



Accelerate machine learning innovation with the right cloud services and infrastructure

Easily prepare data, and build, train, and deploy machine learning applications



Table of contents

Innovate with machine learning	3
Achieve success with AWS machine learning	5
Accelerate every step of the machine learning lifecycle.....	6
Step 1: Prepare data quickly and easily	7
Step 2: Build accurate models across multiple frameworks	9
Step 3: Train models faster and at a lower cost	11
Step 4: Deploy models quickly and cost-effectively	14
Build on a solid foundation for machine learning success	17

Innovate with machine learning

Thanks to advancements in computing power, the decreasing price of storage, and the prevalence of cloud computing, artificial intelligence (AI) and machine learning (ML) have entered the mainstream. Organizations and industries of varying sizes—including those in finance, retail, fashion, real estate, healthcare, and many more—can leverage AI and ML to deliver a wide range of business benefits. These include acquiring new and deeper insights about customers, identifying and responding to cyberthreats, making smarter data-driven decisions, and improving hiring processes.

Because of these benefits, more organizations are making investments in AI and ML. IDC forecasts global spending on AI and ML will increase at a compounded annual growth rate (CAGR) of 26.5 percent from 2022 to 2026, jumping from \$118 billion in 2022 to more than \$300 billion in 2026.¹

ML models are used for many use cases, such as natural language processing (NLP), computer vision (CV), and document processing, which learn from existing data through a process called training to make decisions about new data through a process called inference. Today's large-scale ML models, known as foundation models (FMs), contain hundreds of billions of parameters that are incredibly powerful general-purpose models that can be extended and customized for specific use cases without having to build a model from scratch each time. These models power generative AI applications, such as text summarization, code generation, and image creation, based on natural language prompts.

Some of today's most popular algorithms include:

- **Natural language processing (NLP)** - NLP algorithms analyze language at scale, with the ability to understand context, parse speech, and perform translations in near real time. They are used to create ML applications, such as chatbots, spam filters, voice assistants, and social media monitoring tools.
- **Computer vision (CV)** - CV algorithms process and analyze visual data to detect objects and classify images in ways similar to the human mind—but at exponentially greater speed and scale. They can be used to improve workplace safety, support digital identity verification, and flag inappropriate content.
- **Document processing** - Document processing algorithms extract text, handwriting, and data from documents, going beyond optical character recognition (OCR) to identify and understand data from forms and tables. They can be used to extract information from medical records and automate the processing of financial documents.
- **Generative AI** – FMs, such as large language models (LLMs) and diffusion models, can be used to generate original human-like content, such as coherent prose, images, and videos based on natural language prompts. These can be used for applications such as code generation, text summarization, question answering, and image and video generation.

The potential business value of these applications is substantial—but so are the resource and infrastructure requirements needed to operate them at speed and scale. Training ML models that power these use cases requires large amounts of data, tens of thousands of compute nodes, and enhanced inter-node and intra-node networking.

In response to these requirements, a growing number of organizations are looking to the cloud. The cloud brings together data, low-cost storage, security, and ML services along with high performance computing (HPC) infrastructure for model training and deployment.

How AWS accelerates machine learning success

More ML happens on Amazon Web Services (AWS) than anywhere else, and AWS offers the broadest and deepest portfolio of services to accelerate business transformation. Organizations of all sizes, from Fortune 500 to startups, are benefiting from the ideal combination of high-performance and low-cost ML infrastructure and services from AWS. By running their ML workloads in the cloud, customers get on-demand access to infrastructure and ML tools that can be spun up in minutes, scale from one instance to thousands of instances, and only pay for what they use.

Let's take a look at some examples of AWS customers that are driving results with ML today.



