Fast-track innovation with generative AI and machine learning

Delight your customers with products, services, and customer experiences powered by generative AI and machine learning

Introduction

No longer limited to global technology enterprises and data science specialists, machine learning (ML) has entered the mainstream. Thanks to the cloud, the barriers to widespread use of ML are rapidly disappearing. The cloud brings together data, low-cost storage, security, and ML services, along with high-performance, cost-effective CPU- and GPU-based compute instances with optimized software, which are essential to ML success. The cloud also offers a pay-as-you-go cost model that further enables customers to control costs.

More recently, complex deep learning models consisting of multiple layers of deep neural networks that take inspiration from how the human brain functions necessitate even more powerful compute resources. These advanced models require secure, scalable, and cost-effective CPUs coupled with powerful GPUs with GPU-optimized software and gigabytes or terabytes of storage.

With the cloud, you can either choose fully managed services that automatically manage your infrastructure so you don't need to worry about hardware and software maintenance, or you can opt for self-managed ML lifecycle management to benefit from the scale and the ability to customize infrastructure in a more hands-on way.

Whatever you choose, with the cloud, you don't need to invest in all possible options upfront. Resources are available on demand and are always up to date and ready to provide you with purpose-built ML tools, compute, storage, networking, and the latest infrastructure innovations.

Delight your customers with generative AI and machine learning

There are several common generative AI and machine learning use cases that help customers transform their business

Intelligent document processing

What it is:

Organizations typically have many documents, such as invoices, patient forms, loan applications, and contracts, that contain data, such as applicant names, entities (places or brands), or patient health history, that is essential to their business and requires processing. Intelligent document processing (IDP) applies ML models built using text processing algorithms to extract text from millions of documents, clarify the sentiment of or relationships between that data, and integrate a human step to validate, correct, or augment the ML results for accuracy and compliance. Additionally, with generative AI, users can easily summarize and extract key insights from all these documents.

How it's used:

IDP extracts data from digital documents to perform tasks like processing loan applications, analyzing customer sentiment, determining patient treatments, or filtering out noncompliant purchases from invoices.

The outcome:

ML-powered document processing results in higher accuracy of data and faster data processing. It can also lead to higher customer satisfaction rates, providing more accurate information and helping companies respond to requests faster and more appropriately.

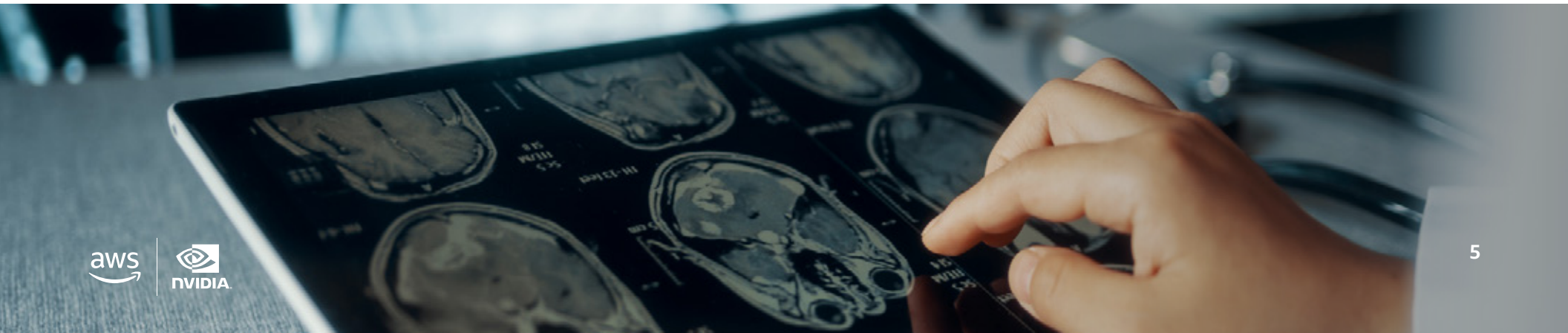
IDP boosts employee productivity, allowing workers to spend more time on business-critical tasks and less time wading through documents for insights and performing manual data entry. Automating document workflows reduces data extraction and analysis complexity, allowing organizations to dedicate less budget and resources to these labor-intensive approaches.

