aws

Accelerate machine learning innovation with the right cloud services and infrastructure

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

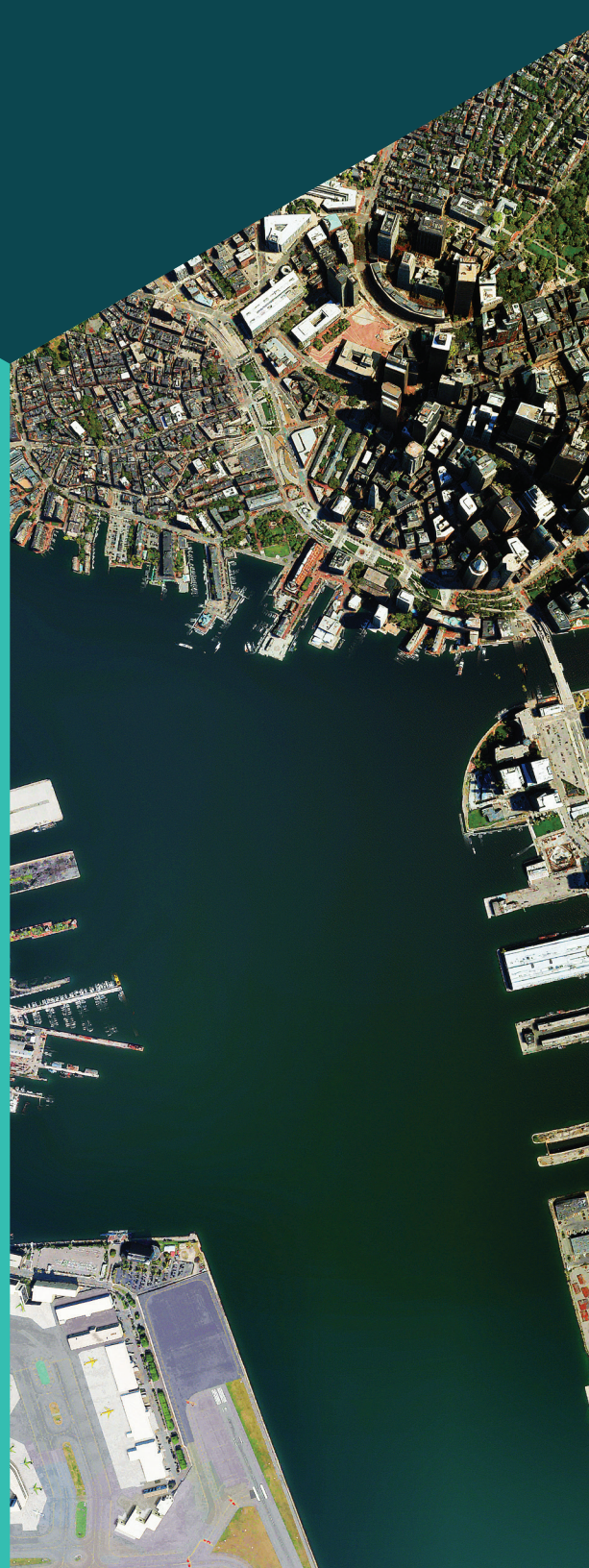


Table of contents

Innovate with machine learning 3

Achieve success with AWS Machine Learning 5

Accelerate every step of the ML lifecycle 6

Prepare data quickly and easily 7

Build accurate models across multiple frameworks 9

Train models faster and at lower cost 12

Deploy models quickly and cost-effectively 15

Build on a solid foundation for ML success 18

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.[1]

Because of the benefits, more organizations are making investments in AI and ML. In fact, IDC predicts that global spending on AI will reach $110 billion by 2024.[2]

One of the reasons ML is increasing in use is because it delivers deeper insights into data. ML works by using computational algorithms, such as natural language processing (NLP), computer vision, and document processing, that learn from existing data, through a process called training, to make decisions about new data, through a process called inference.

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** – Computer vision 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, enable 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, understand, and extract data from forms and tables. They can be used to extract information from medical records and automate processing of financial documents.

[1] "State of AI in the Enterprise, 3rd Edition." Deloitte, 2020.
[2] "Worldwide Artificial Intelligence Spending Guide." IDC, 2020.

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 enable these use cases requires large amounts of data, tens to thousands of compute nodes, and enhanced inter/intra node networking.