Achieve success with AWS machine learning

Tens of thousands of customers have chosen AWS ML to help them realize a wide variety of business results. Here are a few examples:

- **LG AI Research** developed EXAONE, an FM that contains 300 billion parameters. EXAONE was built using **Amazon SageMaker** to complete a wide range of tasks across different industries, such as fashion, manufacturing, research, education, and finance. Using the FM, they developed an AI artist called Tilda, which collaborated with a fashion designer to generate 3,000 images and patterns to design more than 200 outfits for New York Fashion Week 2022. By using SageMaker, LG AI Research reduced costs by approximately 35 percent and increased data processing speed by about 60 percent.
- **NerdWallet** provides tools and advice that make it easy for customers to manage their finances. The company relies heavily on data science and ML to connect customers with personalized financial products. NerdWallet uses a number of AWS services, such as SageMaker and **Amazon Elastic Compute Cloud (Amazon EC2) P3 instances**, to improve performance and reduce the time required for data scientists to train and iterate on ML models from months to just days.
- **Sprinklr** provides a unified customer experience management (Unified-CXM) platform that combines different applications for marketing, advertising, research, customer care, sales, and social media engagement. Sprinklr's Unified-CXM platform uses ML algorithms on unstructured data sourced from many different channels to deliver sentiment and intent insights to its customers. For example, the company's NLP and CV ML models analyze different data formats sourced from social media posts, blog posts, video content, and other content available on public domains across more than 30 channels. With **Amazon EC2 Inf1 instances**, which are powered by **AWS Inferentia**, a high-performance ML inference accelerator, Sprinklr was able to reduce latency by 30 percent. Getting started was easy, and the team is now able to deploy a model using Amazon EC2 Inf1 instances in under two weeks.

Accelerate every step of the machine learning lifecycle

Businesses turn to AWS to break down the barriers across every step of the ML lifecycle. There are four major steps in the ML lifecycle. At each step, ML developers need to support ML governance by, for example, creating policies and establishing controls to ensure model transparency, data privacy, and security.

1. Data science teams need to prepare example data to train a model
2. Then, they need to select which algorithm or framework they will use to build the model
3. Next, models need to be trained to make predictions and tuned frequently to achieve the highest accuracy
4. Finally, models need to be deployed—integrated with their applications, monitored, scaled, and managed in production

AWS offers your choice of infrastructure at every step of the ML workflow. You can customize your infrastructure—including compute, networking, and storage—to fit your performance and budget requirements. You have a broad range of options for high-performing, cost-effective, and scalable infrastructure.

AWS offers the highest-performing ML infrastructure powered by GPUs and purpose-built ML accelerators **AWS Trainium** and **AWS Inferentia**.

AWS Trainium enables up to 50 percent savings on training costs over comparable Amazon EC2 instances. And AWS Inferentia2 enables up to 70 percent better price performance over comparable Amazon EC2 instances.

The easiest and fastest way to use the **AWS ML infrastructure** is SageMaker, a fully managed service that brings together a broad set of capabilities such as data labeling, data preparation, feature engineering, statistical bias detection, automatic machine learning (AutoML), training, tuning, hosting, explainability, monitoring, and workflows. **Amazon SageMaker JumpStart** provides hundreds of built-in algorithms, pretrained FMs, and pre-built solutions that customers can deploy with just a few clicks.

AWS Neuron SDK also makes it easy to extract the full performance of AWS Trainium and AWS Inferentia accelerators by integrating natively with popular ML frameworks, such as PyTorch and TensorFlow. Customers can continue using their existing frameworks and application code when using the Amazon EC2 Trn1n, Trn1, Inf2, and Inf1 instances based on these accelerators.

Customers can also use **AWS Deep Learning Containers** (docker images preinstalled with deep learning frameworks) with **Amazon Elastic Kubernetes Service** (Amazon EKS) and **Amazon Elastic Container Service** (Amazon ECS). In addition, the **AWS Deep Learning AMIs** (DLAMI) provide preconfigured environments to build deep learning applications quickly by providing ML practitioners and researchers with the infrastructure and tools needed to accelerate deep learning in the cloud at any scale.

Now that you have a general idea of how the ML development process works—and how AWS can help—let's dive into each of the four stages in greater detail.

