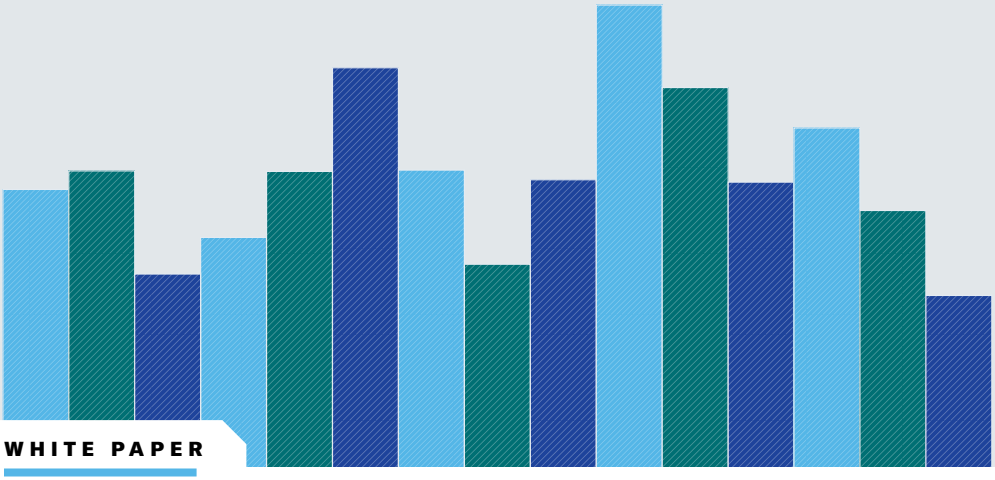
Harvard Business Review
ANALYTIC SERVICES



WHITE PAPER

Rethinking Cloud Strategies for Advanced AI

Sponsored by

Microsoft Azure NVIDIA

SPONSOR PERSPECTIVE

Generative AI is changing how businesses operate in a fast-moving world. AI amplifies human ingenuity, transforming business workflows to boost employee productivity and wellbeing, revolutionizing customer interactions with enhanced self-service options and personalized engagement, and speeding up innovation by equipping R&D and engineering teams with self-learning tools to continuously improve. AI can help researchers and health care professionals develop new therapies and personalized care plans for patients. It enhances manufacturing and supply chain operations with process automation and customization, improves coding and design efficiency in software development, and more. As generative AI continues to shape the business landscape, companies must ensure they are taking a strategic approach to cloud adoption and have the data and technology infrastructure necessary to make use of these advanced technologies.

To develop and use AI, companies need to understand how their business strategy, product strategy, and infrastructure are connected. They also need technology that can manage large AI workloads and operate securely and reliably at scale. In 2023, we witnessed remarkable progress in AI infrastructure. From new virtual machines to improved services, Microsoft has laid the foundation for AI breakthroughs that involve custom-made silicon and powerful supercomputers. AI infrastructure is also becoming more varied and widespread than ever before. From cloud to edge, CPUs to GPUS, and application-specific integrated circuits (ASICs), the AI hardware and software landscape keeps growing at a fast pace.

New research from Harvard Business Review Analytic Services shows that large AI models needed twice as much computing power every 10.7 months from 2016 to 2022. This fast growth shows the high demand for computational power and the importance of increasing computing resources to use AI's full potential. With strong IT infrastructure, organizations can use advanced AI technologies and compete in a more data-driven landscape.

Many companies currently use infrastructure components that are not designed for advanced technologies like generative AI. They need a modern platform that lets them adapt to new tools and methods that are constantly evolving. This report illustrates how companies can adopt strategic cloud use benefitting from purpose-built, full-stack infrastructure to meet the demands of machine learning-driven applications that require extensive data and computing power.

As AI becomes more integrated and the tech stack changes, it is vital to plan for uncertainties and changes that affect cloud computing. Even cloud-native companies face questions about their future cloud strategy. This involves adjusting to new security models, handling evolving skill needs for AI-enabled activities, and maximizing cloud resources with changing workload trends. Moreover, businesses will need to follow emerging rules to successfully operate in the changing regulatory environment of AI-enhanced cloud services. This report, sponsored by Microsoft, can help readers think through expanding their computing resources to fully benefit from AI's potential.

