MOOR
INSIGHTS & STRATEGY

THE CASE FOR A MANAGED AI AND ML MODEL INFRASTRUCTURE

INTRODUCTION

As artificial intelligence (AI) and machine learning (ML) gain market momentum, foundation models are being consumed and managed differently. Increasingly, organizations want to build and manage their own models that balance cost, reliability, and specificity of model outputs. Over the past 12 months, this balance has become easier to achieve for these reasons:

- **More diverse and efficient foundation models** — The recent proliferation of different sizes and types of AI and ML models has lowered the barriers for executing industry- or organization-specific training and fine-tuning. This leads to higher quality and more accurate outputs from existing foundation models or new models derived from them.
- **Improved tooling and IT operational alignment** — We have seen a maturing of data scientist tools in the form of both notebooks and IDEs. The industry has further improved tooling by taking steps to better align traditional IT tools such as Kubernetes and CI/CD pipelines with AI and ML needs. This has removed barriers between data scientists and IT ops resources.
- **More robust infrastructure** — Another key ingredient of the reduced costs of custom models has been the rampant pace of expansion and improvement for AI and ML infrastructure, including the development of GPUs, TPUs, and accelerators. This has led to a drastic reduction in costs, such that some training implementations cost a small fraction of what they did even a year ago.

While these factors are sure to further increase the pace of development for AI and ML applications, challenges still remain if an organization wants to host its own AI or ML infrastructure, including:

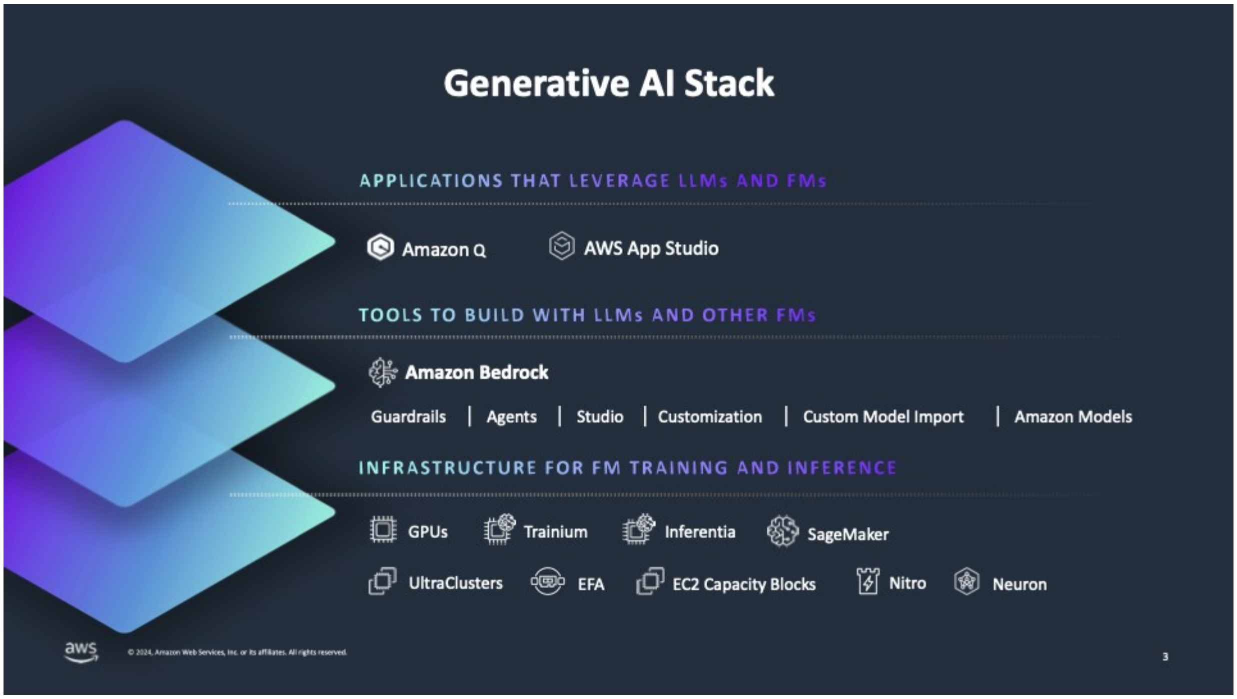
- **Choice** — An internally built and managed AI and ML environment may limit an organization’s choices for infrastructure, networking, models, and tooling due to availability or budgetary constraints. Also, rapid changes in the market may prompt the organization to regret recent decisions.

- **Costs** — The costs of change as infrastructure matures are well understood. But AI and ML infrastructure is a particularly complex and high-cost effort that requires constant human attention. There are also costs associated with how the GPUs are utilized and governed amid competing priorities and projects.
- **Time to market** — An additional challenge associated with large GPU-driven neural networks is managing reliability issues as components inevitably fail. These failures raise the potential for training runs to be halted or lost, which slows time to market.

So, although the barriers to entry for building and training AI and ML models continue to shrink, organizations may want to consider a managed end-to-end offering such as Amazon SageMaker HyperPod to improve time to market and overcome the complexities and costs associated with on-premises models.

The Amazon Web Services Approach to AI

AWS's comprehensive generative AI technology stack supports a wide range of use cases. Choosing the best way to deploy these capabilities depends on the level of flexibility and robustness an organization wants in its own AI strategy. AWS presents its capabilities in three layers.



Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs

Amazon Q | AWS App Studio

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs | Trainium | Inferentia | SageMaker

UltraClusters | EFA | EC2 Capacity Blocks | Nitro | Neuron

aws © 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3

Amazon Web Services generative AI stack (Source: AWS)

The top layer includes applications that provide varying degrees of AI-based assistance to end users. These users can be developers or skilled business analysts looking to create new applications more rapidly by employing AI assistive tools.

