IDC

Realize Superior Business Outcomes, Developer Efficiency, and Accelerated Innovation with High-Performance, Cost-Efficient, and Easy-to-Use Machine Learning Infrastructure

RESEARCH BY:

Sriram Subramanian
Research Director, Infrastructure Systems, Platforms and Technologies Group, IDC

October 2021 | IDC Doc. #US48194621

Table of Contents

Click on titles or page numbers to navigate to each section.

Executive Summary

Enterprises across industries and geographies are increasingly adopting artificial intelligence (AI) and machine learning (ML) technologies to enable new use cases, improve customer experience, increase employee productivity, and accelerate business innovation.

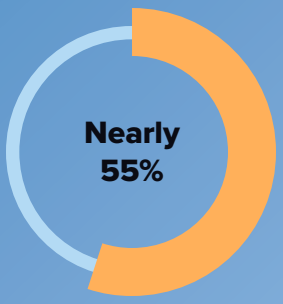
According to the IDC's *Worldwide Semiannual Artificial Intelligence Tracker* (1H20 release, published January 2021), the worldwide revenue for the AI/ML market, including software, hardware, and services, is forecast to continue its growth trajectory with a 16.4% year-over-year increase in 2021 to $327.5 billion, reaching more than $500 billion with a five-year compound annual growth rate (CAGR) of 17.5% by 2024. IDC studies* also observe a direct correlation between AI/ML adoption and superior business outcomes. Machine learning in general and more specifically AI capabilities such as natural language processing (NLP), conversational AI, and computer vision are at the forefront of transformative customer and employee experiences. Respondents cited that IT automation, intelligent task/process automation, automated threat analysis/investigation, and automated customer service agents are some of the top AI use cases across industry verticals.

* AI StrategiesView 2021

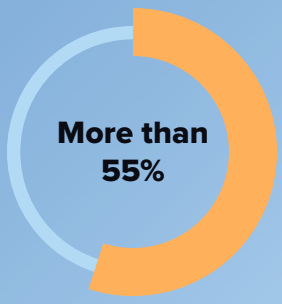
Customers are increasingly adopting cloud platforms to leverage scale, agility, and the choice of infrastructure and cloud services options. IDC studies* confirm this trend, with about 55% of the respondents citing that their AI/ML applications/solutions are deployed on the public cloud. Customers are also increasingly leveraging hardware accelerators for their AI/ML needs, with more than 55% of the respondents indicating using hardware accelerators for most of their AI/ML needs. Containerized deployments of ML models have become the standard method of model deployment—owing to the scale, consistency, and portability they enable across platforms throughout the ML life cycle (from experimentation to production).

This white paper discusses how AWS enables customers to accelerate their AI/ML innovation by providing a breadth of infrastructure choices for machine learning/deep learning (DL) training and inference needs that are optimized for price performance. This paper also provides an overview of the Amazon EC2 DL1 instances, built from the ground up for deep learning training, and provides recommendations on selecting the right AWS ML infrastructure and services based on the use case.

Increasing Adoption of Cloud Platforms among Customers



of respondents say their AI/ML applications/solutions **are deployed on the public cloud.**



of respondents **indicate using hardware accelerators** for most of their AI/ML needs.

* AI StrategiesView 2021

Situation Overview

AI/ML Adoption Trends

It is not a bit hyperbolic to claim that AI/ML adoption among enterprises has been rising mercurially in recent years. With this explosive growth of AI/ML adoption, IDC observes the following key trends:

Enterprises are leveraging AI/ML capabilities to improve customer experience, enhance employee productivity, identify new business opportunities, and accelerate innovation. Various IDC studies show that IT automation, supply chain and logistics, and intelligent process automation (IPA) are the top use cases across industries and geographies. AI/ML capabilities such as NLP, conversational AI, and image recognition are transforming these use cases by providing actionable insights sooner, improved accuracy, and better contextual intelligence.