Learn more about Microsoft Azure AI Infrastructure.

Omar Khan
General Manager, Azure Infrastructure
Microsoft

Rethinking Cloud Strategies for Advanced AI

Artificial intelligence (AI) is providing companies with capabilities that would have been unimaginable only a short time ago, including advances in self-driving vehicles, rapid drug development in health care, natural language processing, and better risk management in finance. Advanced AI technologies such as generative AI—which are self-learning algorithms that can create code, text, video, and audio content—are requiring companies to take a more strategic view of their cloud adoption so they have the IT foundation required to make full use of state-of-the-art AI.

The AI InfrastructureView 2021 benchmark survey in April 2021 by Needham, Mass., technology research firm IDC identified inadequate purpose-built infrastructure capabilities as the most frequent reasons that AI projects fail. Thus, forward-looking companies are increasingly using a purpose-built, full-stack cloud infrastructure for the intense requirements of these new AI-powered applications, which require massive amounts of data and computing power for training models. Witness Wayve, a London startup, is creating autonomous driving technology based on cutting-edge technology, such as the latest advances in computer vision. "Advanced AI, the latest and greatest, is absolutely pivotal to what we're doing," says Jamie Shotton, chief scientist at Wayve. "We have to train the algorithm on petabytes and potentially greater amounts of data that we've captured from our fleet of cars, which is a radically different approach to autonomous self-driving than anyone has done before."

Indeed, the computing requirements for large-scale AI models doubled every 10.7 months from 2016 to 2022, according to a study presented at the 2022

HIGHLIGHTS

Advanced artificial intelligence (AI) technologies such as generative AI—which are **self-learning algorithms that can create code, text, video, and audio content**—are requiring companies to take a more strategic view of their cloud adoption so they have the IT foundation required to make full use of state-of-the-art AI.

Forward-looking companies are increasingly using a **purpose-built, full-stack cloud infrastructure** for the intense requirements of these new AI-powered applications.

Today, companies often have **inelastic processing power, insufficient storage capacity, ineffective resource allocation management, and inadequate networking capabilities** for generative AI.

"What's unique about generative AI is copyright ambiguities, lack of truthfulness and accuracy issues, and increase in misuse and biases," says Ritu Jyoti, group vice president, worldwide AI and automation research, with IDC's software market research and advisory practice.

International Joint Conference on Neural Networks in Padova, Italy. Using the cloud for these computing requirements can enable organizations to rapidly scale their infrastructure as necessary.

Generative AI raises key issues around not just scale and performance but responsibility, too, according to Ritu Jyoti, group vice president, worldwide AI and automation research, with IDC's software market research and advisory practice. "What's unique about generative AI is copyright ambiguities, lack of truthfulness and accuracy issues, and increase in misuse and biases," she says. "You need to make sure on the back end that you get the appropriate logs and alerts, so if something goes wrong, you can correct the action right away. Essentially, along with being proactive, you need to embrace generative AI with all the right guardrails."

This report will examine the opportunities and demands state-of-the-art AI presents and how both are causing companies to reexamine their cloud strategy. It will explore how companies are effecting change within their industries by implementing AI infrastructure that allows them to operate in an efficient and cost-effective manner. It will also address the many strategic, technological, and human issues that accompany the redefining nature of advanced technology like generative AI.

The New World of Artificial Intelligence

Savvy customers are increasingly happy with AI-infused offerings like intelligent virtual assistants that can automate support, thus quickly meeting and even anticipating their needs. Companies then are using AI to create new offerings, predict trends, and, for other strategic purposes, reap greater benefits from their overall digital transformation, according to a December 2021 study conducted by Technology Horizon Research that was sponsored by the London-based consultancy EY. Nearly two-thirds (63%) of these advanced AI users report their digital transformation is exceeding expectations, compared to only 37% of companies that are using AI for operational improvements such as cutting costs. **FIGURE 1**

