AI & Machine Learning

Key requirements for an MLOps foundation

September 1, 2020

Craig Wiley
Director of Product Management, Cloud AI and Industry Solutions

AI-driven organizations are using data and machine learning to solve their hardest problems and are reaping the rewards.

"Companies that fully absorb AI in their value-producing workflows by 2025 will dominate the 2030 world economy with +120% cash flow growth,"[1] according to McKinsey Global Institute.

But it's not easy right now. Machine learning (ML) systems have a special capacity for creating technical debt if not managed well. They have all of the maintenance problems of traditional code plus an additional set of ML-specific issues: ML systems have unique hardware and software dependencies, require testing and validation of data as well as code, and as the world changes around us deployed ML models degrade over time. Moreover, ML systems underperform without throwing errors, making identifying and resolving issues especially challenging. Put another way—creating an ML model is the easy part—operationalizing and managing the lifecycle of ML models, data and experiments is where it gets complicated.

Unifying ML system development and operations

Starting with **AI Platform Pipelines**: we announced a hosted offering for building and managing ML pipelines on AI Platform earlier this year. We now have a fully managed service for ML pipelines that will be available in preview by October this year. With the new managed service, customers can build ML pipelines using TensorFlow Extended (TFX's) pre-built components and templates that significantly reduce the effort required to deploy models.

We offer a Continuous Evaluation service in our platform that samples prediction input and output from deployed ML models, then analyzes the model's performance against ground-truth labels. If the data needs human labeling, it also helps customers assign human reviewers to provide ground truth labels to evaluate model performance. We are excited to announce a **Continuous Monitoring** service that will monitor model performance in production to let you know if it is going stale, or if there are any outliers, skews, or concept drifts, so teams can quickly intervene, debug, or retrain a new model. This will simplify the management of models at scale, and help data scientists focus on models that are at risk of not meeting business objectives. Continuous Monitoring is expected to be available to customers by the end of 2020.

The foundation of all these new services is our new **ML Metadata Management** service in AI Platform. This service lets AI teams track all the important artifacts and experiments they run, providing a curated ledger of actions and detailed model lineage. This will enable customers to determine model provenance for any model trained on AI Platform for debugging, audit, or collaboration. AI Platform Pipelines will automatically track artifacts and lineage and AI teams can also use the ML Metadata service directly for custom workloads, artifact and metadata tracking. Our ML Metadata service is expected to be available in preview by the end of September.

productivity of users by eliminating redundant steps in feature engineering. The Feature Store will also provide tooling to mitigate common causes of inconsistency between the features used for training and prediction.

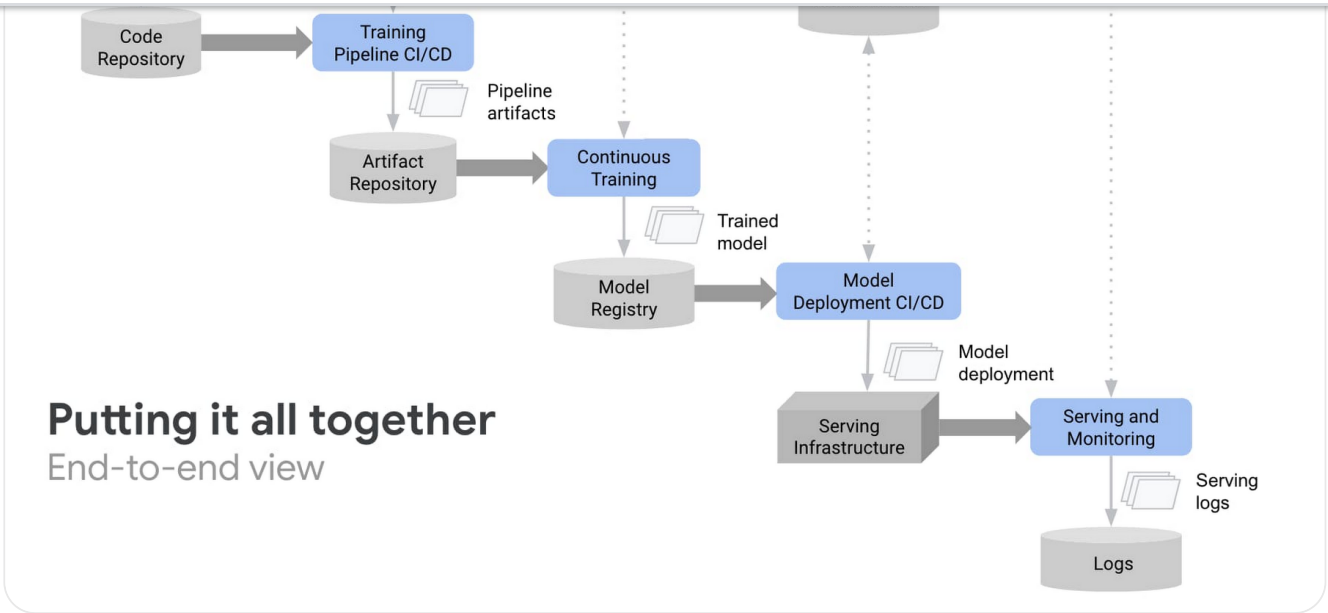
Bridging ML and IT

DevOps is a popular and common practice for developing and managing large-scale software systems that grew over decades of experience and learning in the software development industry. This practice provides benefits such as reducing development cycles, increasing deployment velocity, and ensuring dependable releases of high-quality software.

Like DevOps, MLOps is an ML engineering culture and practice that aims at unifying ML system development (Dev) and ML system operation (Ops). Unlike DevOps, ML systems present unique challenges to core DevOps principles like Continuous Integration and Continuous Delivery (CI/CD).

In ML systems:

- Continuous Integration (CI) is not only about testing and validating code and components, but also testing and validating data, data schemas, and models.
- Continuous Delivery (CD) is not only about a single software package or a service, but a system (an ML training pipeline) that should automatically deploy another service (model prediction service).
- Continuous Training (CT) is a new property, unique to ML systems, that's concerned with automatically retraining candidate models for testing and serving.
- Continuous Monitoring (CM) is not only about catching errors in production systems, but also about monitoring production inference data and model



Practicing MLOps means that you advocate for automation and monitoring at all steps of ML system construction, including integration, testing, releasing, deployment and infrastructure management. The announcements we're making today will help simplify how AI teams manage the entire ML development lifecycle.

Our goal is to make machine learning act more like computer science so that it becomes more efficient and faster to deploy, and we are excited to bring that efficiency and speed to your business. To learn more about MLOps and see how our customers are using the platform, check out the An Introduction to MLOps on Google Cloud session at Next OnAir, and our documentation on Continuous delivery and automation pipelines in machine learning and Architecture for MLOps using TFX, Kubeflow Pipelines, and Cloud Build.

---

1. Excerpted from "Notes from the AI frontier: Modeling the impact of AI on the world economy," Sept 2018, McKinsey Global Institute.