In response to these issues, 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 compute infrastructure for model training and deployment.

How Amazon Web Services (AWS) delivers ML 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 increasingly adopting AWS because AWS offers the ideal combination of high-performance and low-cost infrastructure services and machine learning services optimized for ML. 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 to thousands of instances, and only pay for what they use.

Let's take a look at some examples of AWS customers who 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 just a few examples:

- **The National Football League** (NFL) has partnered with AWS to create Next Gen Stats, a program that engages fans by combining old and new data to reveal insights about game dynamics. There are now over 20 unique Next Gen Stats that are used on-screen during games in real time—by broadcasters, partners, fans, and more. The most exciting stats use predictive ML models built on **Amazon SageMaker** to highlight athleticism like never before.
- **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 Amazon SageMaker and Amazon Elastic Compute Cloud (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.
- **Freddy's Frozen Custard & Steakburgers**, a fast-casual restaurant chain headquartered in Wichita, Kansas, turned to data science to find a better way to evaluate the quality of its restaurants. Using Domo AutoML, powered by SageMaker Autopilot, Freddy's IT team was able to derive business value from ML in just weeks—something that took months previously. With ML tools at the ready, the team used 5x larger datasets for more accurate predictions, ramped up quickly with ML modeling, and proved their hypothesis around staffing.

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

Accelerate every step of the ML lifecycle

Businesses turn to AWS because we systematically break down the barriers to ML across each step of the lifecycle. There are four major steps in the ML lifecycle:

1. Developers need to prepare example data to train a model.
2. Then, they need to select which algorithm or framework they'll use to build the model.
3. Models next 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 the choice of a broad range of options for high-performing, cost-effective, and scalable infrastructure.

The easiest and fastest way to use this infrastructure is **Amazon SageMaker**, a fully managed service that brings together a broad set of capabilities such as labeling, data preparation, feature engineering, statistical bias detection, AutoML, training, tuning, hosting, explainability, monitoring, and workflows.

Customers can also use the **AWS Deep Learning Containers** (AWS DL Containers). AWS DL Containers are Docker images pre-installed with deep learning frameworks that make it easy to deploy custom ML environments quickly by letting you skip the complicated process of building and optimizing your environments from scratch. In addition, the **AWS Deep Learning AMIs** provide pre-configured environments to quickly build deep learning applications by providing ML practitioners and researchers with the infrastructure and tools to accelerate deep learning in the cloud at any scale. You can quickly launch Amazon EC2 instances pre-installed with popular deep learning frameworks and interfaces such as TensorFlow, PyTorch, Apache MXNet, Chainer, Gluon, Horovod, and Keras to train sophisticated, custom AI models, experiment with new algorithms, or learn new skills and techniques.

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.

Prepare data quickly and easily

The challenge

Preparing data for use in an ML model is a time-consuming process, and many customers say they spend about 80 percent of their time on data preparation tasks, like data collection and cleansing. Data preparation has always been considered tedious and resource-intensive due to the inherent nature of data being “dirty” and not ready for ML in its raw form. “Dirty” data could include missing or erroneous values, outliers, etc.

Stand-alone data preparation tools help ease the process, but their value is limited. While most of these tools provide data transformation, feature engineering, and visualization, few offer built-in model validation. They treat data preparation as an extract, transform, load (ETL) workload, making it tedious to iteratively prepare data, validate the model on test datasets, deploy it in production, and then go back to ingesting new data sources and performing additional feature engineering.

Further complicating things, most ML engineering teams need to write code for common data preparation tasks needed for ML—or integrate with stand-alone ETL frameworks that are managed by other organizations.

The solution

Amazon SageMaker Data Wrangler reduces the time it takes to aggregate and prepare data for ML from weeks to minutes. With Amazon SageMaker Data Wrangler, you can simplify the process of data preparation and feature engineering and complete each step of the data preparation workflow, including data selection, cleansing, exploration, and visualization from a single visual interface. Using Amazon SageMaker Data Wrangler's data selection tool, you can choose the data you want from various data sources and import it with a single click.

Amazon SageMaker Data Wrangler contains over 300 built-in data transformations so you can quickly normalize, transform, and combine features without having to write any code. With Amazon SageMaker Data Wrangler's visualization templates, you can quickly preview and inspect that these transformations are completed as you intended by viewing them in Amazon SageMaker Studio, the first fully integrated development environment (IDE) for ML.