Customer success

Nearly 90 percent of US radiologists operate at or over capacity, according to a Mayo Clinic study—and **Rad AI** wants to help lighten their workloads. The company trains ML models to read detailed documents for radiologists and automatically summarize results, which physicians use to identify patient ailments and devise treatment plans. Rad AI chose to migrate its document summary applications from older GPU-based **Amazon Elastic Compute Cloud** (Amazon EC2) instances to the latest **Amazon EC2 P4d** instances powered by NVIDIA A100 Tensor Core GPUs. By deploying its application on Amazon EC2 P4d instances, Rad AI improved its ML inference times by 60 percent, delivering faster, more accurate reports to radiologists and improving patient outcomes.

Thomson Reuters is one of the world's most trusted providers of answers, with teams of experts who bring together information, innovation, and insights to unravel complex situations for organizations around the globe. Thomson Reuters has over 150 years of rich human-annotated data on law, tax, news, and other segments, and in 2018, the company chose **Amazon SageMaker** to accelerate its research and development efforts. Since deploying SageMaker, Thomson Reuters has been able to take advantage of such advanced capabilities as on-the-fly answer generation, long-text summarization, and fully interactive, conversational question answering. These capabilities enable Thomson Reuters to build comprehensive assistive artificial intelligence (AI) systems that can guide users toward the best solution for all their information needs.

By deploying its application on Amazon EC2 P4d instances, Rad AI improved its ML inference times by 60%, delivering faster, more accurate reports to radiologists and improving patient outcomes.



Computer vision

What it is:

Computer vision (CV) allows machines to identify people, places, and things in images with accuracy at or above human levels and with much greater speed and efficiency. CV automates extraction, analysis, classification, and understanding of useful information from a single image or a sequence of images. The image data can take many forms, such as single images, video sequences, views from multiple cameras, or three-dimensional data.

How it's used:

CV is used to help companies increase their brand reputation and increase safety by detecting inappropriate content in media and entertainment. In healthcare, CV applications analyze medical images and identify ones that require additional analysis, thereby improving patient outcomes. And in manufacturing, companies are using CV to automate defect detection of components on the assembly line.

Customer success

The University of Oxford houses 21 million objects in the collections of its **Gardens, Libraries & Museums (GLAM)**—artifacts and specimens that are among the world's most significant. Utilizing SageMaker, the University of Oxford is currently using CV powered by NVIDIA GPU-based **Amazon EC2 P3 instances** to build an enhanced image recognition system and accelerate the process of cataloging its extensive coin collection. Analyzing a coin, which previously took volunteers anywhere from 10 minutes to hours, is expected to take just a few minutes once the image recognition system is in place.[1]

Aerobotics is an agri-tech company operating in 18 countries around the world, based out of Cape Town, South Africa. Its mission is to provide intelligent tools to feed the world. It aims to achieve this by providing farmers with actionable data and insights on its platform, Aeroview, so it can make the necessary interventions at the right time in the growing season. Its predominant data source is aerial drone imagery: capturing visual and multispectral images of trees and fruit in an orchard.

Aerobotics uses SageMaker to improve its Tree Insights product, which provides per-tree measurements of important quantities like canopy area and health and provides the locations of dead and missing trees. Farmers use this information to make precise interventions like fixing irrigation lines, applying fertilizers at variable rates, and ordering replacement trees.

Analyzing a coin, which previously took volunteers anywhere from 10 minutes to hours, is expected to take just a few minutes once the image recognition system is in place.[1]

[1] "University of Oxford Introduces a Sector-Leading Image Recognition ML Prototype to Augment Digitization in Numismatics," AWS, 2021

Personalized recommendations

What it is:

Consumers today expect real-time, curated experiences across digital channels as they consider, purchase, and use products and services. ML algorithms can be used to scale and create personalized customer experiences tailored to individual preferences and behaviors across channels.

How it's used:

Organizations can build applications capable of delivering a wide array of personalized experiences, including specific product recommendations, personalized product re-ranking, and customized direct marketing powered by generative AI. With ML-based personalization, organizations can go beyond rigid, static rules and use recommendation systems that deliver highly personalized recommendations to customers.

The outcome:

ML can help organizations deliver highly personalized experiences—resulting in improvements in customer engagement, conversion, revenue, and margin—and create differentiation in a digital world.