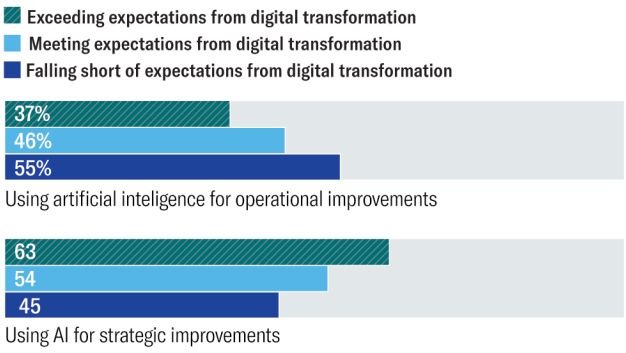
AI-enabled applications are a refinement of traditional programming techniques. These approaches include classical deterministic computing, which is an algorithm that always produces the same output when given the same inputs, and scientific technical high-performance computing, which uses supercomputers and other advanced technology to solve complex computational problems. "Many industries are looking at how classical deterministic or scientific technical high-performance computing can be used in conjunction with AI or machine learning to come up with a blended model that's more efficient," says Addison Snell, CEO of Intersect360 Research, a market intelligence, research, and consulting advisory practice in Sunnyvale, Calif. "The analytics part may discard outcomes that are really unlikely, allowing companies to save huge amounts of computation by focusing on the best solutions. This is leading to the new types of applications that are doing everything from reducing shoplifting in stores to identifying which credit card applicants deserve a lower or higher rate because they present risks."

The enhanced ability to use analytics to ultimately home in on the best solutions and develop new types of applications is enabling organizations to make breakthroughs that are reshaping how industries are addressing problems and finding opportunities.

FIGURE 1

Artificial Intelligence Propels Digital Transformation

Focus on strategic rather than operational improvements brings better results



Source: Technology Horizon Research survey, December 2021

For example, today, fashion designers can spend six to 12 months of intense work coming up with new styles, unsure if their creations will sell or not. But Fashable, based in Viana do Castelo, Portugal, built a generative AI application that can create dozens of original AI-generated clothing designs in minutes, without the need for actual material.

The algorithm ingests data from multiple sources to learn about trends, styles, and clothing types. Using social media to do A/B tests directly with customers lets designers gauge interest and forecast demand for their particular creations before going into production. "We can share the collection with customers before they are produced, avoiding the problem of overstock," says Orlando Ribas Fernandes, CEO and cofounder of Fashable.

Wildlife Protection Services (WPS), a nonprofit organization in Golden, Colo., is using AI to deal with issues related to habitat loss, forest fragmentation, and poaching and other illegal uses of wildlife. "Conservation is a huge challenge globally, and we're not necessarily winning the war," says Eric Schmidt, executive director of the organization.

To improve its odds in the fight, WPS is arming itself with AI models that search images from thousands of camera feeds, looking for humans and vehicles that may be engaged in suspicious activities or animals that may be encroaching on human populations. It sends alerts to wardens and rangers at national parks, private reserves, and other protected areas in sub-Saharan Africa, Southeast Asia, and Latin America.

Health care is another field being transformed by advanced AI. "Around six years ago, we identified that AI was having a disruptive influence in several adjacent verticals, most notably in radiology," says Rui Lopes, director of new technology assessment at Elekta, a Stockholm-based Swedish maker of precision radiation therapy solutions. "We realized the tsunami of AI innovations that were happening in the computer vision and text recognition fields were eventually going to find their way into the medical field, as well."

While radiology is used to diagnose cancer and has long embraced AI, Elekta focused on a related but more involved area: radiotherapy, which is used to treat cancer. Patients with cancer might need as few as five or as many as 45 radiotherapy treatments, which must be constantly adjusted to ensure cancer cells are attacked, while healthy cells are spared. Currently, many people around the world do not have access to this therapy—not because of a lack of technology but because of insufficient medical personnel from the diverse disciplines that must collaborate to ensure the correct adjustments are made to treatment plans.

"AI is well suited to address this problem because we can embed some of that intelligence, and the protocolized ways of working, into the devices themselves to increase the access to treatment for a larger swath of patients worldwide," Lopes says. "This provides not just personalization of care but democratization of a standard of care, allowing more advanced protocols to be deployed in regions of the world that lack the human capital to do so now."

"Many industries are looking at how classical deterministic or scientific technical high-performance computing can be used in conjunction with AI or machine learning to come up with a blended model that's more efficient," says Addison Snell, CEO of Intersect360 Research.

Cutting-Edge AI Infrastructure

AI projects require a level of computer power and scalability that is difficult and costly to implement on premises. Companies often lack the right systems to handle the demands of complex model training and keep pace with the rapid technology changes in the AI space.