STEP 1

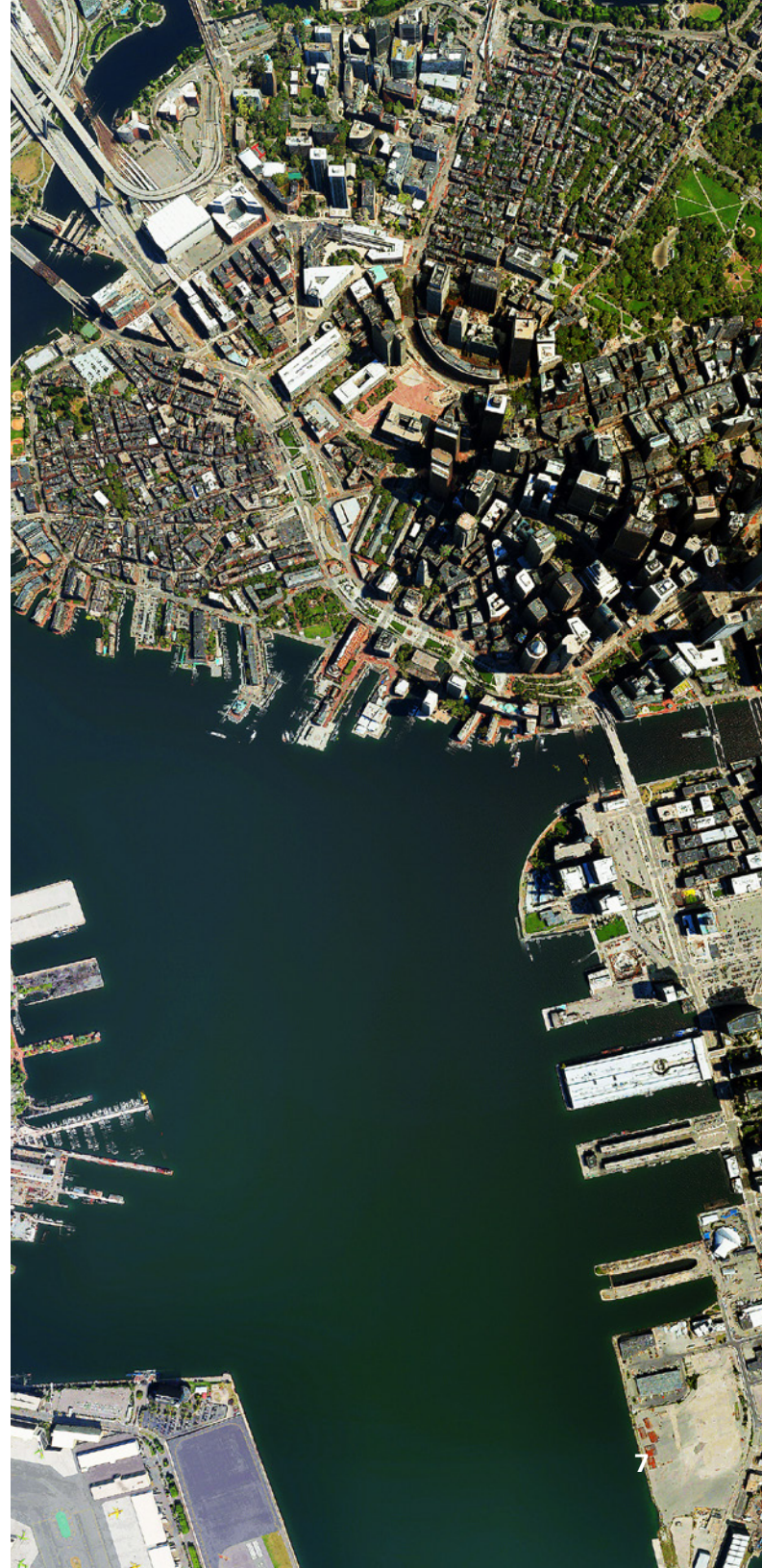
Prepare data quickly and easily

The challenge

Data is the fuel for ML. But even with the right data strategy in place, managing data can be the most time consuming and challenging part of building ML models. Many customers say they spend about 80 percent of their time on data preparation tasks, such as data collection, cleansing, and labeling.