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…(and) operationalize our ML workflows faster."***[3]
>
> **Caleb Wilkinson,** Lead Data Scientist, INVISTA

[3] Amazon SageMaker Data Wrangler

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 TensorFlow, PyTorch, and Apache MXNet 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, Amazon SageMaker is the solution for you. Amazon SageMaker includes a dozen pre-built algorithms that can be deployed on the framework of your choice. Using **Amazon SageMaker Studio**, you can build models in a single, visual interface, which can improve data science team productivity by up to 10 times.[4]

Amazon 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 Amazon 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 up or down the available resources, 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 automatic machine learning (AutoML) to build your models, **Amazon SageMaker Autopilot** automatically builds the best machine learning models based on your data. You can also use **Amazon SageMaker JumpStart** to quickly and easily bring ML applications to market. Solutions can be used out of the box or customized for a specific business problem, eliminating the need to train an ML model and string together different components of an ML application.

Accelerate time-to-deploy for over 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, which are preconfigured with all necessary AWS services required to launch into production, including an AWS CloudFormation template and reference architecture.

[4] Amazon SageMaker Studio product page. Accessed May 11, 2021.

AWS is the cloud of choice for popular ML frameworks. Today, 92 percent of cloud-based TensorFlow and 91 percent of cloud-based PyTorch run on AWS[5] and AWS participates in the community to add new functionality to these frameworks. For example, TorchServe, now the default model-serving library on PyTorch, was built and is maintained by AWS in partnership with Facebook.

> ***Amazon SageMaker is an important component of enabling Cerner to deliver on our intent to deliver value for our clients through AI/ML. Additionally, Amazon SageMaker provides Cerner with the ability to leverage different frameworks like TensorFlow and PyTorch, as well as the ability to integrate with various AWS services."***[6]

Dr. Tanuj Gupta, MD, Vice President, Cerner Intelligence

[5] Barr, Jeff. "AWS re:Invent 2020 Machine Learning Keynote." AWS, 2020
[6] Amazon SageMaker customers

Train models faster and at lower cost

The challenge

After your models are built, training and tuning is the next step of the workflow. With the emergence of more complex ML applications, advancements that we never thought possible, like autonomous systems or machines that understand speech, are now a reality. Unfortunately, these innovations require complex, compute-intensive training and tuning—with some models having hundreds of billions of parameters.

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 ML investments at your company.

The solution

AWS offers a range of CPU, GPU, and custom accelerator-based Amazon EC2 instances that fit the requirements of your ML use cases. For use cases such as conversational AI and video tagging that have high memory and network performance requirements, Amazon EC2 P4d and Amazon EC2 P3 instances are ideal.

Amazon EC2 P4d instances are the highest-performing deep learning training instances on AWS. They deliver 2.5 times better performance and 60 percent lower costs to train ML models compared to previous generation Amazon EC2 P3 instances—so you can train the most complex multi-node ML models with high efficiency.

Amazon EC2 P3 instances deliver high-performance and cost-effective deep learning training and are ideal when you need to train medium-to-large models and for single node distributed training use cases.

AWS not only offers the highest-performing, most cost-effective infrastructure but is also rapidly investing in infrastructure to keep up with the changing needs of customers. And AWS is introducing new chips and instances to lower the cost of training. **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 current GPU-based instances and are ideal for natural language processing and computer vision use cases.

> ***"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."***[7]
>
> **Junya Inada,** Director of Automated Driving (Recognition), TRI-AD

[7] Amazon EC2 P4d Instances

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 hardware	Network bandwidth	Storage	Extra features
Amazon EC2 P4d	8 GPU A-100	NVIDIA	400 Gbps EFA, GPU-Direct RDMA	8 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	Support all major ML frameworks
Amazon DL1	8 Gaudi Accelerators	Habana Labs, Intel	400 Gbps, ENA	8 TB NVMe	**Habana SynapseAI SDK**

Amazon SageMaker is the fastest and easiest way to take advantage of these instances. The service offers one-click training by setting up a distributed compute cluster, performing the training, and tearing down the cluster when complete. Amazon SageMaker also offers the easiest methods for training large deep learning models and datasets. Using partitioning algorithms, Amazon SageMaker's distributed training libraries automatically split large deep learning models and training datasets across AWS GPU instances in a fraction of the time it takes to do manually.