Data scientists and machine learning engineers are constantly pushing the boundaries of technology to create larger and more complex models that provide higher accuracy predictions in an increasing number of problem spaces. Evolving use cases such as object detection, natural language processing (NLP), image classification, conversational AI, and time series data are better served by deep learning—a subset of machine learning algorithms that mimics the working of the human brain, using multiple layers of artificial neural networks. Deep learning models are large and complex, consisting of billions of parameters. This growth in model complexity translates to an exponential growth in the underlying compute infrastructure required to train and run these models, resulting in increased costs.

Despite the claims on industry standardization around few ML platforms, IDC observes that end users are using a plethora of ML platforms. Recent IDC studies* show that Spark ML, Apache MXNet, TensorFlow, and PyTorch are the most used ML frameworks for model build/training. However, it is interesting to note that end users seek similar capabilities such as accelerated model development, better model monitoring, and a faster path to production across these platforms.

Challenges with AI/ML Adoption

Along with the aforementioned trends, IDC observes that AI/ML adoption is not without challenges. IDC also observes that high failure rates in AI/ML initiatives are not uncommon. In a recent study, respondents were asked about their significant challenges to implementing AI/ML. About 56% of the respondents cited* cost as their most significant challenge to implementing AI/ML solutions.

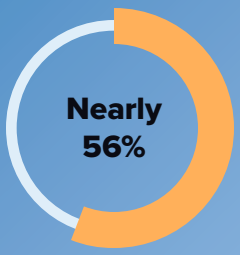
Respondents also cited lack of expertise, lack of production-ready data pipelines, and lack of integrated development environments as other challenges that impact successful execution of their AI/ML initiatives.

Cost

Cost is the most important challenge to implementing AI/ML initiatives. Across the life cycle of ML models—during building, training, testing, and inferencing, cost optimization plays a significant role in determining key decisions—including the choice of infrastructure, deployment location, number of training iterations, and frequency of inferencing.

Model training is an iterative process—multiple iterations of training the model on the sample data set may be required to get the model to converge to the desired accuracy. Often, data scientists also tune the model several times and retrain to improve accuracy. Every iteration of model training increases the cost—both the cost of model training and the opportunity cost of not bringing out the product to the market sooner.

About 74% of the respondents to a recent survey* indicated running 5–10 iterations of model training before reaching the required accuracy.



of respondents **say cost is the most significant challenge** to implementing AI/ML solutions.

Cost optimization plays a significant role in key decision-making.

* AI StrategiesView 2021

More than half of the respondents also indicated rebuilding models weekly or more frequently, with 26% of them rebuilding daily or hourly. As models grow more in size and complexity, IDC expects the number of training iterations needed to achieve the required accuracy and hence the cost for model training to increase. The increasing cost to train models combined with the lack of integrated development environments also limits the ability to experiment and retrain the models more frequently.

High-performing and cost-efficient machine learning infrastructure enables customers to train their models quickly without impacting their budgets. Cloud-based as-a-service infrastructure options help customers lower the cost of machine learning workflows by eliminating up-front hardware investments. Pay-as-you-go pricing models also provide more granularity and visibility into costs, thereby enabling customers to take more control of their training costs. Finally, as customers gain more maturity in their AI/ML adoption on the cloud, they are better positioned to leverage platform and cost optimizations provided by the cloud service provider.

Lack of Agility

Agility in machine learning workflows refers to the ability to run model training iterations and to tune/change the model itself as often as required to get the model to converge to the desired accuracy. It is similar to what is commonly referred to as developer agility in the software development life cycle.

Delays in infrastructure provisioning introduced by traditional IT provisioning processes slow down machine learning workflows, thereby limiting the ability to iterate as quickly as possible. Customers can use on-demand, self-provisioning cloud-based infrastructure to run training iterations as often as required without being slowed by traditional IT provisioning processes. Self-service provisioning capabilities are also foundational to building continuous integration (CI)/continuous deployment (CD) pipelines for machine learning models that enable updating models as quickly as possible when needed.

Lack of Infrastructure Choices in On-Premises Environments

One of the most common myths around machine learning is the indispensability of GPUs for model training and inferencing. While GPUs certainly accelerate machine learning workflows, they are not the only choice of compute infrastructure available. Nor are these the most cost effective always.

