Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

TRENDS REPORT

# The Rise Of The AI Cloud

## AI Is Reshaping Your Cloud Strategy

March 7, 2024

By Lee Sustar, Mike Gualtieri with Lauren Nelson, Arthur Ross, Kate Pierpont

FORRESTER®

## Summary

The AI boom continues, and cloud providers are at the heart of it, with the Microsoft/OpenAI partnership sparking similar efforts from rivals. Cloud providers provide scale for large language models (LLMs), generative AI (genAI), and GPU resources that most organizations cannot afford on their own. Cloud providers' click-and-go AI managed services, along with infusions of AI into existing services, put new capabilities into users' hands. But cloud isn't ideal for every AI implementation and enterprises want to know how to best revamp their cloud strategy. This report provides guidance on integrating AI into your cloud strategy and examines the potential and limitations of running AI workloads in the cloud.

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

# AI Launches A New Era For Cloud

Cloud hit the IT world over a decade ago. Its early days were filled with interesting market dynamics, surprising new innovations, and tons of exciting developments. But as cloud settled as a more mature market with few players, there weren't a lot of new developments to keep tabs on or new players to watch — until the start of the AI boom. With increased excitement over the fast development of GPT-3, every technology market refocused their sights on AI innovation. Possibly the biggest benefactor? Cloud providers. AI restates the value proposition of the cloud. Since its inception, cloud enabled enterprises to test innovative ideas with little capital expenditure or risk of failure — benefits driven by cloud's economic model, its wealth of pre-built services, and its ability to deliver specialized infrastructure readily. Each value has been utilized significantly as enterprises harness cloud to train and deploy LLM for genAI use cases.

This means that enterprises now have access to new services to accelerate their genAI efforts. However, cloud is not always the answer. Tech leaders should ensure that their cloud providers can adequately address their workload requirements and focus on the ongoing dynamics as they unravel.

# Make AI Central To Your Cloud Strategy

Many enterprises have a decade of cloud experience under their belts and honed operating models and adjusted strategies to meet significant scale. Most recently, companies have been pivoting their strategies and architectures to build on cloud-native technologies like containers and turnkey development and abstracted application services rather than simple virtual machines (VMs), storage, and networks. AI is another pivotal moment for cloud. Enterprises must undergo yet another refresh to understand the capabilities required (i.e., activities, competencies, underlying technologies), operating unit and governance adjustments, and its role in their broader strategy/vision. This journey may even include onboarding a new AI cloud.

To chart your way forward with AI in the cloud, you must start with your company's own objectives for AI usage, rather than snap up the latest vendor offerings. Forrester recommends taking a five-step approach: 1) keep up to date with the market dynamics; 2) recognize the advantage and disadvantages of cloud for AI; 3) understand the varying AI workload requirements; 4) clarify your cloud priorities before purchase; and 5) update your cloud strategy and operating model to reflect this pivotal change.

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

## The Cloud Market Is Flush With AI Activity

In 2023 and continuing into 2024, there has been intense competition between AWS, Google, and Microsoft centered around AI. Current and prospective customers are left to sort through growing lists of AI services to find what is relevant, practical, affordable and — crucially — available right now. But big promises and backlogged orders for GPU resources have created a path for new AI cloud competitors and provided new disruption opportunities for longstanding cloud providers that have long lagged the leaders. Taking advantage of these opportunities may require increasingly multicloud vendor strategies. Cloud customers considering AI at scale should set aside previous assumptions and take a fresh look at their options.