There are two types of data: structured and unstructured. Structured data is highly organized quantitative data easily decipherable by ML. However, structured data makes up only a small percentage of all data. Unstructured data is qualitative, including things like images, handwritten notes, and geospatial data. It's extremely valuable but much harder to use for ML. Most of the insights for ML are embedded in unstructured data, but unstructured data analysis is often beyond the capabilities of many existing data management tools; for example, when a physician needs to analyze information from X-rays, MRIs, and written prescriptions.

Further complicating things, most ML engineering teams need to write code for common data preparation tasks needed for ML—or integrate with standalone extract, transform, and load (ETL) frameworks managed by other organizations.



The solution

SageMaker helps process both structured and unstructured data. **Amazon SageMaker Ground Truth Plus** helps customers easily create high-quality training datasets without having to build labeling applications or manage labeling workforces. SageMaker Ground Truth Plus also helps reduce data labeling costs by up to 40 percent and meet your data security, privacy, and compliance requirements. You simply upload your data, and SageMaker Ground Truth Plus creates data labeling workflows and manages workflows. For geospatial data, ML practitioners can access geospatial data sources, purpose-built processing operations, pretrained ML models, and built-in visualization tools to run geospatial ML faster and at scale.

For structured data, **Amazon SageMaker Data Wrangler** drastically simplifies the preparation of structured data with a no-code visual interface. SageMaker Data Wrangler contains more than 300 built-in data transformations, so you can quickly normalize, transform, and combine features without having to write any code. With the SageMaker Data Wrangler visualization templates, you can quickly preview and inspect that these transformations were completed as you intended by viewing them in **Amazon SageMaker Studio**, the first fully integrated development environment (IDE) for ML. You can also simplify your data workflows with a unified notebook environment for data engineering, analytics, and ML. Create, browse, and connect to **Amazon EMR** clusters and AWS Glue Interactive Sessions directly from SageMaker Studio notebooks. Monitor and debug Spark jobs using familiar tools, such as Spark UI, right from the notebooks. Use the built-in data preparation capability powered by SageMaker Data Wrangler directly from the notebooks to visualize data, identify data quality issues, and apply recommended solutions to improve data quality and model accuracy without writing a single line of code.

Once your data is prepared, you can build fully automated ML workflows with **Amazon SageMaker Pipelines** and save them for reuse in the **Amazon SageMaker Feature Store**.



With Amazon SageMaker Data Wrangler, we can now interactively select, clean, explore, and understand our data effectively, empowering our data science team to create feature engineering pipelines that can scale effortlessly to datasets that span hundreds of millions of rows... with Amazon SageMaker Data Wrangler, we can operationalize our ML workflows faster.”²

Caleb Wilkinson, Lead Data Scientist, INVISTA

STEP 2

Build accurate models across multiple frameworks

The challenge

Once you have training data available, you need to choose an ML algorithm with a learning style that meets your needs. This can be difficult, as there are dozens of algorithms to choose from. ML frameworks, such as PyTorch and TensorFlow, make development easier—but they are typically best suited for specific algorithms. This often results in the need to manage and build across a mix of algorithms and frameworks, which can be complex, error-prone, and resource intensive.

Building models also requires lots of experimentation and iteration. Most teams use Jupyter Notebooks to build models and share work across teams. Unfortunately, as more models are developed, sharing work and scaling become more difficult.

The solution

If you want to use pre-built algorithms and a fully managed service to build efficient, accurate, and powerful ML models, SageMaker is the solution for you. SageMaker includes a dozen pre-built algorithms that can be deployed on the framework of your choice. Using SageMaker Studio, you can build models in a single visual interface, which can improve data science team productivity by up to 10 times.³

SageMaker Studio gives you complete access, control, and visibility as you train your model. You can quickly upload data, create new notebooks, and adjust ML experiments. All ML development activities—including notebooks, experiment management, automatic model creation, debugging, and model and data drift detection—can be performed within SageMaker Studio.