* AI StrategiesView 2021

A recent IDC survey* shows that end users use a variety of infrastructure — general-purpose CPUs, GPUs, and chips purpose built for deep learning — for their training and inferencing needs. Customers also prefer the ability to port models across multiple deployment options and locations, as the model moves from experimentation to production.

Customers who deploy their AI/ML models on premises are tied to their infrastructure investments and will not be able to experiment with a variety of infrastructure options quickly and easily to best meet their performance and budget needs. On public cloud infrastructure, end users have more choice of compute options — general-purpose, compute-optimized, memory-optimized, and accelerated instances. Public cloud service providers also provide more choices of ML infrastructure: CPU-based instances, GPU-based instances, and accelerated instances based on custom chips such as Amazon EC2 DL1 instances powered by Gaudi accelerators from Habana Labs, an Intel company. With a range of choices, customers have the flexibility to use different compute infrastructure/instances based on their model training and inference needs.

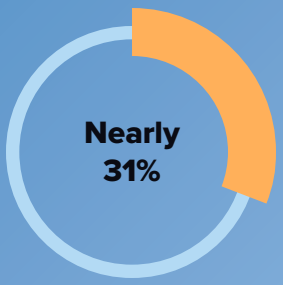
How Cloud Helps Scale AI Adoption

Implementing ML capabilities into a business application consists of four stages. Developers need to prepare and label sample data to train their models. They need to choose the right frameworks to build their models. The models then need to be trained to make predictions and tuned frequently to increase prediction accuracy. Finally, models need to be deployed, a process called inference — integrated into their applications, monitored, scaled, and managed in production. To successfully deploy their ML applications through the ML life cycle, businesses need to make significant investments in infrastructure, resources, and skill sets. Further:

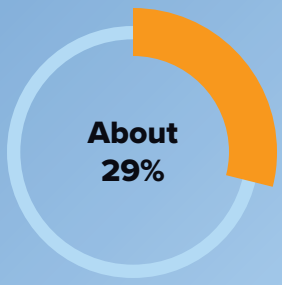
Cloud platforms provide managed infrastructure services as well as purpose-built and ML-optimized cloud services and tools that enable customers to get started on their AI/ML projects quickly and scale their existing ML applications.
Customers get on-demand access to data management, workflow management, and high-performance, low-cost, and easy-to-use infrastructure services for every stage of the ML life cycle. About 31% of the respondents indicated* the lack of ML infrastructure expertise to be one of their top 3 biggest challenges that impacted the success of their AI projects. Cloud-based ML infrastructure helps customers

* AI StrategiesView 2021

Need for Significant Investments in Infrastructure, Resources, and Skill Sets



of respondents say **the lack of ML infrastructure expertise is one of their top 3 challenges.**



of respondents say **the lack of access to specialized hardware is one of their top 3 biggest challenges.**

overcome this challenge by providing infrastructure services without the need to manage the underlying infrastructure. The elasticity and scale provided by the cloud platforms enable customers to run model training without any up-front investments in ML infrastructure.

About 29% of the respondents indicated* the lack of access to specialized hardware as one of their top 3 biggest challenges that impacted the success of their AI projects. Cloud platforms offer flexibility and choice of ML infrastructure options, including CPU-based, GPU-based, and custom accelerated instances.

The self-service provisioning capabilities enabled by cloud-based infrastructure enable data scientists to provision the right compute resources required for model training on demand. This enables them to train machine learning models as often as required, without being limited by resource or process constraints.

The flexibility and choice of compute options enabled by cloud service providers enable customers to drive down the cost through optimal and efficient use of the right compute resources. By selecting the right compute resource most cost efficient for each step in the ML workflow, customers can drive down the cost of training and deploying their ML workloads.

Through managed ML infrastructure capabilities, elasticity and scale, choice of infrastructure options, and self-service capabilities, cloud-based infrastructure helps customers scale their AI/ML adoption by providing what they want—faster time to insights with more agility, flexibility, and cost optimizations.