Customer success

Multinational ecommerce company **Zalando** decided to standardize its ML workloads in the cloud to improve customer experiences, boost productivity, and push the needle in its business. With SageMaker, Zalando can steer campaigns better, generate personalized outfits, and deliver more engaging customer experiences—while increasing engineer and data scientist productivity by 20 percent.[2]

NerdWallet is a personal finance startup that provides tools and advice that make it easy for customers to pay off debt, choose the best financial products and services, and tackle major life goals like buying a house or saving for retirement. The company relies heavily on data science and ML to connect customers with personalized financial products. Previously, NerdWallet would provide customers with a list of potential credit cards they might like, but it had no way to forecast the likelihood of acceptance. With SageMaker, the company can more effectively match customers to the right financial products for them.

The use of SageMaker and Amazon EC2 P3 instances with NVIDIA V100 Tensor Core GPUs has also improved NerdWallet's flexibility and performance and has reduced the time required for data scientists to train ML models from months to days.

With SageMaker, Zalando can steer campaigns better, generate personalized outfits, and deliver more engaging customer experiences—while increasing engineer and data scientist productivity by 20%.[2]

[2] AWS Customer Success Story: Zalando, AWS

Generative AI

What it is:

Generative AI is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. Like all AI, generative AI is powered by ML models—very large models that are pretrained on vast amounts of data and commonly referred to as foundation models (FMs).

How it's used:

You can take advantage of ML for your business quickly and apply it to a broader set of use cases with generative AI. Apply generative AI across all lines of business, including engineering, marketing, customer service, finance, and sales.

Use generative AI to improve customer experience through capabilities such as chatbots, virtual assistants, intelligent contact centers, and personalization. You can also boost your employees' productivity with generative AI-powered conversational search, content creation, text summarization, and code generation, among others.

Improve business operations with intelligent document processing, maintenance assistants, quality control and visual inspection, and synthetic training data generation. Finally, you can use generative AI to turbocharge production of all types of creative content, from art and music to text, images, animations, and video.

Use cases for generative AI

The use cases for and possibilities of generative AI span all industries and individuals. Here are the most popular applications to date, grouped by industry:

Life Sciences

- **Creating novel protein sequences:** Accelerate drug discovery and research by creating sequences with specific properties for the design of antibodies, enzymes, vaccines, and gene therapy
- **Designing synthetic gene sequences:** Healthcare and life sciences companies can use AI-generated gene sequences for applications in synthetic biology and metabolic engineering, such as creating new biosynthetic pathways or optimizing gene expression for biomanufacturing purposes

Healthcare

- **Creating synthetic patient and healthcare data:** With simulated datasets, organizations can train AI models, simulate clinical trials, or study rare diseases—even when access to real-world data is unavailable or impractical—while complying with the strict security and privacy requirements of the industry
- **Improving patient experience:** Generative AI can personalize patient discharge instructions and treatment plans. Conversational assistants and chatbots can reduce clinician workload, increase patient satisfaction, and help provide proactive healthcare to at-risk communities

Financial Services

- **Improving experiences:** Financial services firms can better serve customers and employees by deploying chatbots that resolve problems faster, personalizing products and recommendations, and automating internal tasks—while still delivering the strong data encryption and privacy controls the industry requires
- **Increasing knowledge-worker efficiency:** Knowledge workers at financial firms can process applications faster, achieve deeper insights into customer behavior, improve collaboration, and deploy powerful training programs and simulations
- **Analyzing market sentiment:** Through faster and more thorough analysis of social media, news articles, and financial data, financial services firms can surface market commentary, identify opportunities sooner, and proactively mitigate risks

Media and Entertainment

- **Speeding up content creation:** From storyboarding and concepting to post-production workflows, media and entertainment companies can automate lower-level tasks to increase production speed and allow creative talent to iterate faster and realize the director's vision
- **Improving music:** Artists can complement and enhance their albums with AI-generated music to create whole new genres
- **Aiding the media supply chain:** Generative AI applications can aid or automate tasks like localization, content moderation, and even content restoration

Education

- **Summarizing text:** Students and teachers can create concise summaries of research documents, lecture transcripts, and class notes to make them easier to search and browse
- **Automating content creation:** AI can transform information into sample test questions, accelerate grading, measure student performance across a wide range of factors, and provide personalized feedback and recommendations to teachers and students
- **Personalizing learning environments:** Educators can create personalized learning pathways for student segments—or even individual students—and leverage simulations and virtual reality to make learning more engaging

