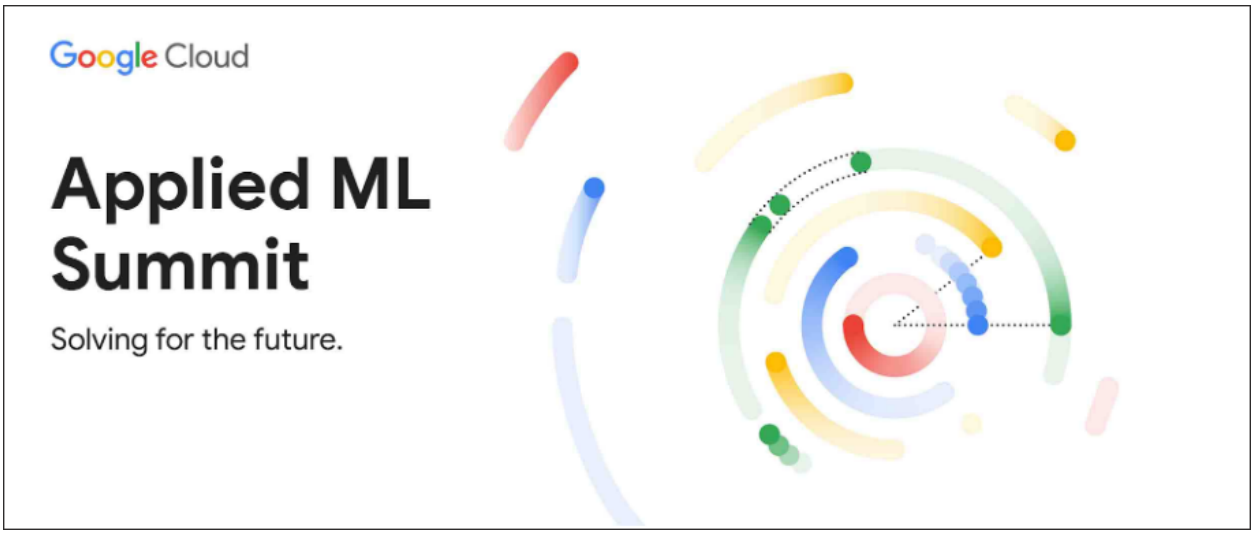
AI & MACHINE LEARNING

Accelerate the deployment of ML in production with Vertex AI



Author: Andrew Moore
Vice President & General Manager: Cloud AI & Industry Solutions

As part of Google Cloud Applied ML Summit, we're announcing a variety of product features and technology partnerships to help you more quickly and efficiently build, deploy, manage, and maintain machine learning (ML) models in production.

Our performance tests found a 2.5x increase in the number of ML predictions generated through Vertex AI and BigQuery in 2021, and a 25x increase in active customers for Vertex AI Workbench in just the last six months. Customers have made clear that managed and integrated ML platforms are crucial to accelerating the deployment of ML in production. For example, Wayfair accelerated large model training jobs by 5-10x with Vertex AI, enabling increased experimentation, reduced coding, and more models making it to production. Likewise, Seagate used AutoML to build a ML model with 98% precision, compared to only 70-80% from their earlier custom models.

Bryan Goodman, Director of AI and Cloud at Ford, said, "Vertex AI is an integral part of the Ford machine learning development platform, including accelerating our efforts to scale AI for non-software experts."

This momentum is tremendous, but we know there is more work to be done to help enterprises across the globe fast-track the digitization of operations with AI.

According to Gartner*, "Only 10% of organizations have 50% or more of their software engineers trained on machine learning skills." [Source: Gartner: Survey Analysis: AI Adoption Spans Software Engineering and Organizational Boundaries – Van Baker, Benoit Lheureux - November 25, 2021]

Similarly, Gartner states that “on average, 53% of [ML] projects make it to production.” [Source: Gartner: 4 Machine Learning Best Practices to Achieve Project Success - Afraz Jaffri, Carlie Idoine, Erick Brethenoux - December 7, 2021].

These findings speak to the primary challenge of not only gaining ML skills or abstracting technology dependencies so more people can participate in the process of ML deployment, but also to applying those skills to deploy models in production, continuously monitor, and drive business impact.

Let’s take a look at how our announcements will help you remove the barriers to deploying useful and predictable ML at scale.

Four pillars for accelerating ML deployment in production

The features we’re announcing today fit into the following four-part framework that we’ve developed in discussions with customers, partners, and other industry thought leaders.

