H2O MLOps

H2O MLOps is a complete system for the deployment, management, and governance of models in production with seamless integration to H2O Driverless AI and H2O open source

The Operations Challenge

Scaling AI for the enterprise requires a new set of tools and skills designed for modern infrastructure and collaboration. Many organizations have learned the hard way that putting AI models into production with manual coding and homegrown systems is time-consuming and risky. After just a few production models, teams using manual deployment and management find they are strapped for resources and have no way to scale. Machine learning operations, also known as MLOps, is a new set of practices and technology that enable organizations to scale and manage AI in production. MLOps technology helps data science and IT teams collaborate and empowers IT teams to lead production machine learning projects.

H2O MLOps

H2O MLOps is a complete system for the deployment, management, and governance of models in production with seamless integration to H2O Driverless AI and H2O open source for model experimentation and training.

Production Model Deployment

H2O MLOps makes it easy to deploy models in production environments based on Kubernetes. H2O MLOps supports all major cloud providers and on-premise Kubernetes distributions like RedHat OpenShift. With H2O MLOps, users can have multiple environments for testing and production located in different places. Users can deploy a new model to the test environment with a few clicks and then deploy it to production once testing completes.

Production Model Monitoring

Production business applications depend on AI models to produce predictable and accurate results. Monitoring for service health like response time and volume is necessary but insufficient. Changes in production data can cause predictive models to be less accurate over time. Detecting these data drifts is critical to identifying which models might need to be updated. H2O MLOps includes monitoring for service levels and data drift with real-time dashboards and alerts when metrics deviate from established thresholds.

Production Model Lifecycle Management

Models running in production may need more frequent updates than other software applications and without downtime. With downstream business applications dependent on their results, production models updates must occur without interrupting service. H2O MLOps gives IT operations teams the tools to seamlessly update models in production, troubleshoot models, and run A/B tests on a test or live production environments.

Production Model Governance

Production environments require particular security and controls to ensure that software is not tampered with or accidentally corrupted. Production operators receive training in production procedures and production controls ensure their compliance through rigorous auditing of access, changes, and events. H2O MLOps includes everything an operations team needs to govern models in production, including a model repository with complete version control and management, access control and logging for legal and regulatory compliance.

Capabilities of H2O MLOps

Production-Ready Scoring Pipelines

H2O MOJOs are production-ready, scoring pipelines produced by H2O Driverless AI and H2O open-source. H2O MOJOs are perfect for production deployment with a small size and low latency for real-time and large-scale batch prediction use cases. In H2O MLOps, MOJOs are automatically deployed into containers and run on Kubernetes.

Shared Model Repository

H2O MLOps includes a shared production model repository with H2O Driverless AI via projects. When data scientists create a project in H2O Driverless AI, their counterparts in production can see the models in the project, and as they collaborate, promote those models onto test or production environments without having to import models. Users can also import H2O open-source models into MLOps for deployment and management.

Version Management and Rollback

H2O MLOps includes a complete version history per deployment and allows users to have multiple versions running simultaneously in development and production environments. As users introduce new versions of models, older versions are archived to maintain a version history and facilitate reproducibility.

Dev - Test - Prod Environments

H2O MLOps uses customer-provisioned Kubernetes environments and supports multiple infrastructure environments simultaneously. MLOps teams can have environments for development, testing, and production, all running in different locations.

Cloud or On-premise Deployment to Kubernetes

With H2O MLOps, production teams can deploy models to the cloud or on-premise Kubernetes. They can even do both from the same system. For example, they can have a development environment on-premise and the test and production environments in a VPC or vice versa.

Service Health Monitoring

Production models are software services that support downstream business applications. H2O MLOps provides a complete set of service monitoring metrics to ensure that each model deployment performs as expected and that IT teams can detect and respond to issues before they become problems for the business.

Real-time Drift Detection

When data changes between training and production, models can become less accurate. This "drift" is tracked by looking at differences between training and production data for each model feature. H2O.ai also offers an AI application for drift detection designed for data scientists, which has detailed views of each feature so data scientists can determine if they want to refit, retrain or build a new production model.

Configurable Alerts

For each deployment, MLOps users can set thresholds and alerts on a variety of metrics. When metrics hit the given point, an alert is triggered to notify the MLOps team and data scientists working on the production project about the issue so they can take appropriate action.

Event Log per Deployment

H2O MLOps includes an event log for each deployment. The log captures all events related to the deployment, including who took action and when it took place.

Seamless Updates for Production Models

Machine learning models can require frequent updates in production. With H2O MLOps, operations can easily replace models with a few clicks, and Kubernetes automatically handles the routing of new requests to the new model while the prior version handles old requests.

Model "Warm-up" Testing

Before putting new model versions into production, operations teams should test the model to ensure that it can perform under production conditions and in a production-like environment. H2O MLOps make it easy to set up for model testing to have warm-up testing environments ready and waiting.

A/B Testing

When introducing a new model or a new model version, operations teams may want to test with live traffic and compare results with the prior version. In H2O, MLOps users can run comparison tests between production champion and a challenger, or test two (or more) models and compare results in a simple A/B test.

Automatic Model Retraining

With H2O Driverless AI and H2O MLOps, users have the unique ability to set up automatic refitting and retraining of models. In H2O MLOps, teams set the conditions for retraining, and when those conditions occur, a request goes to H2O Driverless AI to retrain the model. The model is then placed in the appropriate project folder for testing or automatically promoted to production based on the deployment settings.

Conclusion

H2O MLOps delivers the capabilities operations teams need to deploy, monitor, and manage production models so organizations can finally scale AI across the enterprise. With MLOps in place, IT teams manage production projects, and data science teams can get back to doing data science where they can start new projects and create more value. H2O MLOps automates deployment, removing any need for brittle or manual coding reducing errors, and making processes repeatable. With H2O MLOps, organizations can create standardized testing and update procedures, and have audit logs to reduce risk to meet regulatory needs. H2O MLOps also gives IT teams control over production models and environments to ensure security and manage risk and compliance based on IT and corporate governance practices.