Amazon SageMaker achieves these efficiencies through two techniques: data parallelism and model parallelism. Model parallelism splits models too large to fit on a single GPU into smaller parts before distributing across multiple GPUs to train, and data parallelism splits large datasets to train concurrently in order to improve training speed. **Amazon SageMaker Debugger** also helps you optimize ML models by capturing training metrics in real-time such as data loss during regression and sending alerts when anomalies are detected. This helps you immediately rectify inaccurate model predictions, such as an incorrect identification of an image.

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.[8]

[8] Amazon EC2 Inf1 Instances. Accessed May 11, 2021.

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 computer vision and NLP, the best option is **Amazon EC2 Inf1** instances. Amazon EC2 Inf1 instances, powered by AWS Inferentia, are built from the ground up by AWS and deliver up to 70 percent lower cost and 2.3 times higher throughput per inference than Amazon EC2 GPU-powered instances.

Customers that wish to continue using the NVIDIA ecosystem for their inference due to model, framework, or operator support can leverage **Amazon EC2 G4dn** 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 like 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 chips per instance	Type of hardware	Network bandwidth	Storage	Extra features
Amazon EC2 Inf1	16 AWS Inferentia Chips	AWS Inferentia	100 Gbps	19 Gbps of EBS Bandwidth	**AWS Neuron SDK**, the software supports all leading ML frameworks to migrate models onto Amazon EC2 Inf1 instances with minimal code changes.
Amazon EC2 G4dn	8 GPUs	NVIDIA T4 GPUs	100 Gbps, EFA	1.8 GB NVMe	NVIDIA CUDA libraries
Amazon EC2 C5	96 vCPUs	Intel AVX	25 Gbps	4 x 900 NVMe SSD	Built on Nitro

The solution

As with training, Amazon SageMaker offers the fastest and easiest way to use these instance types. You can one-click deploy your model onto auto-scaling Amazon ML instances across multiple availability zones for high redundancy. Amazon SageMaker will launch the instances, deploy your model, and set up the secure HTTPS endpoint for your application. In addition, Amazon SageMaker provides software innovation to further optimize infrastructure in a scalable and cost-effective way. **Amazon SageMaker multi-model** endpoints enable you to deploy multiple models with a single click on a single endpoint—and then serve them using a single serving container.

Unfortunately, the accuracy of ML models can deteriorate over time, a phenomenon known as model drift. Many factors can cause model drift such as changes in model features. The accuracy of ML models can also be affected by concept drift, the difference between data used to train models and data used during inference. Amazon SageMaker Model Monitor helps you maintain high quality ML models by detecting model and concept drift in real-time and sending you alerts so you can take immediate action.

As a fully managed service, Amazon SageMaker provides access to infrastructure with 99.99 percent public service availability SLA. It takes care of infrastructure setup, management, patching, and updates. The service makes it easy to deploy your trained model into production with a single click—so you can start generating predictions for real-time or batch data. You can one-click deploy your model onto auto-scaling Amazon ML instances across multiple availability zones for high redundancy. Amazon SageMaker will launch the instances, deploy your model, and set up the secure HTTPS endpoint for your application.

Amazon SageMaker helps you scale your model deployment to hundreds of thousands of models using **Amazon SageMaker Pipelines**. Since it is purpose-built for machine learning, Amazon SageMaker Pipelines helps you automate different steps of the ML workflow, including data loading, data transformation, training and tuning, and deployment. With Amazon SageMaker Pipelines, you can build dozens of ML models a week and manage massive volumes of data, thousands of training experiments, and hundreds of different model versions. You can share and re-use workflows to recreate or optimize models, helping you scale ML throughout your organization.

