GENERATIVE AI

Accelerate AI innovation with AWS infrastructure

Empower your company with purpose-built generative AI infrastructure that's secure, performant, and cost-optimized with the broadest and deepest choice of cloud capabilities from AWS

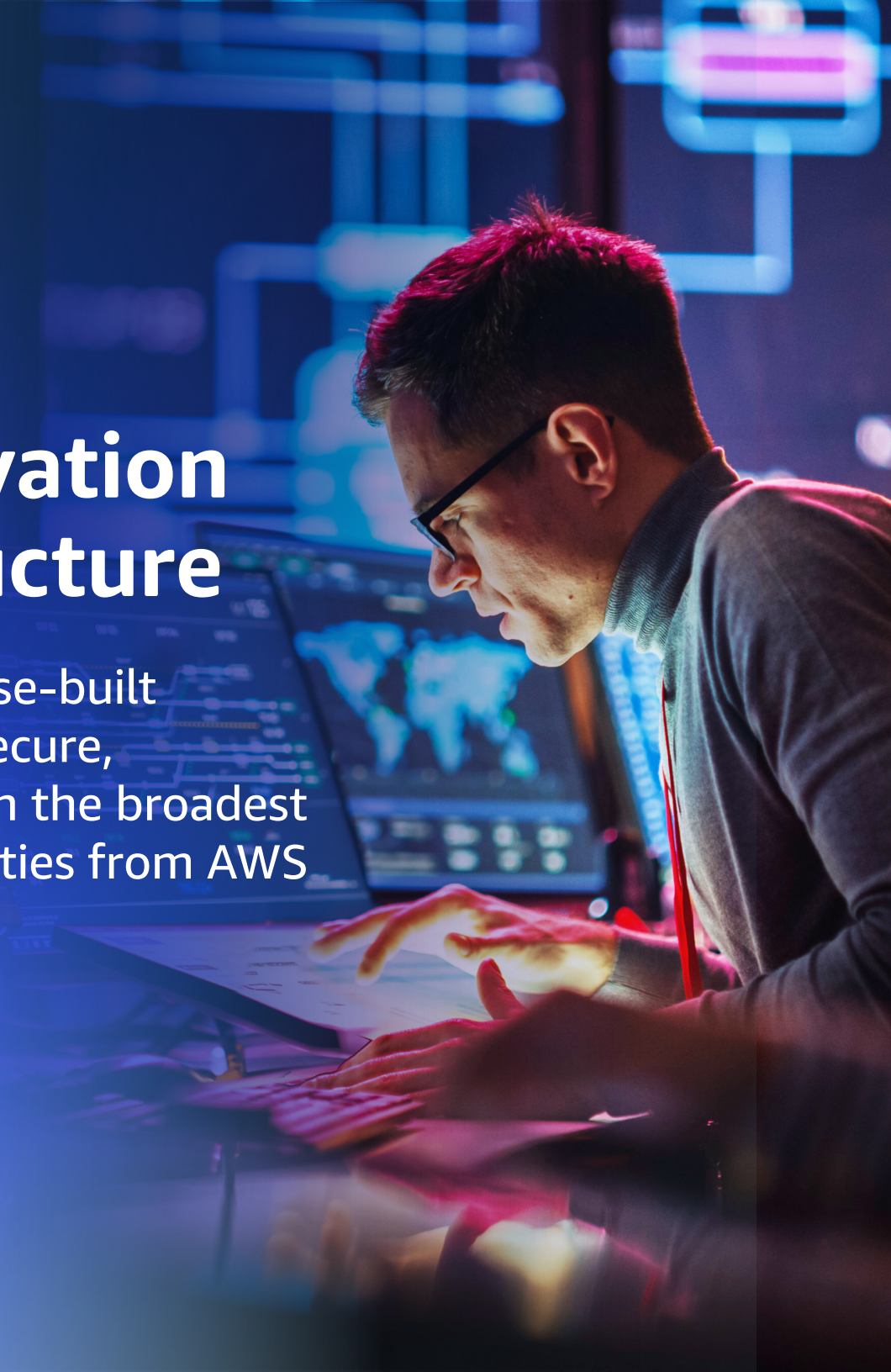


Table of contents

Introduction: The generative AI landscape 3

Challenges across the AI workflow 4

Developing your generative AI solution 6

How AWS helps you innovate with AI 8

Optimize your AI infrastructure 9

Protect data with a secure and private experience 11

Use preferred models and frameworks 12

Next steps 13

Introduction: The generative AI landscape

Generative AI foundation models (FMs) can create new content and ideas, including conversations, stories, images, videos, music, and even software code in response to a prompt. Generative AI is powered by large-scale FMs that can be trained with up to petabytes of data. As these FMs grow, their parameters also continue to increase—upwards of trillions of parameters today. Even a smaller language model can be trained with a few billion parameters, and depending on the use case, that number can go up to 15 billion parameters.

Organizations are looking to reshape industries with generative AI, from healthcare, to entertainment, to finance, to manufacturing. While most organizations are looking to take advantage of generative AI through customized applications that leverage large language models (LLMs) and other FMs, or with a fully managed service that allows them to use industry-leading models as a service, some still want the ability to build and train their own models.

As training and deploying these large-scale FMs continues to evolve, organizations need an unprecedented level of high-throughput, low-latency, and secure infrastructure to train these models in a reasonable time and deploy them for inference, while working to lower costs and maintaining the highest performance possible.

This eBook takes you through the key challenges in training and inferencing models and how the right infrastructure can optimize cost, improve performance, and reduce time to market.

The drive to innovate with AI

More than 30%

annual growth is forecasted for the GPU-accelerated cloud compute market from 2023 to 2035.[1]

8x more