Amazon SageMaker Studio notebooks manage compute instances to view, run, or share a notebook. The underlying compute resources are fully elastic, so you can easily dial the available resources up or down, and the changes take place automatically in the background without interrupting your work. You can also share notebooks with others in a few clicks. They will get the exact same notebook, saved in the same place.

If you prefer to use AutoML to build your models, **Amazon SageMaker Autopilot** automatically builds, trains, and tunes the best ML models based on your data. You can also use SageMaker JumpStart to quickly and easily bring ML applications to market. With SageMaker JumpStart, you can access built-in algorithms with pretrained models from model hubs, pretrained FMs to help you perform tasks such as article summarization and image generation, and pre-built solutions to solve common use cases. In addition, you can share ML artifacts, including ML models and notebooks, within your organization to accelerate ML model building and deployment.

Accelerate the time to deploy for more than 150 open-source models, including one-click deployable ML models and algorithms from popular model zoos. Get started with just a few clicks and easily bring ML applications to market using pre-built solutions and FMs pretrained on terabytes of text and image data. You can perform a wide range of tasks, such as article summarization and text, image, or video generation, which are preconfigured with all necessary AWS services required to launch into production, including an **AWS CloudFormation** template and reference architecture.

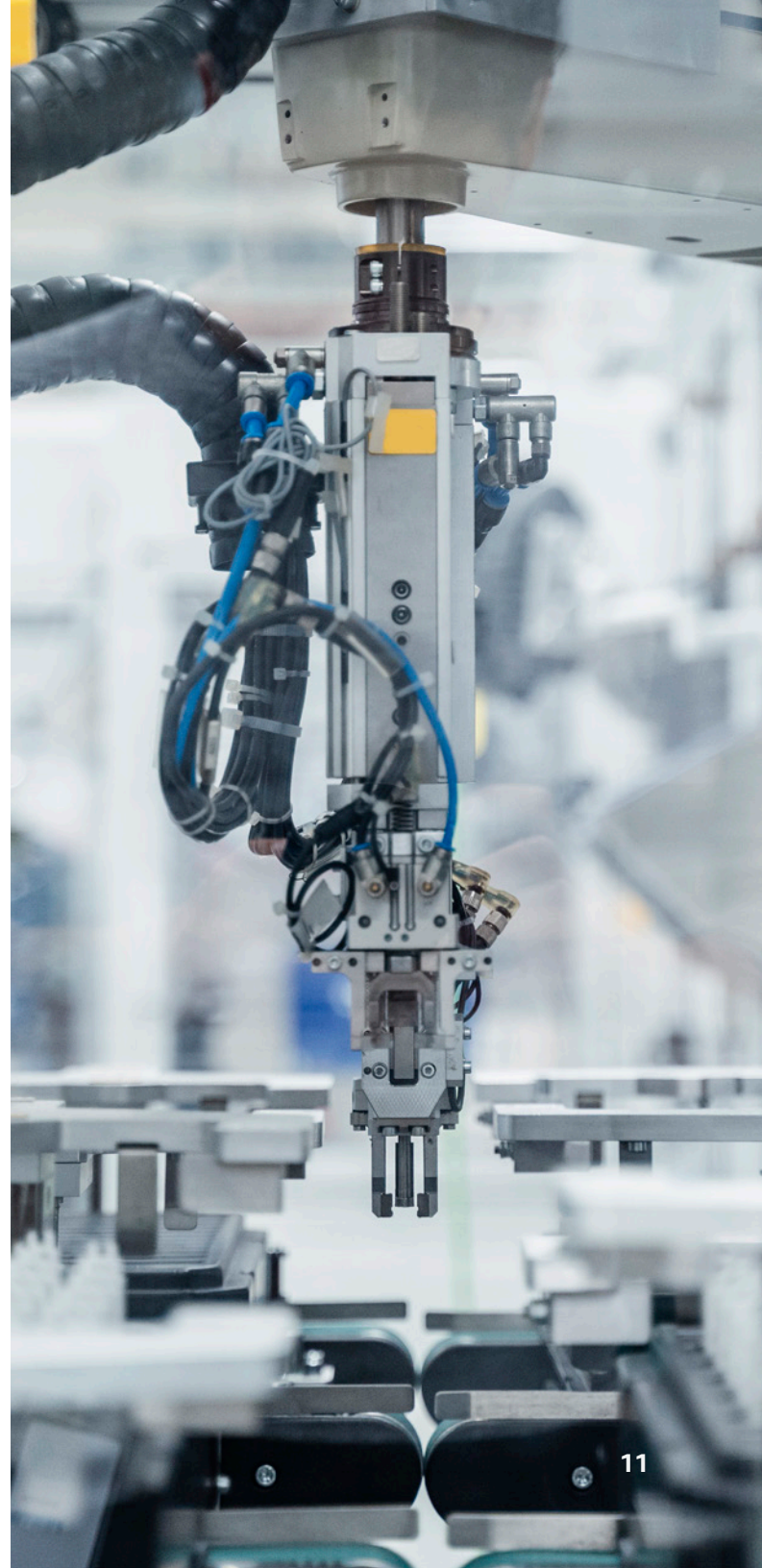
STEP 3

Train models faster and at a lower cost

The challenge

After your models are built, data science teams train the model on their datasets so the models are ready to make accurate predictions on new data. Training is an iterative process, and over time, models need to be retrained and tuned to account for new data or model drift. As deep learning becomes pervasive, models are becoming increasingly complex. Model complexity is doubling every two years, and many bleeding edge models now have a trillion parameters. Training and tuning these large models is computationally intensive and, at times, cost prohibitive.

