What is MLOps?

Machine learning operations (MLOps) are a set of practices that automate and simplify machine learning (ML) workflows and deployments. Machine learning and artificial intelligence (AI) are core capabilities that you can implement to solve complex real-world problems and deliver value to your customers. MLOps is an ML culture and practice that unifies ML application development (Dev) with ML system deployment and operations (Ops). Your organization can use MLOps to automate and standardize processes across the ML lifecycle. These processes include model development, testing, integration, release, and infrastructure management.

Why is MLOps required?

At a high level, to begin the machine learning lifecycle, your organization typically has to start with data preparation. You fetch data of different types from various sources, and perform activities like aggregation, duplicate cleaning, and feature engineering.

After that, you use the data to train and validate the ML model. You can then deploy the trained and validated model as a prediction service that other applications can access through APIs.

Exploratory data analysis often requires you to experiment with different models until the best model version is ready for deployment. It leads to frequent model version deployments and data versioning. Experiment tracking and ML training pipeline management are essential before your applications can integrate or consume the model in their code.

MLOps is critical to systematically and simultaneously manage the release of new ML models with application code and data changes. An optimal MLOps implementation treats the ML assets similarly to other continuous integration and delivery (CI/CD) environment software assets. You deploy ML models alongside the applications and services they use and those that consume them as part of a unified release process.

What are the principles of MLOps?

Next, we explain four key principles of MLOps.

Version control

This process involves tracking changes in the machine learning assets so you can reproduce results and roll back to previous versions if necessary. Every ML training code or model specification goes through a code review phase. Each is versioned to make the training of ML models reproducible and auditable.

Reproducibility in an ML workflow is important at every phase, from data processing to ML model deployment. It means that each phase should produce identical results given the same input.

Automation

Automate various stages in the machine learning pipeline to ensure repeatability, consistency, and scalability. This includes stages from data ingestion, preprocessing, model training, and validation to deployment.

These are some factors that can trigger automated model training and deployment:

- Messaging
- Monitoring or calendar events
- Data changes
- Model training code changes
- Application code changes.

Automated testing helps you discover problems early for fast error fixes and learnings. Automation is more efficient with infrastructure as code (IaC). You can use tools to define and manage infrastructure. This helps ensure it's reproducible and can be consistently deployed across various environments.

Continuous X

Through automation, you can continuously run tests and deploy code across your ML pipeline.

In MLOps, *continuous* refers to four activities that happen continuously if any change is made anywhere in the system:

- *Continuous integration* extends the validation and testing of code to data and models in the pipeline
- *Continuous delivery* automatically deploys the newly trained model or model prediction service
- *Continuous training* automatically retrains ML models for redeployment
- *Continuous monitoring* concerns data monitoring and model monitoring using metrics related to business

Model governance

Governance involves managing all aspects of ML systems for efficiency. You should do many activities for governance:

- Foster close collaboration between data scientists, engineers, and business stakeholders
- Use clear documentation and effective communication channels to ensure everyone is aligned
- Establish mechanisms to collect feedback about model predictions and retrain models further
- Ensure that sensitive data is protected, access to models and infrastructure is secure, and compliance requirements are met

It’s also essential to have a structured process to review, validate, and approve models before they go live. This can involve checking for fairness, bias, and ethical considerations.