> ***"Autodesk is advancing the cognitive technology of our AI-powered virtual assistant, Autodesk Virtual Agent (AVA) by using (AWS) Inferentia. Piloting Inferentia, we are able to obtain a 4.9 times higher throughput over (Amazon EC2) G4dn for our NLU (natural language understanding) models, and we look forward to running more workloads on the Inferentia-based (Amazon EC2) Inf1 instances."***[9]
>
> **Binghui Ouyang,** Sr Data Scientist, Autodesk

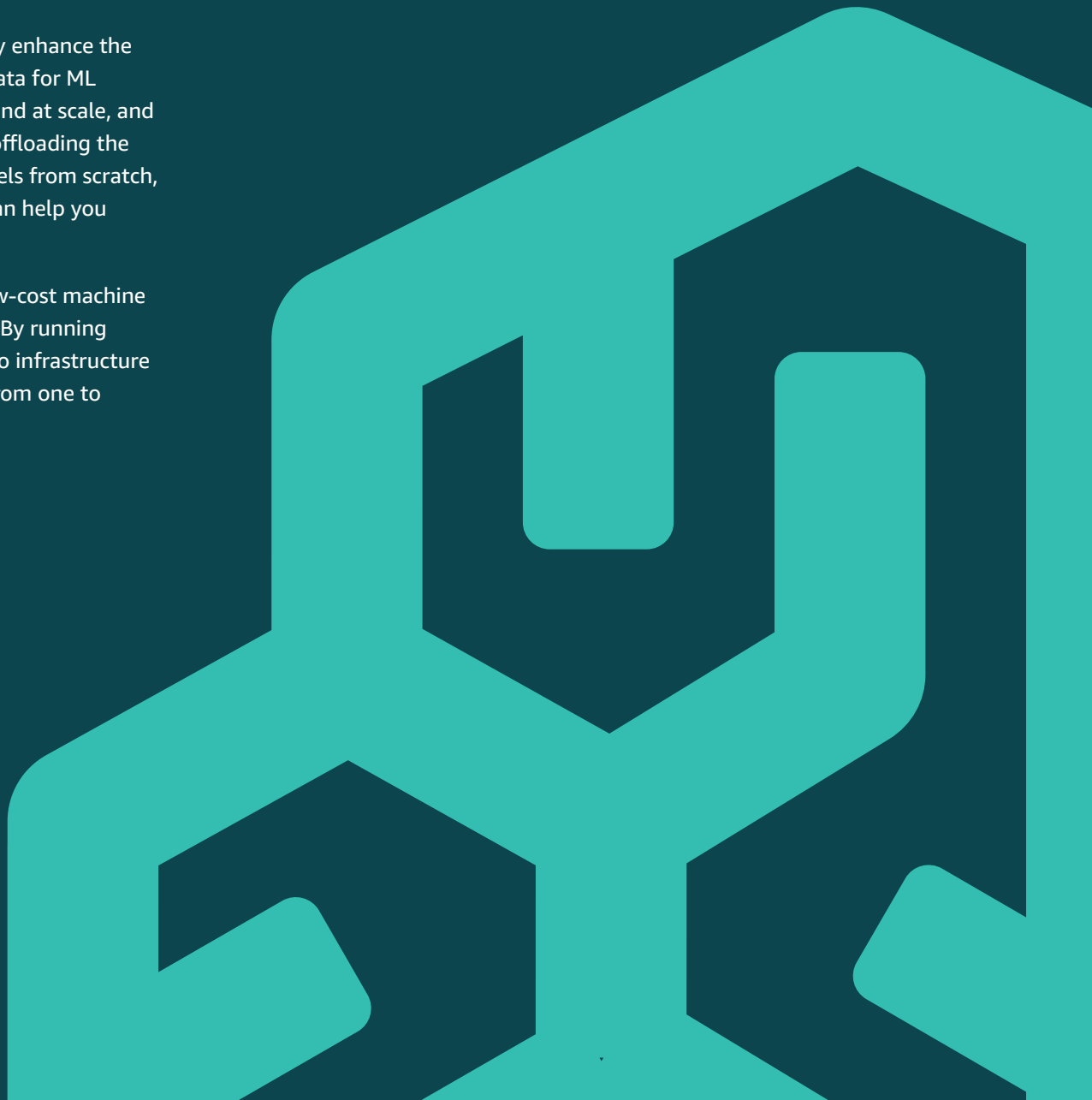
[9] Amazon EC2 Inf1 Instances

Build on a solid foundation for ML success

The right choice of services and infrastructure can substantially enhance the performance of your ML workloads, enabling you to prepare data for ML faster, reliably build sophisticated models, train them 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 machine learning services and infrastructure services optimized for ML. By running your ML workloads in the cloud, you'll get on-demand access to infrastructure and ML tools that can spin up instances in minutes and scale from one to thousands of instances—while only paying for what you use.

Get started with ML ›



aws