* AI StrategiesView 2021

AWS for ML Application Development and Deployment at Scale

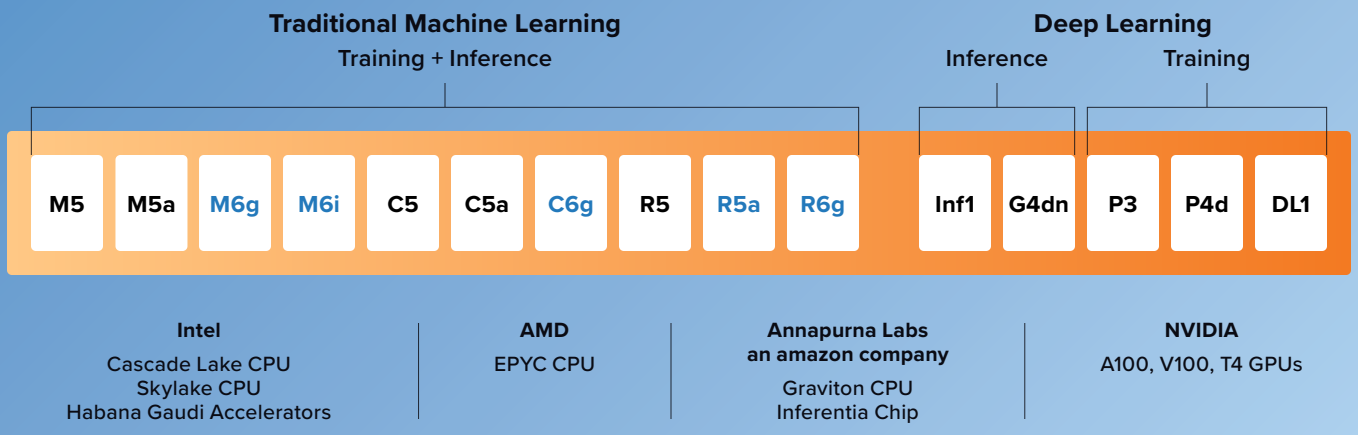
AWS proves to be an ideal platform for ML application development and deployment at scale by providing the agility, flexibility, and cost optimizations that customers want through a wide range of infrastructure choices for ML training and inference needs, optimized price performance, and a lowered barrier to entry.

Breadth of Choice of ML Infrastructure

AWS provides customers a breadth of choice of ML infrastructure in the cloud. This includes GPU-based and accelerator-based instances in partnership with multiple chip vendors including Intel, NVIDIA, and AWS-built Inferentia chips from Annapurna Labs. With Amazon EC2 DL1 instances, AWS expands the ML infrastructure choices to instances built from the ground up for deep learning training needs. Amazon EC2 DL1 instances are powered by Gaudi accelerators that are designed for training deep learning models.

FIGURE 1

AWS Instances for Machine Learning/Deep Learning Needs



Source: AWS, 2021

Lower Barrier to Entry

AWS helps customers accelerate their AI/ML adoption by lowering the barrier to entry of using the cloud to scale ML applications through cost-efficient machine learning infrastructure and a set of managed/self-managed services. Further:

AWS offers low-cost, high-performance, and easy-to-use ML infrastructure services that can significantly lower the cost of ML training in the cloud.

With services such as Amazon SageMaker, Amazon EKS, and Amazon ECS and the availability of prebuilt AMIs, AWS removes the overhead of resource provisioning and infrastructure operations. Developers at all levels of expertise, from beginner to advanced, can get started with building their ML applications easily using popular frameworks such as TensorFlow and PyTorch, models, and tools.

Amazon SageMaker enables customers to build, train, test, and deploy machine learning/deep learning models significantly faster and easier. Amazon SageMaker supports using popular ML frameworks such as TensorFlow and PyTorch and languages such as Python and R natively. It also supports popular languages, runtimes, SDKs, and application life-cycle management frameworks to enable ML application developers to develop applications in their preferred choice of development environments.

By providing a wide range of infrastructure choices for ML training and inference needs, optimized price performance, and a lowered barrier to entry to using the cloud to scale, AWS proves to be an ideal platform for ML application development and deployment that enables customers to innovate faster through AI enablement.