Advanced applications like generative AI must be supported by cutting-edge infrastructure that provides the performance, flexibility, and scalability that these applications demand. According to the IDC report, though, AI infrastructure remains one of the most consequential but the least mature of infrastructure decisions that organizations make as part of their future enterprise. As a result, today, companies often have inelastic processing power, insufficient storage capacity, ineffective resource allocation management, and inadequate networking capabilities for generative AI.

Organizations have three basic approaches to deploying cloud services—infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS)—each of which divides which technology capabilities the cloud provider and company are responsible for managing in different ways. **FIGURE 2**

These approaches are not mutually exclusive, and all three are often used by enterprises and medium-sized companies.

There is SaaS, which is on-demand access to cloud-hosted application software that is ready to use. Another approach is PaaS, which is on-demand access to a cloud-based platform on which companies can develop, run, and manage applications

FIGURE 2

Dividing Duties

Companies are responsible for managing different parts of technology in different cloud models

Company responsibility for managing capabilities indicated in bold.

On Premises	Infrastructure as a Service (IaaS)	Platform as a Service (PaaS)	Software as a Service (SaaS)
Applications	**Applications**	**Applications**	Applications
Data	**Data**	**Data**	Data
Runtime	**Runtime**	Runtime	Runtime
Middleware	**Middleware**	Middleware	Middleware
O/S	**O/S**	O/S	O/S
Virtualization	Virtualization	Virtualization	Virtualization
Services	Services	Services	Services
Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking

Source: Guru99

using tools and programming languages that the cloud vendor supports. The customer controls the applications but not the underlying cloud infrastructure. Then there is IaaS, involving on-demand access to cloud-hosted physical and virtual storage, networking, and servers. The customer controls the applications, storage, networking, and operating systems but not the underlying cloud infrastructure.

Small and medium-sized businesses often favor the SaaS model and buy off-the-shelf components for their AI projects, even if they don't meet 100% of their requirements.

As a small nonprofit, WPS started with a SaaS model. While this move brings technology efficiencies and cost savings, the SaaS model can also raise issues. "If we choose a provider that has a great tool we want to work with, we have to be concerned if the tool will be available in the future" because the provider might discontinue it or go out of business altogether, says Matt Morrisette, a software engineer for the organization.

WPS is also running into non-SaaS issues where it needs regional support to ensure it remains in compliance with the General Data Protection Regulation, better known as GDPR, the law on data protection and privacy in the EU, as its operations expand into new areas. "If you don't host your data in Europe, you can't launch projects there," Morrisette says. "These were things we didn't really think about when we initially launched, but now it's become an issue. We are trying to be more flexible in how we launch AI projects now."

Large enterprises might start with SaaS but move to an IaaS or PaaS model. The choice will be driven by the size of the company, the agility requirements, the available talent, and the use case for the technology. For some use cases, the decision is dictated by the lack of off-the-shelf SaaS solutions. "The one big lightbulb for me was to understand that there was no off-the-shelf, one-size-fits-all solution that's available out there today unless you're in a fairly commodity industry," Lopes says.

Most of the top data-centric organizations are using the cloud to leverage more IaaS and PaaS capabilities, taking advantage of the speed and agility of these models to innovate where they could not have done so with traditional technologies, while reducing costs.

"We use a different mix of tools and technologies, some we build in-house, some we have acquired from partners," Wayve's Shotton says. "Using a managed platform gives us the ability to scale quickly and reliably. It also allows us to focus our efforts doing the research and solving problems around autonomous self-driving rather than building additional tools ourselves."

Preparing for Scale of AI Development

Generative AI requires input from a continuous, larger data set to refine the algorithms. The more data, the more the algorithm learns. To train a model to make predictions, generate text, and handle the other chores of advanced AI, engineers use a process called "inference." This process can be expensive since it needs to run millions of times, consuming large amounts of computing time.

The different stages of AI development—from data analysis and experimentation to model training and inference—can require different capabilities, tools, and AI infrastructure. The workload cost and scale can change quickly from one stage to the next.

"Since advanced AI applications can be costly to train, you need to have rock-solid engineering under the hood to avoid wasting money," Shotton says. "This is resulting in a blurring of the lines between research and production."