As we push the boundaries of ML model performance and capability, the time and cost needed to train models will only continue to grow. This expanding drain on resources can prohibit your organization from taking full advantage of what ML has to offer, slowing innovation and jeopardizing executive support for your ML investments.





Through the use of Amazon EC2 P4d instances, we were able to reduce our training time for object recognition by 40% compared to previous generation GPU instances without any modification to existing codes.”⁴

Junya Inada, Director of Automated Driving (Recognition), TRI-AD



The solution

AWS offers high-performance, cost-effective ML infrastructure for ML training. Choose from the Amazon EC2 range of CPU, GPU, and purpose-built accelerator-based instances to fit the requirements of your ML training use cases. Customers can use [Amazon SageMaker Model Training](#) to take advantage of this infrastructure without the need to manage it.

AWS Trainium – Amazon EC2 Trn1 instances, powered by AWS Trainium, deliver the highest performance on deep learning training of NLP models. They deliver up to 50 percent savings on training costs over comparable Amazon EC2 instances. These instances support up to 1600 Gbps (Trn1n) of Elastic Fabric Adapter (EFA) network bandwidth. They are deployed in Amazon EC2 UltraClusters that enable scaling up to 30,000 AWS Trainium accelerators, which are interconnected with a non-blocking petabit-scale network to provide up to 6.3 exaflops of compute. Customers can use Trn1 instances to train NLP, CV, and recommender models across many applications, such as text summarization, recommendation, and image and video generation.

NVIDIA GPUs – AWS offers a broad range of NVIDIA GPU-based instances. [Amazon EC2 P4d instances](#) are the most performant GPU-based instances for deep learning training. They are well suited to train the most complex multi-node ML models with high efficiency. Amazon EC2 P3 instances are ideal when you need to train medium-to-large models and for single-node distributed training use cases. [Amazon EC2 G5 instances](#) deliver up to a 15 percent lower cost to train than Amazon EC2 P3 instances.

Intel/Habana Gaudi – Amazon EC2 DL1 instances powered by Gaudi accelerators from Habana Labs (an Intel company) are specifically designed for training deep learning models. These instances deliver up to 40 percent better price performance than comparable Amazon EC2 instances and are well suited for NLP and CV use cases.