- **The race is already in progress.** There is no shortage of AI action in the cloud market. The Microsoft/OpenAI combination with ChatGPT for (LLMs) and Copilot technologies set off a rush for NVIDIA GPUs to run AI workloads and demand for Copilot technologies. Microsoft is wading in with its own custom silicon to meet that demand. Cloud rivals have countered after an initial stumble. Google's Bard has been boosted by its Gemini model. AWS is updating its custom Trainium chip as a NVIDIA alternative paired with its Bedrock managed AI service, an alliance with OpenAI competitor Anthropic, and its Q chatbot based on LLMs. Oracle Cloud Infrastructure is in the mix with its SuperCluster offering comprised tens of thousands of NVIDIA GPUs. IBM offers its watsonx and Red Hat OpenShift Data Science for a multicloud, multimodel approach. That's a lot of action. The next step is making it more accessible and more affordable.
- **The market is awash with new players.** The new AI cloud players have plenty of capital and key capabilities. CoreWeave took advantage of the genAI moment with the right infrastructure. Originally established as a blockchain venture based on GPUs for cryptocurrency mining, CoreWeave snapped up distressed assets of failed crypto efforts to amass enough NVIDIA GPUs to suddenly become a player as the cloud giants scrounged for similar resources. NVIDIA's surge in revenue enabled the company to back CoreWeave, which competes with NVIDIA's biggest customers. Vultr, formerly a niche provider, has leveraged NVIDIA GPUs to create AI-cloud-in-a-box solutions in partnership with telcos in various parts of the world. The list of GPU clouds will grow as investors back efforts to cherry-pick pricey AI workloads from the big players.
- **Cloud challengers are in the thick of the contest.** Oracle Cloud Infrastructure (OCI) has leveraged its relationship with NVIDIA and its SuperCluster offering to win business away from the hyperscalers using low cost and specialized data/AI capabilities as levers. IBM Cloud, eclipsed by the hyperscalers years ago, has a

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

new bid for relevance with watsonx/OpenShift Data Science for its "multicloud/multimodel" approach (and, if you're an AWS customer, you can run watsonx there as well). HPE, long a bystander in the cloud market, is making a strategic play with its HPE GreenLake for LLMs. Like the new cloud GPU players, these established tech players are focused on snatching premium AI workloads away from the cloud giants.

- **SaaS players in the game yet hosted on their biggest competitors.** Cloud platform/SaaS providers that run on hyperscaler infrastructure are emerging as AI clouds in their own right. The list includes Salesforce Data Cloud/AI on Salesforce Hyperforce, SAP Business AI on SAP Business Technology Platform, and Snowflake Data Cloud. All of this has put pressure on the hyperscalers. On the one hand, they gain revenue through these opportunities. However, they must counter these disintermediation pressures with their own native AI services even as they maintain high levels of investment for wider infrastructure.

## Cloud Has AI Advantages And Disadvantages

At the heart of cloud's AI advantage is specialized infrastructure geared toward AI use cases. The hyperscalers have a long history of success at deploying fast infrastructure starting with basic x86-based instances and, later, cheaper ARM chips and application-certified stacks for enterprise software. However, deploying GPUs at similar scale presents additional challenges — even for hyperscalers. To successfully deliver AI infrastructure, the goal is to maximize performance for a range of AI workloads including data preparation, model training, and inferencing — regardless of whether the apps are business or customer-facing. To do this, cloud providers are pouring multibillion dollar investments into the 10x AI opportunity to procure or create AIs chips with scale to support the computational demands of LLMs. The cloud providers' goal is to either sell their native services or, at the very least, partner, buy or host solutions that customers may want. Success is paramount for the hyperscalers, as they need to sell these premium services to cover the costs of building out data centers and enabling price-competitive, low-cost (or loss-leading) cloud infrastructure. For their part, enterprises must be intimately familiar with the advantages and disadvantages of cloud AI for their own purposes:

- **Advantage: Pay only for what you use.** We all know that cloud providers have transformed the economics of enterprise IT with consumption-based pricing. That's attractive for organizations at both ends of the AI adoption profile: those new to AI that seek to experiment and assimilate the technology with minimal commitment and organizations with mature cloud operations that want immediate scale and capabilities. Pay-as-you-go AI services meet the needs of organizations

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

wary of overspending on AI.