There are ways to reduce these costs and speed innovations. For example, organizations can build their own models or use pretrained models for some applications and then fine-tune these models using their own customer data.

An "AI-first" approach to infrastructure can accelerate model training and accelerate AI innovation, as well. Cloud infrastructures leverage graphical processing units, an approach designed originally for 3D gaming and which since has incorporated specialized hardware to execute calculations in parallel and significantly reduce training time.

A common infrastructure can increase the velocity of AI innovation. "If you have the data in one place and you offer services around that data, it becomes more compelling for people to use those standardized services," Lopes says. "Instead of spending two weeks having your local workstation churn through creating a model, a developer can do that in a couple of days with the more-powerful computing resources in the cloud. We wanted to unleash our developers to be able to use the data, while still maintaining the right security constraints."

As the Elekta example shows, the training workload will execute much faster in the cloud than on a local workstation, allowing the solution to be found sooner. In addition, some models are too large to even run on a local workstation, making the cloud the only viable option for some advanced AI.

On a small scale, models can be trained in small clusters on workstations on premises. However, according to Adam Moore, director of global cloud solutions for Elekta, this approach can disguise problems that will arise when the models are scaled up to production. "At that point, you will need to radically scale up the amount of data you use," he explains. "By training the models in the cloud, you can identify those problems earlier and build resilience into your compute infrastructure so you avoid hardware failures."

Proper Data Stewardship

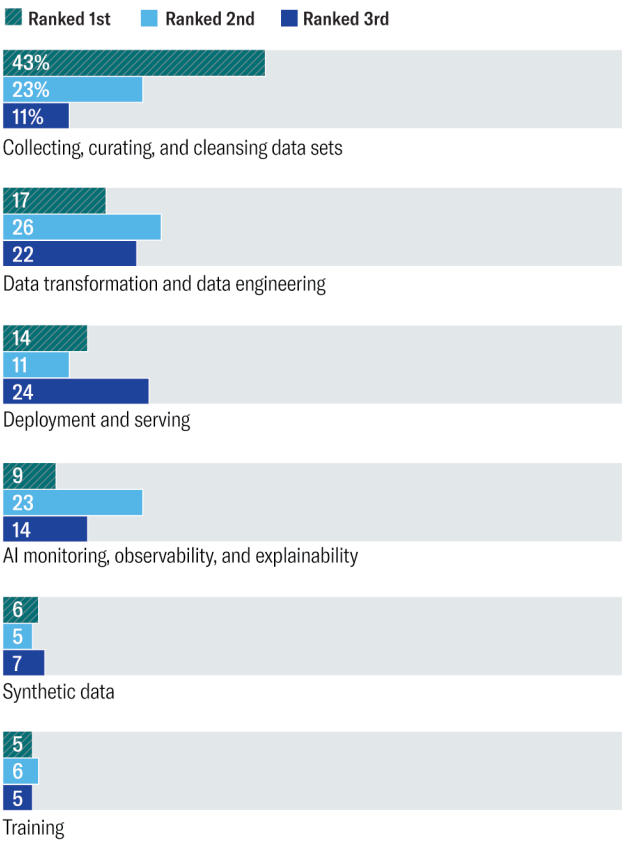
Data is at the core of every AI project. "Collecting, curating, and cleaning data sets" was ranked as the top AI infrastructure element that should receive the most resources, according to the AI Infrastructure Ecosystem, a 2022 study by the AI Infrastructure Alliance, an industry group in San Francisco. **FIGURE 3**

Companies need to be specific about what they are trying to accomplish with their AI initiatives and determine if they have enough of the correct kind of data. They then must choose how to prepare the data for use by AI applications, move it across the network, and store it. "The biggest challenges are not technical but strategic," Intersect360 Research's Snell says. "People have magical thinking that they can throw infinite computing power at a limited amount of imperfect data and learn something they didn't already know."

FIGURE 3

AI Starts with Data

Companies focus their AI resources (talent, time, money) on data first



Source: AI Infrastructure Alliance survey, 2022

"People have magical thinking that they can throw infinite computing power at a limited amount of imperfect data and learn something they didn't already know," says Intersect360's Snell.

