How the Public Clouds are Innovating on AI

Companies without the resources to develop in-house machine learning models are turning to the big cloud providers.

Thinkstock

The three big cloud providers, specifically Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), want developers and data scientists to develop, test, and deploy machine learning models on their clouds. It's a lucrative endeavor for them because testing models often need a burst of infrastructure, and models in production often require high availability.

These are lucrative services for the cloud providers and offer benefits to their customers, but they don't want to compete for your business only on infrastructure, service levels, and pricing. They focus on versatile on-ramps to make it easier for customers to use their machine learning capabilities. Each public cloud offers multiple data storage options, including serverless databases, data warehouses, data lakes, and NoSQL datastores, making it likely that you will develop models in proximity to where your data resides. They offer popular machine learning frameworks, including TensorFlow and PyTorch so that their clouds are one-stop shops for data science teams that want flexibility. All three offer Modelops, MLops, and a growing number of capabilities to support the full machine learning life cycle.

A recent study shows that 78% of enterprise ar-

tificial intelligence (AI) and machine learning (ML) projects are deployed using hybrid cloud infrastructure, so the public clouds have plenty of room to grow. This implies that they will need to continue innovating with new and differentiating capabilities.

That innovation comes in several categories to help enterprises run machine learning at scale, with more services and easier-to-use platforms. Here are some specifics.

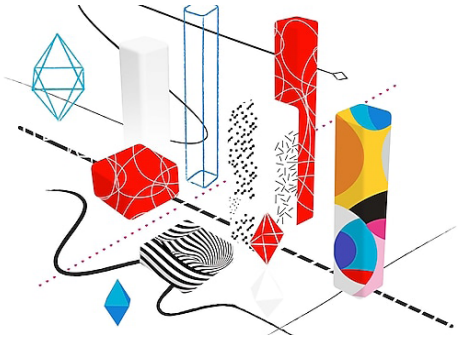
Battle of the AI chips

Machine learning experimentation continues to scale with large and more complex models that require training on vast amounts of data. Microsoft and Nvidia recently announced a massive 530 billion-parameter language processor, while Google claims it trained a 1.6 trillion-parameter model earlier this year.

Training models of this size and complexity can take a long time and become expensive, so the public clouds are innovating with AI chips and infrastructure options. AWS already has Inferentia and Trainium; it recently announced new EC2 instances powered by Habana's Gaudi that offer 40% better price-performance when compared to the latest GPU-powered EC2.

Meanwhile, Google announced TPU v4 earlier in 2021. Its fourth-generation tensor processing unit is demonstrating an average improvement of 2.7 times over TPU v3 performance. Expect more hardware innovations with AI chips and accelerators from Cerebras, Graphcore, Nvidia, and SambaNova.

Chips are not the only AI-enabling infrastructure capability, and all three public clouds have edge computing platforms to help deploy machine learning models for Internet of Things and other streaming applications.



The Roadmap to Business Modernization

Retrieving data. Wait a few seconds and try to cut or copy again.

Battle of the AI services

Most data science teams won't be developing massive-scale AI but do want to create and configure advanced machine learning models. All three cloud providers are developing machine learning services, and I expect these to grow significantly during the next several years.

Below is a brief overview of the machine learning services offered on Azure, GCP, and AWS:

- Microsoft's Cognitive Services include speech services, language services for sentiment analysis, and question and answering services often used in chatbots. Their vision services include facial recognition, and they have decision-support services used for personalization and anomaly detection.
- Microsoft recently announced the OpenAI service that connects to the GPT-3 natural language model that supports search, conversation, text completion, and other services.
- Google Cloud has several document-processing AI services, including DocAI for general document processing and vertical solutions for lending, procurement, contact centers, and contract management.
- AWS machine learning services include Rekognition in computer vision, Textract for

document processing, Lex for chatbots, CodeGuru for code reviews, and Personalize to customize web applications.

- AWS also offers industry-specific AI solutions such as Amazon HealthLake to enable predictions on health data, Amazon Lookout to identify abnormal equipment behavior, and Amazon Fraud Detector for financial services and other industries.

Will we see more machine learning models as a service (MLaaS) from public clouds and other competitors? Dr. Kirk Borne, chief science officer at DataPrime, believes so. "We will see more MLaaS or models-as-a-service offerings because of the growing sophistication of these models and corresponding expense of training them. Fewer organizations will want to invest the time and talent to build their own instances of those pretrained models."

Borne continues, "Huge numbers of small to midsize businesses getting ramped up with ML and AI will find these X-aaS offerings perfectly suited to their time, budget, and strategic requirements. MLaaS also helps address the omnipresent talent gap by taking advantage of pretrained models as a service that use sophisticated and powerful algorithms."

Google Cloud VMware Engine: Migrate, scale and innovate at speed

Every enterprise is striving to adopt a cloud-first strategy - but making that happen is easier said than done. Google Cloud VMware Engine enables business agility without risk, app refactoring or...

Battle to make AI more accessible

The next frontier for public clouds is to enable their machine learning and AI capabilities to organizations that may not have advanced software development and data science teams. They are doing this through low-code technologies that either have built-in machine learning capabilities or help developers interface with their other AI services.

AWS SageMaker's IDE simplifies developing, testing, and deploying machine learning models. The IDE provides several advanced capabilities, including a data wrangler to help data scientists prep data, a feature store to promote collaboration and reuse between data science teams, and devops one-click deployment capabilities. AWS Sagemaker competes with data science platforms such as Alteryx, Dataiku, KNIME, and SAS.

Microsoft offers Azure Machine Learning Studio, a portal that combines no-code and code-first experiences for data scientists. Their more advanced low-code AI offering is AI Builder for the Power Apps platform that enables low-code developers to perform text classification, object detection, and form processing.

Google is taking a similar approach with AutoML for training models. AppSheet's built-in intelligence includes trend predictions, content classification, sentiment analysis, and other features. The public clouds compete with other low-code platforms offering machine learning capabilities, including Creatio, Outsystems, Thinkwise, Vantiq, and others.

It will be interesting to see how the public clouds, startups, enterprise software vendors, chip manufacturers, infrastructure providers, and open source platforms compete on artificial intelligence and machine learning innovation to support bigger models, more services, and easier on-ramps for integrating applications.