- **Advantage: Access to specialized chips.** The cloud giants can get their hands on GPUs more easily than the everyday enterprise. Why? Direct relationships with mass purchases from NVIDIA and AMD — along with their own efforts to meet the demand. AWS has long pushed its custom Trainium and Inferentia chips for AI as a cost-effective alternative to NVIDIA, and Google has championed its tensor processing units (TPUs) for neural network workloads associated with genAI. Microsoft's forthcoming Azure Mia chip marks a further addition to cloud provider AI-focused custom silicon.
- **Advantage: Managed services make it simpler.** Every company is scrambling to figure out their genAI story and broader AI strategy. To accelerate that journey, many enterprises desire solutions that offload some of that responsibility. The cloud providers serve this need well. Turnkey systems are available for both technical and non-technical users: Microsoft Copilot technologies and AI Builder on Microsoft Power Platform and Google Cloud's ready-to-run AI applications. Most vendors offer AI and ML solutions that infuse AI capabilities into the offering (e.g., IBM watsonx, OCI Generative AI). But clouds also serve customers shaping their own AI environments with foundational AI capabilities: AWS Bedrock, Google Vertex AI, and Azure ML.
- **Disadvantage: AI in cloud can get expensive when workloads run hot.** Pay-as-you-go approaches make AI experimentation easy — but AI in production and at scale can translate to high-usage and budget-bursting. The FinOps Foundation points out that when using third-party open model platforms, customers are responsible for managing computational resources. How do you solve for this? The first step is policies that regulate autoscaling, as is normal protocol for cloud usage. To mitigate the risk for AI, seek out all available optimizations and discounts and align platform, configuration, scaling, and purchase methods at the outset. You may have to explore alternative deployment options or avoid design approaches that can artificially increase your costs.
- **Disadvantage: Many applications may require inferencing on the edge.** Public cloud infrastructure was designed with an all-roads-lead-to-Rome approach: by ingesting workloads and data in shared infrastructure, cloud providers provided scale and innovation that transformed enterprise-class IT. That centralized model also works for AI workloads — until it doesn't. Much of the ROI of an AI investment can only be realized by putting inference capabilities close to users and/or customers in edge locations. Cloud providers are tackling this by enabling inference in multiple availability zones. However, these efforts are not always optimal from a cost and/or performance standpoint.

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

- **Disadvantage: Data might have to move to the cloud first.** The data that is most ripe for AI use cases may not be in the cloud at all. On-premises infrastructure holds vast amounts of data which remain essential to all types of organizations. Moving it and/or storing it in the cloud may be expensive. There may even be security or sovereignty restrictions that preclude it from going to the cloud at all. Understand these limitations as you explore the potential use of cloud-based AI services.

## Not All AI Workloads Are Created Equal

AI workloads vary significantly and as such, so do their infrastructure requirements. Enterprises seeking to define their AI strategy and its implications on its broader cloud strategy must understand the following differences between AI workload requirements and how well the cloud market satisfies these needs today.

- **Data prep workloads need throughput.** The task of accumulating, transforming, and formatting data for AI is a burdensome prerequisite to training AI models. Cloud providers address these challenges with their vast data storage and data pipeline processing services. Cloud providers seek to differentiate their AI data workload capabilities with automated scaling and descaling and fast data access for model training and inferencing.
- **Training/tuning workloads need specialized compute.** Training AI models for use cases like LLMs demands specialized AI instances via GPUs or specialized AI chips. All cloud providers offer instances and/or services with NVIDIA GPUs. But they also offer their own custom options. Google has long offered its own design TPUs. AWS offers its own AWS Trainium and instances with Intel Gaudi. Azure recently announced the development for Azure Mia. NVIDIA continues to dominate but other semiconductor companies such as AMD plan to compete with NVIDIA.
- **Inferencing workloads need low-latency concurrency.** Once models are trained, they are used in production; this is known as inferencing. Unlike data prep and training workloads that process huge amounts of data, inferencing is about how quickly a model can return a result. A model may be deployed such that it is called by an application thousands of times per second. This requires specialized hardware to return an answer from a model at the lowest possible latency. It also requires scale-out of inferencing to handle concurrent calls to avoid single-path bottlenecks. Cloud providers offer specialized inferencing services that provide AI instances designed for inferencing that can scale-out. As with AI training, some cloud vendors provide their own chips that support inferencing. For example, AWS

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

Inferentia is designed by AWS for model inferencing.