Many of the issues revolve around data movement. As Snell explains, "Let's say I've decided to use the cloud. Now I've got to move all my data up there, presumably for the analytics to happen there. Is that a secondary data copy? Do I also have the data somewhere else on premises? If I have it in both places, then how do I keep it in sync? If I have it only in one place, then where is it? Where is it backed up? And what happens if I want to move it back on premises or to a different cloud at a different time?" Prospective cloud customers must also understand the ingress/egress fees that come with data movement. This dizzying array of questions has important implications for infrastructure choices.

Snell says that AI initiatives must be based on a long-term data stewardship strategy. This time frame transcends a three-to-five-year service-level agreement (SLA) with a cloud vendor. "You need to maintain long-term control and possession of your data and think of 'long-term' as at least 100 years," he says. "Let's say you are a pediatric hospital, and you take a scan of a two-year-old. You want to keep that data for the life of a patient, so that could be 100 years. If that information could matter to the person's children and grandchildren, you can easily see instances where some data you want to keep forever."

Getting Everyone around the Table

The latest approaches to AI can fundamentally change an organization's processes and business model. For this reason, companies must think beyond just the technology implications to such matters as how advanced AI will impact budgeting and talent strategies.

Tensions can arise between the finance and engineering departments, like whether to spend scarce capital to develop applications quicker. "We must be agile and move quickly if we are to succeed in our mission, but as a startup, we need to work within our financial constraints," Shotton says. "It's a journey, and we partner closely with our finance team to build up from small to larger investments in compute spend."

AI can not only draw on data and resources across the organization but also can impact every aspect of it. Elekta began its AI efforts with a dedicated cloud team to generate the initial spark of enthusiasm, and then brought aboard other departments, including engineering, technology, and security, as well as sales, customer, and product teams, to get their feedback.

"We brought everyone around the table, addressing their concerns and ensuring we had an effective feedback cycle across the entire organization," says Elekta's Moore. "We maintained a clear focus on understanding the value to the customer in everything we did."

The custom-built cloud infrastructure needed to support generative AI can also have tremendous implications for an organization's talent strategy. Just barely behind the always present concerns about cost (56% of respondents) on the issue of AI implementation challenges identified by the IDC AI StrategiesView study was a lack of skilled personnel (55%) and machine learning operators (55%). **FIGURE 4**

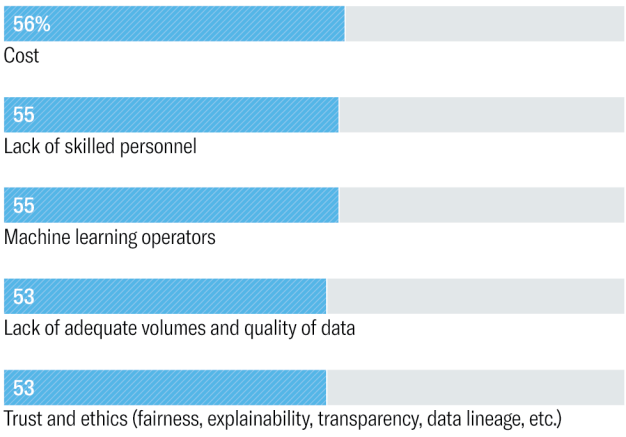
"Sometimes the perception is your on-premises skill set is analogous to what you will need for the cloud," Moore says. "People think it's just taking your computing and data resources to the cloud. From an architectural perspective, there's a number of design patterns and thinking that you need, such as operating at scale and resilience."

In shifting its AI development to the cloud, Elekta employed consultants in DevOps, tools, and other technical areas. "There

FIGURE 4

AI Implementation Challenges

Cost is followed closely by talent issues as top barriers



Source: IDC survey, April 2021

"There will be a lot of hybrid environments—not just the cloud and on-premises, but the cloud and the edge. Edge locations, the intersection of 5G and real-time streaming, are shaping up and will empower newer AI capabilities for businesses," says IDC's Jyoti.

are generalists who will help you do 'a lift and shift' to the cloud, but they don't necessarily understand what you're trying to achieve," Moore says. "We leverage an external system integrator to help us speed up the journey, but we were also focused on making sure that we had the right balance of in-house and external skills so we didn't lose all our skills when we finished the project."

