AWS Machine Learning Infrastructure Helps You Speed Deployment of ML Workloads

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/

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.

**Build**
- Management of large amounts of data
- Amazon EMR processes vast amounts of data quickly at scale.

**Train**
- Single-node training
- Amazon's 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.

**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.

**More machine learning happens on AWS than anywhere else**