- **Models need a mix of these strengths.** Key considerations are how effective cloud providers are at inferencing, the performance of their model compiler, and the capabilities of their hardware abstraction layer software. Cloud providers tout their ability to deliver low-latency concurrency for inference — that is, enabling the preprocessing of a request before the previous request is completed. These providers use a mix of custom silicon and ML instances deployed for high availability. Some also provide capabilities to deploy models on edge hardware.

## Clarify Your AI Cloud Priorities

The last year has seen enterprises scramble to access GPU capabilities as hyperscalers were forced to throttle demand and new players struggled to reach scale. This was defensible as enterprises sprinted for first mover advantage in their respective industries. It is now an antipattern as the new vendor landscape takes shape. Organizations must differentiate workload type according to vendor capabilities.

- **Distinguish AI "in the cloud" from AI "of the cloud."** AI workloads may be best suited to run in a GPU-only cloud startup; the AI "of" your primary cloud may provide advantages for AI-assisted development and automation/intelligence of platform operations, but it may not have the best price/performance for AI workloads.
- **Don't overpay for cloud GPUs.** Sometimes hyperscaler custom silicon is good enough, such as the AWS Trainium/Inferentia offerings and Google TPUs. Other use cases may be compelling for next-generation GPUs such as NVIDIA's GH200 Grace Hopper Superchip that incorporates CPU, GPU, and DPU technologies. Cloud providers are leveraging AI hype to oversell. Take the time to assess and bring in FinOps pros and cloud cost specialists at the outset.
- **Accept out-of-the-box cloud AI when speed and simplicity are key.** You don't need a full-blown, custom genAI deployment to see results from AI; Microsoft Copilot technologies are available in many Azure services and Google Duet AI aims to be an AI "collaborator." AWS's Amazon Q chatbot will enhance a growing number of services. Salesforce and Databricks offer their own off-the-shelf, ready-to-run AI services, too.
- **Consider the wider array of AI infrastructure offerings in the market.** Public cloud isn't always the best option for AI infrastructure. Running round-the-clock workloads can spike cloud costs upward. Investments in on-premises data centers — often seen as an antiquated notion in many "cloud-first" organizations — can make sense for many AI workloads, particularly where there's substantial data

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

gravity in customer-owned legacy infrastructure.

## Modify Your Cloud Strategy To Include AI

AI is your job now too — another silo to bust. AI and cloud strategies can no longer live apart. It is imperative that cloud AI services and workloads are embedded in mainstream cloud operations. This means renewing the cloud operating model, adding in the AI capabilities required (i.e., activities, competencies, underlying technologies), fine-tuning operating unit structures, adjusting governance, and defining AI's role in their broader cloud strategy and vision.

- **Define the role of AI in your broader cloud strategy and vision.** AI cloud services can — and should — go through the normal cloud service adoption framework, governance, and deployment. Firstly, existing investments in on-premises and/or non-sanctioned cloud AI infrastructure must be inventoried. Your cloud strategy AI plank must include a disposition plan to keep, dump, and/or migrate these existing investments without disrupting in-flight and/or production AI projects. The cloud strategy must also recognize the high velocity of AI innovation. AI teams may say they absolutely need X, only to say a few weeks later they don't need X anymore — they need Z. These unique considerations reshape the wider cloud strategy, driving decisions about vendor selection, spending, and personnel to align with organizational goals more closely.

  Yet the demand for cloud AI services changes the metabolism of enterprise IT by placing it in closer relationship with the business and other stakeholders keen to seize the opportunities of AI while keeping within risk appetite. This inevitably reshapes the wider cloud strategy, driving decisions about vendor selection, spending, and personnel to align with organizational goals more closely.