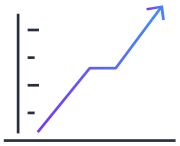
will be spent worldwide on generative AI projects in 2027 than was spent in 2023, according to IDC Research.[2]

[1] "GPU as a Service (GPUaaS) Market revenue to hit USD 30 Billion by 2035," Research Nester, January 2024

[2] "IDC Forecasts Spending on GenAI Solutions Will Double in 2024 and Grow to $151.1 Billion in 2027," IDC, December 2023

Challenges across the AI workflow

The range of infrastructure challenges organizations face are evolving as projects and technologies advance



Increased interest in rapidly adopting new technologies

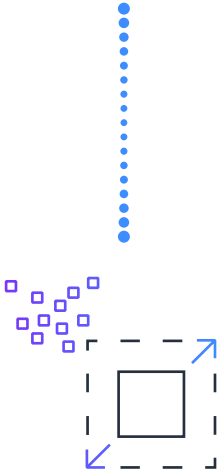
Today, generative AI integrations are being pursued at a fast pace, but organizations still need to thoughtfully address concerns over privacy, security, costs, performance, workforce knowledge and training gaps, and other impacts to avoid risks.

Working with models at any scale

While some FMs have grown astronomically to include trillions of parameters, many organizations are also using smaller and finely tuned models for their specific needs. Organizations need to be able to scale compute, networking, and storage flexibly, to meet diverse and changing requirements.

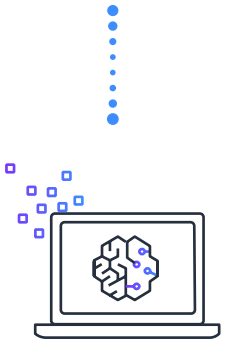
Balancing infrastructure costs while maintaining performance

Training, building, and deploying generative AI models requires an unprecedented level of performance and new technologies with budgets that remain similar year over year. This necessitates a need to find ways to lower costs while maintaining performance. So, you need a broad set of compute accelerators to meet the demand of any generative AI use case.



Data infrastructure modernization, integration, and scalability

Legacy systems inhibit advanced analytics, AI capabilities, and bring substantial capacity constraints requiring organizations to spearhead transformations that optimize value from the cloud. Plus, integrating generative AI systems into existing infrastructure and workflows can be complex and resource-intensive. While initial proof of concepts (PoCs) are easier to complete, scaling solutions and systems to handle increasing workloads and ensuring reliability and performance are not. Infrastructure should offer broad and flexible options to fit each scenario.