AI applications, which devour huge amounts of data from every corner of an organization, can require heightened levels of collaboration. In the past, Elekta's distributed R&D centers would use their own processes, tools, and data repositories. Consequently, the company's data was cached and guarded by different departments.

For Elekta, Lopes, the director of new technology assessment, says this "fractured approach" along business lines provided "the opportunity to reevaluate our infrastructure for how we could consolidate not just our way of working in our processes but also the fuel that drives AI, the data."

Choosing a Cloud for Growth

The diversity of cloud offerings allows organizations to choose the best approach for their unique AI needs. In the past, there were many legacy reasons for why people were hesitant about moving to the cloud. With generative AI technologies, IDC's Jyoti says the question has shifted from whether to use the cloud for AI applications to which cloud provider is best aligned with a company's strategic vision for AI.

"The fight now is over which platform will become the platform of choice for generative AI," Jyoti says. "As companies make their strategic investments, they are going to bet on which cloud they use. We suggest companies work with trusted hyperscalers or suppliers and do their due diligence. But don't sit back because the train has left."

The selection depends on both the capabilities of the cloud vendor and the ecosystem of partners and vendors that is built around the vendor's offerings. The cloud will always be the center of gravity for generative AI because of the scalability, agility, and cost advantages it brings.

In the rapidly changing AI space, organizations need an architecture that provides flexibility to allow them to take advantage of modern technologies and approaches. "There will be a lot of hybrid environments—not just the cloud and on-premises but the cloud and the edge," Jyoti asserts. "Edge locations, the intersection of 5G and real-time streaming, are shaping up and will empower newer AI capabilities for businesses."

The computing power needed for generative AI could be changed by experimentation going on now about the right number of training data sets (something called training tokens) and model size or capacity (known as parameters) that are used. "A greater number of parameters is more compute-intensive, which is more costly," she says. As researchers identify the right balance of data, model architecture, hyper parameters, and inference runtime, the cost equation of advanced AI could shift again, changing the infrastructure requirements and providing more opportunities to discover breakthrough solutions.

Companies should see generative AI as a journey and develop relationships with vendors that will grow as their AI use does. "Everyone won't shut off their on-premises technology today and go in the cloud tomorrow," Jyoti says. "You need to make sure you've matured your SLAs and governance practices, while ensuring there is automation at each step in the process."

Conclusion

Today, most companies rely on an infrastructure foundation designed for general-purpose components rather than developing an infrastructure stack honed for the complex and demanding needs of advanced technologies like generative AI.

"People are going gaga over the ChatGPT aspect of AI, but that's only a small part of generative AI," Jyoti says about the AI-driven chatbot based on language model Generative Pre-trained Transformer (GPT) 3.5. "AI is also about creating new content like video, audio, images, and code. The technology can have a transformational impact on almost every industry. For example, you can use it to not just create marketing blogs, sales proposals, social media posts, summarize customer feedback and/or code generation, but also to design toy images or create digital twins for design and simulation of products. The benefits of generative AI can be remarkable, from improving knowledge worker productivity to increasing innovation velocity and/or accelerating decision velocity."

"Just as word processing programs became commonplace tools to increase general productivity, AI will become a common tool for organizations to increase innovation. AI will be a superpower that will democratize content creation. AI will not replace jobs, it will transform work," says Orlando Ribas Fernandes, CEO and cofounder of Fashable.

The AI journey, as Fashable's Fernandes puts it, is driven by constantly needing new inputs, new training, and new learning that continually increase not just in quantity but also in quality. He points out that the quality-over-quantity aspects could be demonstrated by having an AI algorithm generate personalized clothing lines for individual shoppers. "In the near future, you will have a digital closet of clothing designs that you can ask a manufacturer to produce just for you," Fernandes says. "We will use the metaverse to create physical goods that are exclusive to each person."

Those on the cutting edge of AI say companies should be exploring technologies like generative AI now. "AI will be a tool for every company," Fernandes says. "Just as word processing programs became commonplace tools to increase general productivity, AI will become a common tool for organizations to increase innovation. AI will be a superpower that will democratize content creation. AI will not replace jobs—it will transform work."

Harvard Business Review
ANALYTIC SERVICES

ABOUT US

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticservices@hbr.org.

hbr.org/hbr-analytic-services

Copyright © 2023 Harvard Business School Publishing.

MCCRE37600724