Automotive and Manufacturing

- **Improving product design:** Manufacturers can use AI to optimize the design of mechanical parts—or create entirely new material, chip, and part designs—improving quality and durability, lowering costs, and simplifying production
- **Personalizing in-vehicle experiences:** Virtual assistants and personalized route recommendations can enhance experiences for drivers and passengers
- **Testing and maintaining:** AI can improve product testing by generating information missing from datasheets—and unlock new assisted maintenance use cases to better maintain and service machinery, including products in use by consumers
- **Improving overall equipment effectiveness for factories:** Digitize and capture historical machine maintenance data, repair data, equipment manuals, production data, and potentially data from other manufacturers to generate suggestions for maintenance, repairs, or equipment parameters—and to improve productivity, availability, and quality

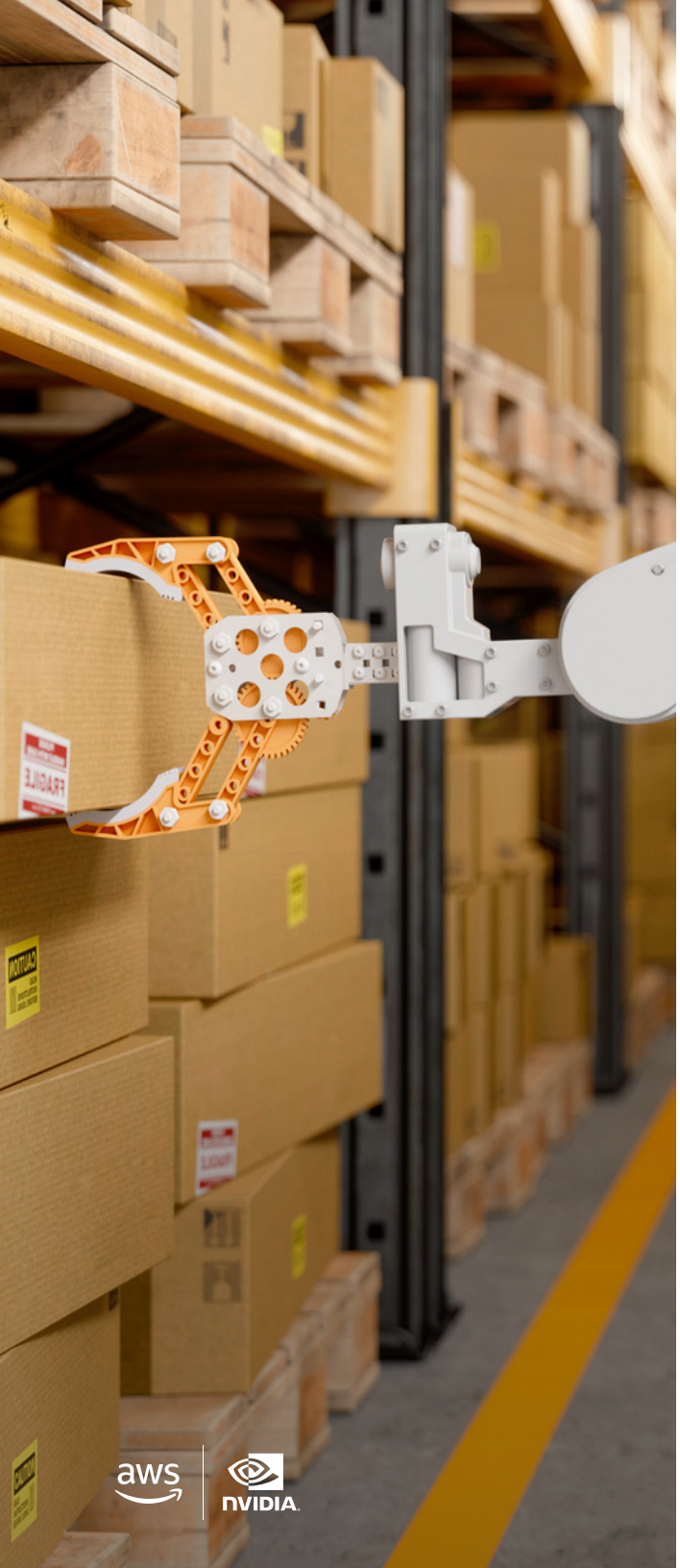
With the rapid growth and rising business value of generative AI, the number of use cases and applications for this transformative technology will only increase over time.