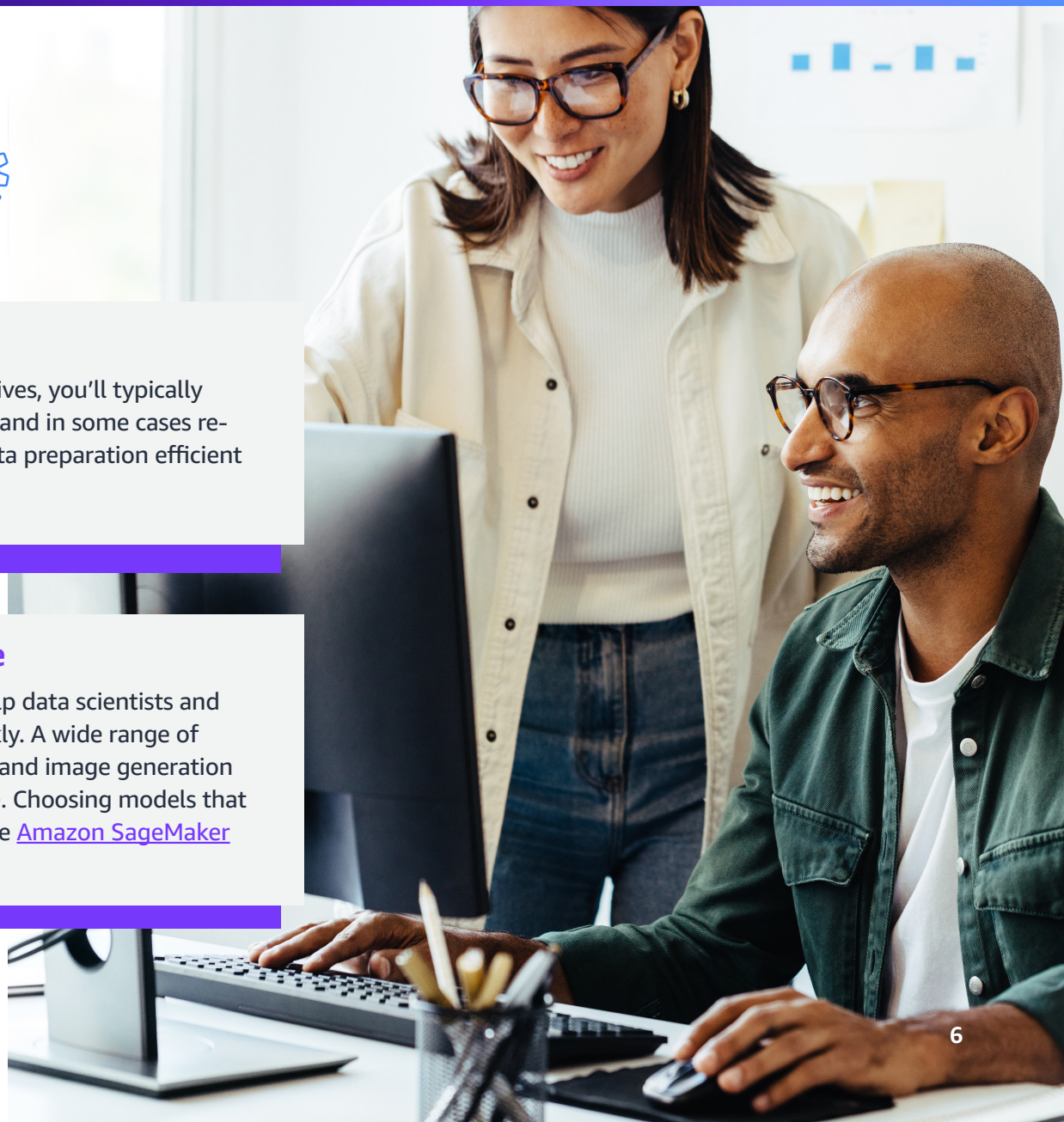


Data sovereignty, data residency, and regulatory considerations

Organizations in highly regulated industries are especially cautious about data security and privacy for generative AI applications, including concerns like exposure of intellectual property or code, data security and privacy, governance, and ensuring compliance. Organizations must navigate complex, uncertain, and ambiguous regulatory landscapes and ensure that their organizations comply with relevant laws and guidelines while exploring their cloud infrastructure solutions.

