AWS Machine Learning Infrastructure Helps You Speed Deployment of ML Workloads

Businesses have found new ways to leverage machine learning for recommendation engines, object detection, voice assistants, fraud detection, and more. The use of machine learning is gaining traction, but long development time, high costs, the need for agility, and high complexity are key barriers that prevent use of machine learning from becoming even more widespread.

More machine learning happens on AWS than anywhere else

Toyota Research Institute · intuit · Celgene · Capital One · Zillow

AWS Machine Learning Infrastructure delivers 4 key benefits

High performance

- For fast model training, Amazon EC2 P3 instances offer the highest performance GPU training instances in the cloud
- For high performance storage access, FSx for Lustre delivers sub-millisecond latencies and throughput

Cost effective

- With a broad choice of services available, you can choose the right ML infrastructure solution for your budget
- Inf1 instances deliver high performance and the lowest cost machine learning inference in the cloud. Other options include C5 for CPU inference, G4 for GPU inference, and Elastic Inference for customized inference

Scalable

- For compute needs, you can scale up or down as needed from one GPU to thousands
- For storage needs, you can scale up or down as needed from one TB to Petabytes of storage

Easy to use

- Amazon SageMaker, a fully managed ML service, is the fastest and easiest way to get started
- Deep Learning AMIs and Deep Learning Containers come pre-installed with ML frameworks and docker images

Services for every stage of your ML workflow

Prepare

Data labeling

Amazon SageMaker Ground Truth offers easy access to labelers through Amazon Mechanical Turk and provides them with built-in workflows and interfaces for common labeling tasks

Management of large amounts of data

Amazon EMR processes vast amounts of data quickly at scale

Shared file storage of large amounts of data

Amazon Simple Storage Service provides long-term durable and readily accessible data storage

Build

Accessing Jupyter Notebooks

Hosted Jupyter Notebooks runs on an EC2 instance of your choice

Getting started using multiple ML frameworks

AWS Deep Learning AMIs let you quickly launch EC2 instances pre-installed with popular deep learning frameworks

Getting started with containers using multiple ML frameworks

AWS Deep Learning Containers come pre-installed with frameworks supporting TensorFlow, PyTorch, and MXNet

Train

Time sensitive large-scale training

Amazon EC2 P3 instances deliver up to 1 petaflop of mixed-precision performance per instance, with up to 100 Gbps of networking throughput

Multi-node training

Elastic Fabric Adapter enables running of applications requiring high levels of inter-node communications

Throughput and latency of storage access

Amazon FSx for Lustre delivers shared file storage with high throughput and consistent low latencies

Deploy

Low-cost, high-throughput inference

Amazon EC2 Inf1 instances feature up to 16 high-performance AWS Inferentia ML chips and deliver the lowest cost inference in the cloud

Inference for models using NVIDIA's CUDA, CuDNN or TensorRT libraries

Amazon EC2 G4 instances are equipped with NVIDIA T4 GPUs, delivering up to 40x better low-latency throughput than CPUs

Inference using Intel AVX-512 VNNI Instructions

Amazon EC2 C5 instances include Intel AVX-512 VNNI which helps speed up typical machine learning operations like convolution

AWS Machine Learning Infrastructure services are high-performing, cost-effective, agile, and easy-to-use for your machine learning workloads.

To learn more, visit https://aws.amazon.com/machine-learning/infrastructure/