Habana Gaudi Accelerator-Based DL1 Instances: A Quick Look

DL1 instances—Amazon EC2 instances powered by Habana Gaudi accelerators—are one such example of a cost-efficient machine learning infrastructure option. They are built from the ground up for training deep learning models. Combining Gaudi's purpose-built architecture with 400Gbps of networking throughput, 4TB of NVMe SSD storage, and Habana's integrated software stack makes these instances ideal for training deep learning models. They can be launched using AWS Deep Learning AMIs, Amazon EKS and ECS, or Amazon SageMaker.*

* Support for Amazon SageMaker for Amazon EC2 DL1 instances is coming soon.

Integrated Software Stack

Developers at any level of expertise can get started with deep learning training easily with DL1 instances. Developers who want to use their ML life-cycle management software and tools can leverage AWS Deep Learning AMIs and Containers that come preconfigured with Habana's SynapseAI SDK. They can leverage Amazon ECS or Amazon EKS for containerized workloads or Amazon SageMaker for a powerful, fully managed experience.*

The Synapse AI Software Suite includes an SDK, Habana's graph compiler and runtime, TPC kernel library, firmware and drivers, and other developer tools that simplify building and training models. DL1 instances support leading ML frameworks such as TensorFlow and PyTorch, enabling developers to continue using their preferred ML workflows. Customers can access optimized models such as Mask R-CNN for object detection and BERT for natural language processing on Habana's GitHub repository to quickly build, train, and deploy their models. Using SynapseAI, developers can also easily migrate existing models running on CPU- or GPU-based instances to DL1 instances with minimal code changes.

Value Proposition

Based on the studies conducted by AWS and Intel/Habana, DL1 instances can deliver up to 40% better price performance than current-generation GPU-based EC2 instances for training deep learning models. By providing such a superior price performance, AWS lowers the barrier for customers to begin exploring and scaling their deep learning models in the cloud. By providing the SynapseAI SDK, integration with leading ML frameworks, and tools to migrate models easily from GPU-based and CPU-based instances, DL1 instances enable customers to get started easily on their deep learning training. Customers can leverage cloud scale, agility, and a range of workflow and data management cloud services to accelerate their AI/ML transformation.

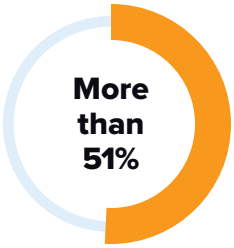
* Support for Amazon SageMaker for Amazon EC2 DL1 instances is coming soon.

Future Outlook

Increased Adoption of Cloud-Based ML Infrastructure

With the increasing adoption of AI/ML in the cloud, IDC forecasts that about 51% of the revenue from AI/ML life-cycle software will be cloud based by 2025.* This likely indicates that the majority of ML infrastructure would be cloud based by then. As more enterprise workloads are being migrated to the cloud, IDC also expects to see increased adoption of cloud-based ML services such as Amazon SageMaker.

Cloud-based ML infrastructure provides customers with a wide range of infrastructure options, positions them to leverage other cloud-based services such as databases, and enables them to accelerate innovations quickly by providing end-to-end automation for the entire ML pipeline. The wide range of infrastructure options provided also enables customers to leverage multiple ML models simultaneously as they grow in their AI adoption maturity.



of the revenue from AI/ML life-cycle software will be cloud based by 2025 as forecasted by IDC.

* IDC Semiannual Software Tracker, 2H20 Release

Opportunities and Challenges for AWS

Opportunities: Meeting Where the Customer Is

While AI/ML adoption is increasing globally, and AI/ML applications are growing exponentially, not all enterprises are ahead in their adoption maturity. As with cloud adoption, most enterprises are lagging in AI/ML adoption. This is reflected in multiple IDC studies where respondents cite lack of expertise as one of the top challenges with AI/ML adoption.

IDC observes that enterprises behind in their AI/ML adoption maturity need assistance in their AI transformation journeys. With a wide portfolio of infrastructure, machine learning as a service (MLaaS)/ML platform, and AI services; several successful user stories; and expertise across multiple industries, AWS is uniquely poised to meet where the customers are at in their AI-enabled digital transformations.