The middle layer of tools is geared toward developers who want to integrate AWS-hosted large language models (LLMs) into their applications and agents. This layer provides developers with model choice without needing to retool applications or agents. Additionally, this layer provides services such as Guardrails that can be shared across multiple models.

The bottom layer of infrastructure is the base for all other AWS capabilities. However, it is also exposed to AWS customers that seek to build, train, and deploy their own AI and ML models.

A Closer Look at Infrastructure for Training and Inference

AI and ML have demonstrated massive potential in areas such as customer support, application development, and process optimization. And using off-the-shelf foundation models for these purposes can reap some benefits. However, for AI and ML to achieve the best results, more enterprises are leveraging foundation models and training them to match the context of specific industry or organizational needs.

Improved context is just one benefit of customized models. For instance, training a smaller foundation model with contextually specific content often leads to more accurate outcomes at a lower cost than using a generic off-the-shelf model.

AWS has both the hardware infrastructure and the tooling to create, train, and maintain these models. Its hardware infrastructure includes the latest GPUs, custom silicon, reliable compute clusters, and high-performance storage. Amazon SageMaker offers an end-to-end solution covering all phases of the model development lifecycle, including large-scale data preparation, training of foundation models, inference, and governance.

This combination of infrastructure and SageMaker provides data scientists and ML engineers more choice and faster time to market.

SageMaker and the Maturation of AI and ML

In the IT sphere, 2024 will be remembered as a pivotal year in the long history of AI. The rapid adoption of AI has been propelled by both general market awareness and a lowering of technology barriers. Although Amazon SageMaker has been part of the maturation of AI and ML technology since 2017, 2024 also saw a number of significant milestones in its history. As the year draws to a close, SageMaker offers:

- **More choice** — SageMaker offers more than 250 foundation models as well as support for deployment capabilities such as NVIDIA NIMs.
- **Enhanced productivity** — Adding Amazon Q Developer capabilities into SageMaker provides AI-based assistance to data scientists, leading to better and more consistent work products.
- **Better optimization** — New inference techniques are leading to approximately 2x higher throughput while reducing costs by about 50%.
- **Increased integration** — Integrations with the AWS Elastic Kubernetes Service (EKS) and governance tools such as DataZone help break down barriers between AI and ML workflows and other established IT processes.
- **Greater control** — SageMaker HyperPod gives organizations an end-to-end managed AI and ML infrastructure to further reduce training time and provide more granular control of underlying compute instances.

SageMaker HyperPod and Reducing Time to Market

Using SageMaker HyperPod is a great way to stand up AI and ML infrastructure without having to procure, build, deploy, and power hundreds or thousands of compute nodes. Key capabilities within HyperPod also allow for faster training. They include:

- **Resiliency and persistent clusters** — Training large models can extend into weeks or months in some cases. When that occurs, it's important that the work is completed on a resilient infrastructure because a single failure can halt the entire training process. A key capability of HyperPod is the training of models on a self-healing cluster, so in the event of a failure, new nodes can pick up seamlessly and without interruption. This capability alone can reduce training time by up to 20%, which is why some of the largest AI companies such as Hugging Face and Perplexity leverage HyperPod to train their models and reduce time to market.
- **Scale and distributed training** — As the use of models increases either in terms of larger datasets or adding new models, the ability to scale training workload deployments up to thousands of accelerators is a major concern. However, it's not just about increasing the scale of the infrastructure: Software-based scaling capabilities, such as distributed libraries and the ability to split models and training data across the cluster, can improve performance by as much as 20%.

- **Optimized cluster utilization** — SageMaker HyperPod clusters can be managed and orchestrated using Slurm or Amazon EKS to make it easy to scale across AI accelerators, including GPUs and AWS Trainium. In addition to providing the improved performance mentioned above, this also reduces the effort required from data scientists and IT operations resources.

OTHER BENEFITS OF SAGEMAKER HYPERPOD

Besides enabling quicker model training and deployment—with attendant cost savings—using HyperPod offers additional benefits. Providing management and governance tools to AI and ML ops teams leads to more optimized and streamlined deployments. For example, in addition to automatic management capabilities, SageMaker HyperPod supports manual customization via SSH into the cluster. Providing choice in terms of infrastructure and tooling to data science means more flexible solutions and minimum disruption for the team.

From an ML ops perspective, jobs can be scheduled and infrastructure can be shared among teams based on the needs of the business. Additionally, HyperPod now supports Amazon EKS to enable containerization and orchestration of AI and ML workloads in a more granular way. These capabilities allow teams to best optimize when and how work is completed such that it best meets the budgets, timing, and functional requirements of the organization.

Choice of infrastructure also creates significant advantages. HyperPod reaps all of the benefits of SageMaker's ability to work with over 250 models and a wide range of notebooks and IDEs. It also supports the addition of frameworks, containers, and debugging tools as the team requires. This level of flexibility enables teams to leverage a managed experience without change or disruption to their existing work patterns.

CONCLUSION

Amazon SageMaker HyperPod extends SageMaker's market-leading capabilities by mitigating the complexity and costs associated with AI and ML model development. HyperPod has been tested and proven by some of the biggest players in AI and ML and can be scaled to meet each customer's needs and technology requirements. Moor Insights & Strategy believes that organizations that are overwhelmed—or simply stalled—by the complex task of building and maintaining AI and ML infrastructure should know that SageMaker is well worth investigating as a solid option for addressing this challenge.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Jason Andersen, Vice President and Principal Analyst, Application Platforms, DevOps, OS

PUBLISHER

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

AWS commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.