⁴ [Amazon EC2 P4d Instances](#), March 2023

The solution

Refer to the chart below to compare AWS infrastructure options optimized for ML training and tuning.

Instance type	Maximum chips per instance	Type of accelerator	Network bandwidth	Storage	Extra features
Amazon EC2 Trn1	16 AWS Trainium Accelerators	AWS Trainium	1600 Gbps EFA (Trn1n) 800 Gbps EFA (Trn1)	8 TB NVMe	Can be deployed on Amazon EC2 UltraClusters comprised of more than 30,000 AWS Trainium accelerators, high-speed networking, and high-throughput, low-latency storage Supports popular ML frameworks with AWS Neuron SDK
Amazon EC2 P4d	8 GPU A-100	NVIDIA	400 Gbps EFA, GPU-Direct RDMA	8 TB NVMe	Can be deployed on Amazon EC2 UltraClusters comprised of more than 4,000 GPUs, high-speed networking, and high-throughput, low-latency storage
Amazon EC2 P3	8 GPU Tesla V100	NVIDIA	100 Gbps, EFA	1.8 TB NVMe	Supports all major ML frameworks
Amazon EC2 DL1	8 Gaudi Accelerators	Habana Labs, Intel	400 Gbps, ENA	8 TB NVMe	Supports popular ML frameworks with Habana SynapseAI SDK

SageMaker reduces the time and cost to train and tune ML models using built-in tools to manage and track training experiments, automatically choose optimal hyperparameters, debug training jobs, and monitor the utilization of underlying system resources and network bandwidth. SageMaker can automatically scale infrastructure up or down based on your training job requirements, from one accelerator to thousands or from terabytes to petabytes of storage. And because you pay only for what you use, you can manage your training costs more effectively.

To train deep learning models faster, you can use the [Amazon SageMaker Training Compiler](#) to accelerate the model training process by up to 50 percent through graph- and kernel-level optimizations that make more efficient use of accelerators. Moreover, you can add either data parallelism or model parallelism to your training script with a few lines of code, and the SageMaker distributed training libraries will automatically split models and training datasets across Amazon EC2 instances to help you complete distributed training faster.

STEP 4

Deploy models quickly and cost-effectively

The challenge

Once you've trained and optimized your model to your desired level of accuracy and precision, it's time to put the model into production to make predictions. This is known as the prediction or inference step of ML.

A model that takes several hundred milliseconds to generate text translations, apply filters to images, or generate product recommendations can make an app feel sluggish or frustrating to use, driving users away. By speeding up inference, you can reduce the overall app latency and deliver a smooth experience.

Up to 90 percent of the infrastructure cost for developing and running an ML application is spent on inference—making the need for high-performance, low-cost ML inference infrastructure critical.⁵

The solution

AWS offers a breadth of high-performance, cost-effective, and easy-to-use instances for ML inference. For highly sophisticated models, such as LLMs or diffusion models, **Amazon EC2 Inf2 instances** powered by AWS Inferentia2 are the best option. Inf2 instances deliver up to 40 percent better price performance, up to three times higher throughput, and up to eight times lower latency over comparable Amazon EC2 instances. **Amazon EC2 Inf1 instances**, powered by first-generation AWS Inferentia, are well suited for smaller NLP and vision models. They deliver up to 70 percent lower cost and 2.3 times higher throughput than comparable Amazon EC2 instances.

Customers that wish to continue using the NVIDIA ecosystem for their inference due to model, framework, or operator support can leverage **Amazon EC2 G5 instances** for high-performance inference. If you are looking for inference for models that take advantage of Intel AVX-512 Vector Neural Network Instructions, **Amazon EC2 C5 instances** can help speed up typical ML operations, such as convolution, and automatically improve inference performance over a wide range of deep learning workloads.

Use the chart below to compare AWS infrastructure options optimized for ML inference.

