiative, you get the opportunity to look at the latest research, technology, and innovations in terms of edge-native applications, platforms, and networks. What are you most excited about that you're seeing?

Satya: The truth of the matter is edge computing *is* more expensive than cloud computing, and it is always going to be. You're never going to be able to hide the additional cost. The only way you can make it profitable is by delivering end value to the customer that *more than makes up for the premium*. If what you're doing is taking what's running in the cloud and just moving it into the edge, the cost of moving it is small, but the value you are giving the end user, customer, is very little.

This is why with edge-native applications starting with the business problem is going to be crucial. Let me give you an example. The class of applications I'm most excited about is what I would call "industrial troubleshooting." When an aircraft that is delayed at the gate costs the airline an estimated $600 per minute. Multiply that cost by the average one-hour delay across all delayed flights, and that's a multimillion-dollar business problem. (Data has shown that in 2019 alone, delays cost airlines US$8.3 billion.[16]) Edge computing and AI with some kind of troubleshooting to find an employee at the gate to minimize the delay and save the airline money and the customer time—that's the kind of thinking that is going to lead to the most exciting changes in the next few years.

Myke: We've talked about manufacturing already where the potential to drive business value is significant and tied to the Department of Defense regulatory compliance requirements, such as NIST SP 800-171. Additionally, in health care, we're starting to see wearables, IoT, robotics, edge computing, and the cloud come together to manage smart beds that automatically adjust when patients need to move based on real-time analytics at the edge while maintaining that data in aggregate in the cloud for larger diagnostic and personalized health care goals.

Endnotes

1. Deloitte, "Deloitte Cloud Institute announces the launch of its *fellowship program*," press release, November 10, 2020.
2. Mahadev Satyanarayanan et al., "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Computing* 8, no. 4 (2009): pp. 14–23.
3. Mahadev Satyanarayanan, "The emergence of edge computing," *Computer* 50, no. 1 (2017): pp. 30–39.
4. Ken Carroll and Mahesh Chandramouli, *Scaling IoT to meet enterprise needs: Balancing edge and cloud computing*, Deloitte Insights, June 20, 2019.
5. Mahadev Satyanarayanan, Wei Gao, and Brandon Lucia,"The computing landscape of the 21st century," *ACM Digital Library, February 2019.*
6. Mahadev Satyanarayanan et al., "The role of edge offload for hardware-accelerated mobile devices," *ACM Digital Library, February 2021.*
7. *The edge of cloud: A discussion with Mahadev Satyanarayanan, the Carnegie Group University professor of computer science*, by the Deloitte Center for Integrated Research, is an independent publication and has not been authorized, sponsored, or otherwise approved by Apple Inc. Apple and iPhone are trademarks of Apple Inc., registered in the United States and other countries.
8. Hope Reese, "How a connected motorcycle could save thousands of lives," TechRepublic, February 24, 2020.
9. "The smart factory @ Wichita," a video collaboration between Deloitte and Wichita State University in their new 60,000-square-foot space that will feature an end-to-end smart production line, space for smart ecosystem sponsors, and experiential labs.
10. Naima Hoque Essing et al., *5G edge as an operations transformation platform: How a 5G edge platform can help leaders across industries transform operations*, Deloitte Insights, February 16, 2021.
11. Microsoft News Center, "Satya Nadella email to employees: Embracing our future: Intelligent cloud and intelligent edge," March 29, 2018.
12. Yousef Khalidi, "Microsoft partners with the industry to unlock new 5G scenarios with Azure Edge Zones," Microsoft Azure, March 31, 2020.
13. Business Wire "AWS announces AWS Wavelength," press release, December 3, 2019.
14. Laura Barber, "AWS and Vodafone business bring edge computing closer to organizations in Europe," AWS, December 1, 2020.
15. Scott Howe, "AWS re:Invent recap: Edge computing innovation with the AWS Snow Family," AWS, December 11, 2020.
16. Federal Aviation Administration, "Cost of delay estimates 2019," July 8, 2020.

Acknowledgments

A special thanks to **Mahadev Satyanarayanan** for sharing his valuable insights. The authors would also like to thank the following individuals for contributing their thoughts and ideas to this article: **Jay Parekh**, **Jack Fritz**, **Dan Littmann**, **Rahul Bajpai**, **Robert Kasegrande**, **Jonathan Holdowsky**, and **Brenna Sniderman**.