- **Build out your cloud-related AI capabilities.** Mature cloud strategies formally define the capabilities their team supports and maintains. This list describes the activities it supports, what success looks like, metrics, and the underlying technologies needed to achieve that capability. Examples of added AI capabilities might include: 1) Provide AI infrastructure workloads: data prep, model training, and inferencing; 2) minimize cloud costs with strong governance practices and chargebacks; and 3) evaluate the impact, latency, fault-tolerance, and use of pre-trained, externally called AI models (such as LLMs) on existing application infrastructure. Each cloud AI capability should include detailed descriptions, metrics for success, and a complete list of technologies such as collaboration software, databases, and cloud platforms.

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

- **Update your vendor management protocols to include new players.** Sourcing and vendor management teams are just getting accustomed to working with the major cloud platforms. AI presents a new challenge, as it may shake up your primary cloud vendor choice, add in new vendors, or even push your team to procure even more quickly with new vendors. To do this well, work with your procurement team early. Change in protocol may require executive support. For example, an upstart GPU-only cloud presents a different vendor risk than a trillion-dollar hyperscaler.
- **Prepare cloud platform teams to support AI at scale.** ModelOps — the tools, technology, and practices that enable cross-functional AI teams to efficiently deploy, monitor, retrain, and govern AI models in production — is well established. IT teams are typically tasked with supporting the infrastructure for data engineering and data warehousing to enable data science pros to do the rest. Now, the new cloud AI services draw data scientists and cloud platform teams closer together to deliver the innovation enabled by genAI. At the same time, cloud platform teams will be tasked with overseeing the SaaS-based abstracted AI services. Cloud platform team capacity will be a constraint on cloud AI service adoption without additional resources or skills acquisition.
- **Align your AI cloud to your security, governance, risk, and compliance (GRC) program.** Cloud security and governance practices are getting more sophisticated as the workloads running on cloud platforms continue to push boundaries. AI will challenge your existing procedures for reasons such as model provenance, privacy regulations, and third-party risk. Securing genAI requires a multipronged effort. It is easier to secure smaller functional models on organizational data than to lock down large-scale foundational models. The AI security agenda involves identifying key dependencies, assessing third-party risks of suppliers of genAI, and training teams to deal with new attacks on prompt engineering.
- **Be rigorous about cloud cost management.** A major element of cloud governance programs is cost governance. Rapidly scaling infrastructure can quickly get your organization into excess spending. Mature FinOps practitioners are no strangers to these concepts, but your AI stakeholders may be new participants in this world. Don't let their high tolerance for big spending on experimentation and early-stage efforts become baked into operating costs. The FinOps Foundation's assessment of key cost drivers in cloud AI services is a good reference for your planning.

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Forrester Report Copy Prepared Exclusively For Ira Kantor With Forrester Research GSA, Inc. Distribution and reproduction are prohibited.

FORRESTER®

# We help business and technology leaders use customer obsession to accelerate growth.

FORRESTER.COM

### Obsessed With Customer Obsession

At Forrester, customer obsession is at the core of everything we do. We're on your side and by your side to help you become more customer obsessed.

### Research

Accelerate your impact on the market with a proven path to growth.

- Customer and market dynamics
- Curated tools and frameworks
- Objective advice
- Hands-on guidance

Learn more.

### Consulting

Implement modern strategies that align and empower teams.

- In-depth strategic projects
- Webinars, speeches, and workshops
- Custom content

Learn more.

### Events

Develop fresh perspectives, draw inspiration from leaders, and network with peers.

- Thought leadership, frameworks, and models
- One-on-ones with peers and analysts
- In-person and virtual experiences

Learn more.

FOLLOW FORRESTER

### Contact Us

Contact Forrester at www.forrester.com/contactus. For information on hard-copy or electronic reprints, please contact your Account Team or reprints@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
Tel: +1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

© 2024 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.