Instance type	Maximum accelerators per instance	Type of hardware	Network bandwidth	Storage	Extra features
<u>Amazon EC2 Inf2</u>	12 AWS Inferentia2 Accelerators	AWS Inferentia2	100 Gbps	40 Gbps of EBS Bandwidth	Distributed inference with high-speed connectivity between accelerators; well suited for ultra-large models with hundreds of billions of parameters Supports popular ML frameworks with <u>AWS Neuron SDK</u>
<u>Amazon EC2 Inf1</u>	16 AWS Inferentia Accelerators	AWS Inferentia	100 Gbps	19 Gbps of EBS Bandwidth	Supports popular ML frameworks with <u>AWS Neuron SDK</u>
<u>Amazon EC2 G5</u>	8 NVIDIA A10G Tensor Core GPUs	NVIDIA	100 Gbps	7.6 NVMe	Supports all major frameworks and NVIDIA libraries
<u>Amazon EC2 C5</u>	96 vCPUs	Intel AVX	25 Gbps	4 x 900 NVMe SSD	Built on Nitro

The solution

SageMaker helps you take advantage of the above-mentioned broad selection of ML infrastructure and provides model deployment options to help meet your needs, whether real time or batch. Once you deploy a model, SageMaker creates persistent endpoints to integrate into your applications to make ML predictions. It supports the entire spectrum of inference, from low latency (a few milliseconds) and high throughput (hundreds of thousands of inference requests per second) to long-running inference for use cases, such as NLP. Whether you bring your own models and containers or use those provided by AWS, you can implement MLOps best practices using SageMaker to reduce the operational burden of managing ML models at scale.

For use cases with intermittent and unpredictable usage patterns, **Amazon SageMaker Serverless Inference** allows you to deploy ML models on pay-per-use pricing without worrying about servers or clusters. When deploying your model, simply select the serverless option, and SageMaker automatically provisions, scales, and turns off compute capacity based on the volume of inference requests, so you don't need to manage complex scaling policies and forecast traffic demand upfront.

Amazon SageMaker Inference Recommender helps you choose the best available compute instance and configuration to deploy ML models for optimal inference performance and cost. SageMaker Inference Recommender automatically selects the compute instance type, instance count, container parameters, and model optimizations for inference to maximize performance and minimize cost.

SageMaker model deployment features are natively integrated with MLOps capabilities, including Amazon SageMaker Pipelines (workflow automation and orchestration), **Amazon SageMaker Projects** (templates for standardizing developer environments for data scientists and continuous integration and

continuous delivery [CI/CD] systems for MLOps engineers), SageMaker Feature Store (feature management), **Amazon SageMaker Model Registry** (model and artifact catalog to track lineage and support automated approval workflows), **Amazon SageMaker Clarify** (bias detection), and **Amazon SageMaker Model Monitor** (model and concept drift detection).

As a result, whether you deploy one model or tens of thousands, SageMaker helps offload the operational overhead of deploying, scaling, and managing ML models while getting them to production faster.



We launched a large-scale AI chatbot service on the Amazon EC2 Inf1 instances and reduced our inference latency by 97% over comparable GPU-based instances while also reducing costs. As we keep fine-tuning tailored NLP models periodically, reducing model training times and costs is also important. Based on our experience from successful migration of inference workload on Inf1 instances and our initial work on AWS Trainium-based EC2 Trn1 instances, we expect Trn1 instances will provide additional value in improving end-to-end ML performance and cost.”⁶

Takuya Nakade, CTO, Money Forward, Inc.

Build on a solid foundation for machine learning success

The right choice of services and infrastructure can substantially enhance the performance of your ML workloads—you can prepare data for ML faster, be equipped to reliably build sophisticated models, train the models quickly and at scale, and deploy them in powerful, cost-efficient ways. Whether you're offloading the bulk of development to a fully managed service, creating models from scratch, or anything in between, the right services and infrastructure can help you complete ML projects faster and achieve greater results.

AWS offers the ideal combination of high-performance and low-cost infrastructure and services optimized for ML. By running your ML workloads in the cloud, you will get on-demand access to infrastructure and ML tools that can spin up instances in minutes and scale to thousands of instances—while only paying for what you use.

Get started with ML ›