About the authors

Myke Miller | mykemiller@deloitte.com

Myke Miller is the dean of Deloitte Consulting LLC's Cloud institute where he draws on more than 20 years of experience to deliver curated and collaborative learning experiences focused on key cloud roles. As managing director for Deloitte's Cloud Engineering Group, he helps clients transform legacy infrastructure into innovative and secure platforms for business growth. He specializes in the energy and resources industry and the power and utilities sector.

Diana Kearns-Manolatos | dkearnsmanolatos@deloitte.com

Diana Kearns-Manolatos is a senior manager in Deloitte's Center for Integrated Research where she analyzes market shifts and emerging trends across industries. Her research focuses on cloud and the *future of workforce*. Additionally, she draws on almost 15 years of award winning marketing communications expertise to align insights with business strategy. She speaks on technology and women in leadership, and holds a bachelor's and master's degree from Fordham University.

David Linthicum | dlinthicum@deloitte.com

As the chief cloud strategy officer for Deloitte Consulting LLP, David Linthicum is responsible for building innovative technologies that help clients operate more efficiently while delivering strategies that enable them to disrupt their markets. He is widely respected as a visionary in cloud computing—he was recently named the number one cloud influencer in a report by Apollo Research. He is a graduate of George Mason University.

Contact us

Our insights can help you take advantage of change. If you're looking for fresh ideas to address your challenges, we should talk.

Industry leadership

David Linthicum
Managing director, chief cloud strategy officer | Deloitte Consulting LLP
+ 1 703 216 6676 | dlinthicum@deloitte.com

David Linthicum is the chief cloud strategy officer for Deloitte Consulting LLP and responsible for building innovative technologies that help clients operate more efficiently while delivering strategies that enable them to disrupt their markets.

Myke Miller
Managing director, cloud engineering | Dean of Deloitte's Cloud Institute
+ 1 612 599 4267 | mykemiller@deloitte.com

Myke Miller is the dean of Deloitte's Cloud institute where he delivers innovative, curated, and collaborative learning experiences, focused on key cloud roles to differentiate Deloitte's workforce in the cloud era.

The Deloitte Center for Integrated Research

Diana M. Kearns-Manolatos
Senior manager, subject matter specialist | Deloitte Services LP
+ 1 212 436 3301 | dkearnsmanolatos@deloitte.com

Diana Kearns-Manolatos is a senior manager with Deloitte Services LP's Center for Integrated Research where she focuses on cloud eminence and leads the annual Deloitte *MIT Sloan Management Review* Future of Workforce study.

About the Deloitte Center for Integrated Research

Deloitte's Center for Integrated Research focuses on developing fresh perspectives on critical business issues that cut across industries and functions, from the rapid change of emerging technologies to the consistent factor of human behavior. We look at transformative topics in new ways, delivering new thinking in a variety of formats, such as research articles, short videos, in-person workshops, and online courses.

Connect

To learn more about the vision of the Deloitte Center for Integrated Research, its solutions, thought leadership, and events, please visit www.deloitte.com/us/cir.

The Deloitte Cloud Institute

While cloud is a pathway to what's possible for everyone, it's also complex and rapidly changing. At Cloud Institute, we are building the industry's most skilled workforce across the entire cloud and technology life cycle—to outpace change, navigate complexity, and help our clients, and society, truly realize cloud's potential with courage and imagination. Through a curriculum designed to support and encourage lifelong learning, practitioners at any level can upskill with hands-on courses, deep training, and learning pathways from strategy to human capital, and everything in between. Cloud Institute empowers our people to discover their own possible—and that means better outcomes and new value for our clients.

Deloitte.
Insights

Sign up for Deloitte Insights updates at www.deloitte.com/insights.

Follow @DeloitteInsight

Deloitte Insights contributors
Editorial: Kavita Saini, Aparna Prusty, and Nairita Gangopadhyay
Creative: Jagan Mohan and Sonya Vasilieff
Promotion: Hannah Rapp
Cover artwork: Jaime Austin

About Deloitte Insights

Deloitte Insights publishes original articles, reports and periodicals that provide insights for businesses, the public sector and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte Insights is an imprint of Deloitte Development LLC.

About this publication

About Deloitte