Challenges: Complexity

Based Lack of data science expertise and lack of clarity are among the top reasons cited for failures in AI/ML initiatives. With so many variables at play—the type of compute and storage infrastructure, resource provisioning, data management, type of algorithms, hyperparameter tuning, and scoring—to name a few, it is no wonder AI/ML life-cycle management is daunting. Selectig the right instance type and the right services for the right use case among a plethora of choices and orchestrating them all together is no simple task.

While deep learning models are promising for evolving use cases such as object detection, image classification, and natural language processing, it is still early stages of adoption of deep learning techniques. Although AWS enables a high-performing, cost-efficient, and easy-to-use ML infrastructure, given the complexity involved with deep learning models, the end users may need a lot more hand-holding in their early stages of deep learning.

Recommendations for the Technology Buyer

Start with Business Use Cases

As with any technology innovation, IDC recommends starting with business use cases while developing and deploying AI/ML applications. Without the right goals, success metrics, and KPIs, one is bound to face disappointments. Having unrealistic expectations is the number one reason cited for failures in AI/ML initiatives.

For example, identify whether you need to build AI/ML capabilities in-house, you can leverage cloud-based services, or you can procure AI-enabled COTS/SaaS offerings. One needs to consider various aspects such as in-house expertise, availability of AI data, appetite for investments, propensity for capex versus opex, and availability of COTS/SaaS offerings.

Leverage Cloud to Scale and Iterate

Different use cases need different types of machine learning/deep learning models—these models vary based on their algorithms, accuracy, capacity, and complexity. The infrastructure requirements for machine learning/deep learning models differ based on these attributes. The accuracy of a given model can be increased through multiple iterations, leveraging different learning algorithms, or using different training sets. Cloud-based infrastructure provides the choices and scale needed for iterating as required.

IDC recommends selecting the right infrastructure (instance types) based on model attributes, performance requirements, and cost considerations. For training, key considerations are model size, time, and cost to train models. For inference, low latency and high throughput are critical to delivering an ideal real-time user experience. As model complexity grows, the underlying compute infrastructure costs for ML inference have also grown exponentially.

IDC recommends leveraging cloud-based infrastructure to train machine learning/deep learning models and iterate as often as necessary to meet required accuracy needs, without being limited by lack of resources or operational complexities.

Tame the Heterogeneity Through a Unified Life-Cycle Management Platform

IDC recommends embracing the reality of a heterogeneous mix of AI use cases and machine learning/deep learning models and taming such heterogeneity through unified life-cycle management.

A unified life-cycle management service such as Amazon SageMaker enables end users to build, train, test, and deploy machine learning models at scale. Amazon SageMaker provides features to enable all phases of an ML pipeline end to end, thereby accelerating innovations through AI enablement. Amazon SageMaker enables end users to mix and match ML training algorithms, thereby providing them with more control and flexibility. Using Amazon SageMaker also relieves end users of the overhead of resource provisioning, thereby making MLOps easier.

IDC recommends taming the heterogeneity of AI/ML use cases through a unified life-cycle management platform such as Amazon SageMaker.

About the Analyst

Sriram Subramanian
Research Director, AI/ML Lifecycle Management Software, IDC

Sriram Subramanian is a Research Director within IDC's AI and Automation Software Research group covering AI/ML Lifecycle Management Software. His coverage on AI/ ML Lifecycle management includes the tools, technologies, and end-user needs for building, training, tuning, running, and scaling the end-to-end life cycle for AI/ML solutions from experimentation to production. Major themes of his research include MLOps, Trustworthy AI, AI Build, and Data Labeling. He advises vendors on product strategy, market positioning, and messaging; and buyers on implementing end to end scalable ML pipelines and AI-enabled digital transformation best practices to meet their business needs.

More about Sriram Subramanian

IDC Custom Solutions

This publication was produced by IDC Custom Solutions. As a premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets, IDC's Custom Solutions group helps clients plan, market, sell and succeed in the global marketplace. We create actionable market intelligence and influential content marketing programs that yield measurable results.