Developing your generative AI solution

5 steps for delivering generative AI projects into production



1 Data collection and preparation

After you've identified a use case and set objectives, you'll typically need to source large datasets, cleanse the data, and in some cases re-process it. You'll need scalable tools to make data preparation efficient and manageable.

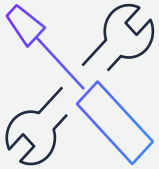
2 Selecting models and architecture

Pre-built models and solution templates can help data scientists and machine learning practitioners get started quickly. A wide range of publicly available and fine-tunable FMs for text and image generation are available from libraries such as HuggingFace. Choosing models that work with accelerated compute and ML tools like Amazon SageMaker can help you innovate faster.

3 Model training

Data is typically split into sets for training, validation, and testing. The model is trained through multiple runs in which weights are adjusted, problems are identified, and tracking metrics like model accuracy are refined. FMs are often trained on petabytes of data and may be too large to fit in a single GPU. You'll need purpose-built ML silicon or GPUs, in clusters with up to thousands of nodes. As a result, much of your training budget is likely to be spent on infrastructure. You'll also need access to the latest ML frameworks and libraries, plus performant and secure technologies that speed up networking and minimize latency.

4 Fine-tuning and optimizing models



Your compute capacity and resource needs will vary depending on the type of fine tuning or optimizations you choose—from full fine tuning to parameter-efficient fine tuning. You'll also need access to tools and software that help you maximize performance.

5 Deployment

As you prepare to deploy FMs for inference, your infrastructure needs will change. Inference can account for a large portion of the total cost of generative AI in production, so you'll need to implement infrastructure that reduces the inference cost at scale. Compute needs are also different from the training stage, because nodes can be distributed, rather than clustered. You may find complexities in achieving the low latency needed for real-time inference—required by interactive use cases like chatbots—or the throughput needed for batch inference of large datasets.

How AWS helps you innovate with AI

Choosing the right AWS infrastructure is essential for effectively building, training, and deploying AI models. AWS offers the broadest and deepest set of capabilities to help you build great solutions

AWS has been investing in purpose-built infrastructure to deliver more AI performance at lower costs, longer than any other provider. With the most complete range of infrastructure, services, and tools, you can meet the needs of every generative AI workflow and easily scale up or down as your needs change.

Innovate faster by leveraging the latest technologies and infrastructure that is optimized for high performance—helping you drive all of your generative AI and deep learning applications and business needs.

Leverage proven, purpose-built infrastructure with specialized ML silicon, networking, and storage, offering the right infrastructure you need for your generative AI workflows. Develop new generative AI applications and bring new use cases to the cloud previously not possible due to costs, performance, or capacity constraints.

Enable IT and operations teams to focus on innovation instead of reacting to security issues and protect highly sensitive data with greater ease.

AWS infrastructure delivers more

AWS has been developing custom silicon for more than 10 years. Starting with the AWS Nitro System to our Graviton CPUs to our Trainium and Inferentia AI chips.

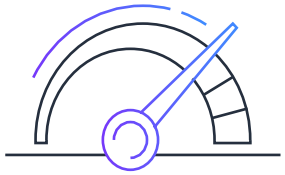
AWS infrastructure is 3.6 times more energy efficient than the median of surveyed U.S. enterprise data centers.

Benefit from our 12 years of collaboration

AWS was the first to bring NVIDIA GPUs to the cloud more than 12 years ago, and now AWS is partnering with NVIDIA to bring the world's largest AI supercomputer to the cloud (Project Ceiba)—with over 414 exaflops of compute performance.

Optimize your AI infrastructure

With AWS you can accelerate generative AI with an extensive choice of purpose-built infrastructure resources, including compute, network, and storage solutions. AWS purpose-built infrastructure is uniquely designed from the ground up to accelerate innovation, enhance security, and improve performance while lowering costs.



Accelerate compute at any scale

- **Achieve up to 50 percent savings on training costs** with Trainium-based Amazon EC2 Trn1 instances over comparable EC2 instances.
- **Achieve 40 percent lower inference costs** with Inferentia2-based EC2 Inf2 instances, versus comparable EC2 instances.
- **Achieve 4 times higher throughput and up to 10 times lower latency** with Amazon EC2 Inf2 instances powered by second-generation Inferentia2, compared to previous generation Inferentia-based instances.
- **Make cost savings of up to 40 percent** by using P5 instances over P4 instances.
- **Improve performance per watt by up to 50 percent** with EC2 instances powered by ML accelerators such as Trainium and Inferentia2.
- **Train models up to 40 percent faster** with Amazon SageMaker HyperPod.