Data augmentation

Generative AI's ability to create synthetic data, made possible by a process called label-efficient learning, is proving useful across all modalities and use cases. Synthetic data can be used to train AI models when real-world data is nonexistent, restricted, or simply unable to address corner cases with high accuracy. Foundation models used to create generative AI applications can reduce labeling costs by either automatically producing additional augmented training data or by learning an internal representation of the data that facilitates training AI models with less labeled data.

Use cases on the rise

In addition to these everyday cases, we've observed several emerging cases quickly gaining momentum



Autonomous systems lower costs and increase productivity

Autonomous systems use many different ML models to sense their environment and operate without human intervention. Autonomous systems rely on sensors, actuators, complex algorithms, ML systems, and powerful processors to execute software quickly.

Robots are one example of autonomous systems. As a purveyor of cutting-edge technologies, **Amazon Robotics** has long known that using AI and ML to automate key aspects of the fulfillment process represented extraordinary potential gains—so in 2017, it devoted teams to accomplishing just that. As the company iterated on its ML project, it turned to Amazon Web Services (AWS) and SageMaker, a managed service that helps data scientists and developers prepare, build, train, and deploy high-quality ML models quickly. This freed the Amazon Robotics team from the difficult task of standing up and managing a fleet of NVIDIA GPUs for running inferences at scale across multiple regions. As of January 2021, the solution saved the company nearly 50 percent on ML inferencing costs and unlocked a 20 percent improvement in productivity with comparable overall savings.

Saving lives with predictive healthcare

Timely diagnosis of severe medical conditions is often delayed due to insufficient data on a patient or because there's not enough time to analyze and correlate large patient datasets. **CloudMedx** is helping to address this challenge by developing ML models that understand how different diseases, symptoms, and medications are related to each other to help predict disease progression and determine the likelihood that a patient may have a complication.[3]

Machine learning to improve wastewater management

Wastewater management is vital for public health—and for conserving one of the most precious resources on earth. **Opseyes** leveraged AI to develop the first rapid microscopy test for wastewater treatment plants. This test allows users to instantly check plant conditions, quickly remedy threats of contamination, and avoid downtime.[4]

The power of machine learning for musicians

Sunhouse, an AI startup founded by musicians and technologists, is creating an ML-based system to empower drummers to turn drum sets into entire production suites for touring, composing, and jamming. By using custom drum sensors and AI-powered acoustic mapping, Sunhouse's technology makes drums into an expressive tool for composing and performing with samples, effects, and midi. Sensory Percussion gives drummers the ability to control electronics with their drumsticks, opening up a completely new avenue for creativity and music making. Since its inception, Sunhouse's solution has become widely used by live performers in the New York jazz community and many of today's leading drummers, including Marcus Gilmore and Wilco's Glenn Kotche.[5]

[3] Darrow, B., "This AI Software {CloudMedx} Aims to Conquer Heart Disease," CloudMedx, October 2016
[4] Caufield, B., "In the Drink of an AI: Startup Opseyes Instantly Analyzes Wastewater," NVIDIA Blog, February 2021
[5] Finkle, L., "Drum Roll, Please: AI Startup Sunhouse Founder Tlacael Esparaza Finds His Rhythm," NVIDIA Podcast, March 2021

Solutions from AWS and NVIDIA

AWS and NVIDIA have collaborated for over 12 years to continually deliver powerful, cost-effective, and flexible GPU-based solutions for customers. **NVIDIA GPU-based Amazon EC2 instances** deliver the high-performance, cost-optimized infrastructure needed to train ML and deep learning models quickly and accurately. With the latest NVIDIA A100 Tensor Core GPU-powered **Amazon EC2 P4d instances**, developers can reduce the time to train their models from days to minutes. You can run the most complex multi-node training at high efficiency and optimized cost. This enables you to experiment, train, and tune your models quickly to accelerate innovation.

These instances are the first in the cloud to support 400 Gbps instance networking. Amazon EC2 P4d instances provide an average of 2.5 times better performance for deep learning models compared to previous generation Amazon EC2 P3 and Amazon EC2 P3dn instances.