Providing freedom of choice

Data scientists work most effectively when they have the freedom to choose the ML frameworks, deployment instances, and compute processors they’ll work with. To this end, we partnered with NVIDIA earlier this year to launch One Click Deploy of NVIDIA AI software solutions to Vertex AI Workbench. NVIDIA’s NGC catalog lets data scientists start their model development on Google Cloud, speeding the path to building and deploying state-of-the-art AI. The feature simplifies the deployment of Jupyter Notebooks from over 12 complex steps to a single click, abstracting away routine tasks to help data science teams focus on accelerating ML deployment in production.

We also believe this power to choose should not come at a cost. With this in mind, we are thrilled to announce the availability of Vertex AI Training Reduction Server, which supports both Tensorflow and PyTorch. Training Reduction Server is built to optimize bandwidth and latency of multi-node distributed training on NVIDIA GPUs. This significantly reduces the training time required for large language workloads, like BERT, and further enables cost parity across different approaches. In many mission-critical business scenarios, a shortened training cycle allows data scientists to train a model with higher predictive performance within the constraints of a deployment window.

Meeting users where they are

Whether ML tasks involve pre-trained APIs, AutoML, or custom models built from the ground up, skills proficiency should not be the gating criteria for participation in an enterprise-wide strategy. This is the only way to get your data engineers, data analysts, ML researchers, MLOps engineers, and data scientists to participate in the process of ML acceleration across the organization.

To this end, we’re announcing the preview of Vertex AI Tabular Workflows, which includes a glassbox and managed AutoML pipeline that lets you see and interpret each step in the model building and deployment process. Now, you can comfortably train datasets of over a terabyte, without sacrificing accuracy, by picking and choosing which parts of the process you want AutoML to handle versus which parts you want to engineer yourself.

Elements of Tabular Workflows can also be integrated into your existing Vertex AI pipelines. We've added new managed algorithms including advanced research models like TabNet, new algorithms for feature selection, model distillation and much more. Future noteworthy components will include implementation of Google proprietary models such as Temporal Fusion Transformers, and Open Source models like XGboost and Wide & Deep.

Uniting data and AI

To fast track the deployment of ML models into production, your organization needs a unified data and AI strategy. To further integrate data engineering capabilities directly into the data science environment, we're announcing features to address all data types: structured data, graph data, and unstructured data.
First up, for structured data, we are announcing the preview of Serverless Spark on Vertex AI Workbench. This allows data scientists to launch a serverless spark session within their notebooks and interactively develop code.

In the space of graph data, we are excited to introduce a data partnership with Neo4j that unlocks the power of graph-based ML models, letting data scientists explore, analyze, and engineer features from connected data in Neo4j and then deploy models with Vertex AI, all within a single unified platform. With Neo4j Graph Data Science and Vertex AI, data scientists can extract more predictive power from models using graph-based inputs, and get to production faster across use cases such as fraud and anomaly detection, recommendation engines, customer 360 , logistics, and more.

In the space of unstructured data, our partnership with Labelbox is all about helping data scientists leverage the power of unstructured data to build more effective ML models on Vertex AI. Labelbox's native integration with Vertex AI reduces the time required to label unstructured image, text, audio, and video data, which helps accelerate model development for image classification, object detection, entity recognition, and various other tasks. With the integration only available on Google Cloud, Labelbox and Vertex AI create a flywheel for accelerated model development.

Managing and maintaining ML models

Finally, our customers demand tools to easily manage and maintain ML models. Data scientists shouldn't need to be infrastructure engineers or operations engineers to keep models accurate, explainable, scaled, disaster resistant, and secure, all in an ever-changing environment. To address this need, we're announcing the preview of Vertex AI Example-based Explanations. This novel Explainable AI technique helps data scientists identify mislabeled examples in their training data or discover what data to collect to improve model accuracy. Using example-based explanations to quickly diagnose and treat issues, data scientists can now maintain a high bar on model quality.

Ford and Vertex AI

As mentioned, we've seen our customers achieve great results with our AI and ML solutions. Ford, for example, is leveraging Vertex AI across many use cases and user types.

“We’re using Vertex AI pipelines to build generic and reusable modular machine learning workflows. These are useful as people build on the work of others and to accelerate their own work,” explained Goodman.

“For low code and no code users, AutoML models are useful for transcribing speech and basic object detection, and we like that there is integrated deployment for trained models. It really helps people get things into use, which is important. For power users, we are extensively leveraging Vertex AI’s custom model deployment for our in-house models. It’s ideal for data scientists and data engineers not to have to master skills in infrastructure and software. This is critical for growing the community of AI builders at Ford, and we’re seeing really good success.”

Customer stories and enthusiasm propel our efforts to continue creating better products that make AI and ML more accessible, sustainable, and powerful. We’re thrilled to have been on this journey with you so far, and we can’t wait to see what you do with our new announcements. To learn more, check out additional expert commentary at our Applied ML Summit, and visit our Data Science on Google Cloud page to learn more about how Google Cloud is helping you fast-track the deployment of ML in production.