Minimize latency with optimized networking

- **Enable lightning-fast inter-node communication for high-performance AI applications** with up to 3,200 Gbps of Elastic Fabric Adapter (EFA) networking, providing low-latency, high-bandwidth networking throughput.
- **Reduce latency by 16 percent and support up to 20,000 GPUs** with Amazon EC2 UltraCluster 2.0, a flatter and wider network fabric specifically optimized for ML accelerators. It offers up to 10 times more overall bandwidth than alternatives.
- **Increase network efficiency and optimize job scheduling** with the Amazon EC2 Instance Topology API. With insights into the proximity between your instances, it helps you strategically allocate each job to the instance type that best fits your requirements.

Optimize storage for throughput, low latency, and reduced costs

AWS offers a comprehensive choice of cloud storage options that meet every need in AI workflows, from delivering the performance needed to keep accelerators highly utilized to reducing the cost of long-term storage.

- **Amazon FSx for Lustre** can help you accelerate machine learning with maximized throughput to compute resources and seamless access to training data stored in Amazon S3.
- **Amazon S3 Express One Zone** provides the lowest-latency cloud object storage available, with data access speed up to 10 times faster and request costs up to 50 percent lower than Amazon S3 Standard.
- **Amazon S3** is built to retrieve any amount of data from anywhere, offering industry-leading scalability, data availability, security, and performance. Use S3 to create a centralized repository or data lake that allows you to store all your structured and unstructured data at any scale.

Protect data with a secure and private experience

AWS's top priority is safeguarding the security and confidentiality of workloads. Our approach to AI infrastructure security is built on the principles of completely isolating any data from risk

Control data and AI infrastructure securely

Built on a foundation of the AWS Nitro System, AWS safeguards even your most sensitive data. Nitro System is designed to enforce restrictions so that nobody, including anyone at AWS, can access your workloads or data running on your accelerated computing EC2 instances or any other Nitro-based EC2 instance.

The level of security protection offered is so critical that we've added it in our AWS Service Terms to provide an additional assurance to all of our customers and has been validated by the NCC Group, an independent cybersecurity firm.

Use preferred models and frameworks

A number of AI frameworks, SDKs, and model libraries have established themselves as fundamental to the development and deployment of today's AI systems. With AWS, your infrastructure is ready to easily support these frameworks with optimized performance.

- **AWS compute instances support major ML frameworks** such as TensorFlow and PyTorch. They also support model libraries and toolkits, such as Hugging Face, to enable a broad range of use cases. You can continue using your existing workflows and get started with only a few lines of code changes.
- **Accelerate deep learning in the cloud** with optimizations for ML frameworks and toolkits, which come pre-installed in the AWS Deep Learning AMIs (AWS DLAMIs) and AWS Deep Learning Containers (AWS DLCs).
- **Optimize machine learning on Trainium and Inferentia accelerators** with AWS Neuron, an SDK that supports libraries such as Megatron-LM.

"Amazon EC2 P4d instances offer 400 Gbps high-performance networking on EFA. The GPU-to-GPU networking speed directly impacts the ability to scale efficiently and remain cost effective when scaling to hundreds of GPUs."

Opher Lieber, Technical Lead for Jurassic, AI21 Labs

AI21 Labs trained a 178-billion-parameter large language model using Amazon EC2 P4d instances and PyTorch. The company scaled to hundreds of GPUs efficiently and cost effectively, supported distributed training and model parallelism on PyTorch, and built knowledge for developing models at scale.

Link to full story ›

Next steps

Generative AI offers unprecedented potential for innovation—but you need a secure, efficient, and highly capable foundation to truly make it work for your organization.

With AWS, your compute, networking, and storage infrastructure are primed for the future of AI. You can train models with purpose-built ML silicon, get projects up and running faster with comprehensive managed services, and ensure nobody can compromise the security of your AI data.

From pre-training to inference and beyond, no other platform helps you accelerate AI innovation like AWS.

Get started with generative AI infrastructure on AWS ›

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