Amazon EC2 P4d instances are also deployed in hyperscale clusters called EC2 UltraClusters, comprised of the highest-performance compute, networking, and storage in the cloud. Each EC2 UltraCluster is one of the most powerful supercomputers globally, enabling you to run your most complex multi-node ML model training. You can quickly scale from a few to thousands of NVIDIA A100 Tensor Core GPUs in EC2 UltraClusters based on ML project needs.

Amazon EC2 G5 instances are the latest generation of NVIDIA GPU–based instances that can be used for a wide range of ML use cases. They deliver up to three times better performance for ML inference and up to 3.3 times higher performance for ML training compared to Amazon EC2 G4dn instances.

Customers can use G5 instances to get high-performance and cost-efficient infrastructure to train and deploy larger and more sophisticated models for natural language processing (NLP), CV, and recommender engine use cases.

G5 instances feature up to eight NVIDIA A10G Tensor Core GPUs and second-generation AMD EPYC processors. They also support up to 192 vCPUs, up to 100 Gbps of network bandwidth, and up to 7.6 terabytes of local NVMe SSD storage.

Available in the AWS Marketplace, **NVIDIA AI Enterprise** is a library of full-stack software, including over 100 frameworks, pretrained models, AI workflows, and infrastructure optimization. It streamlines the development and deployment of production-ready applications for generative AI, speech AI, vision AI, cybersecurity, and more. NVIDIA AI Enterprise addresses the complexities of organizations building and maintaining their own high-performance, secure, cloud-native AI software platform by offering continuous monitoring for vulnerabilities with regular and timely patching of critical and common vulnerabilities and exposures (CVEs), API stability, and enterprise support with SLAs and access to NVIDIA AI experts.

Start leveraging powerful ML infrastructure now with Amazon SageMaker

The easiest and fastest way to benefit from powerful ML infrastructure on AWS is to deploy **Amazon SageMaker**. This fully managed service brings together a broad set of essential capabilities, such as data labeling, data preparation, feature engineering, statistical bias detection, AutoML, training, tuning, hosting, explainability, monitoring, and workflows.

Amazon SageMaker JumpStart is an ML hub offering algorithms, models, and ML solutions. With SageMaker JumpStart, customers can discover, explore, and deploy open-source FMs that are not available in **Amazon Bedrock** such as OpenLLaMA, RedPajama, Mosaic MPT-7B, FLAN-T5/UL2, GPT-J-6B/Neox-20B, and Bloom/BloomZ. We are continuously adding more models and have doubled our FMs available this year alone. For customers who want to create their own FMs, Amazon SageMaker provides managed infrastructure and tools to accelerate scalable, reliable, and secure model building, training, and deployment.

The NVIDIA AI platform on AWS—which includes NVIDIA GPU–accelerated instances, NVIDIA GPU–optimized software, and NVIDIA AI Enterprise—can help developers significantly accelerate the development of ML algorithms in SageMaker. The platform enables developers to run model training and inference faster, deploy ML applications sooner, and realize cost savings.

To further accelerate AI model deployment and lower inference costs, SageMaker has integrated **NVIDIA Triton Inference Server**. This enables features that maximize the performance of both CPU and GPU instances on AWS—such as multi-framework support, dynamic batching, and concurrent model execution.

To achieve these benefits, NVIDIA offers NVIDIA AI Enterprise in the **AWS Marketplace**—a comprehensive collection of GPU-optimized libraries, pretrained AI models for CV, conversational AI and recommenders, application frameworks, and inference-serving solutions like Triton Inference Server—to simplify the deployment of ML models on CPUs and GPUs. The software in NVIDIA AI Enterprise is constantly optimized, secure, and supported by NVIDIA, allowing developers to easily access the latest NVIDIA innovations. Regular releases give users access to the latest features and performance improvements.

Put machine learning to work for your business

By running ML workloads in the cloud, enterprises get on-demand access to the most powerful GPU instances and ML tools that can be spun up in minutes, scale from one to thousands of instances, and keep infrastructure costs under control.

NVIDIA-powered AWS services and infrastructure are available for organizations of all experience levels—from those that are seasoned in building ML workloads and want to manage their infrastructure to those that prefer a fully managed approach. AWS supports your organization with compute, networking, storage, and ML tools across each step of the ML development lifecycle, including collecting and preparing data, choosing the right algorithm, tuning the model for maximum accuracy, and deploying and monitoring model performance and quality over time.

Learn what's possible with AWS and NVIDIA ›

Get started with Amazon SageMaker ›