indium
Make Technology Work

MLOps on AWS: Enabling faster Time-to-market

White Paper

www.indiumsoftware.com

What is MLOps?

One of the key technologies enabling the digital transformation of businesses is artificial intelligence/machine learning. And if anything, the Covid-19 pandemic has speeded up this AI adoption, with a PwC survey report showing it going up by 18% since last year and another 54% accelerating towards it. Unfortunately, though, most of these remain as mere "pretty shiny objects" that do not deliver on the promise. This will compel businesses to build a framework to operationalize their AI projects, according to Forbes.

One of the key strategies to scale up proof of concept would be taking the DevOps route for ML projects and implement MLOps strategies.

In the DevOps methodology of software development, businesses adopt automation, platform design, and culture to ensure rapid delivery of high-quality service to improve business value and responsiveness.

The fast-paced, iterative IT service delivery of DevOps helps linking legacy apps with newer cloud-native apps and infrastructure. By integrating DevOps with ML, and by data science adopting agile software development practices, not only can the process become more efficient but also accelerate time-to-market.

With the rapid adoption of artificial intelligence and machine learning workloads, DevOps practices are being applied to these projects and are referred to as MLOps or ML Operations. This can help transform your business at scale using AI/ML projects, accelerating the delivery time, reducing defects, and improving the productivity of data scientists. It encompasses:

- Project management
- CI/CD
- Quality assurance
- Release management

Benefits of MLOps

Incorporating MLOps for ML projects helps businesses speed up development and time-to-market because of:

- **Improved Productivity** - Data engineers and data scientists get access to curated data sets in self-service environments that speed up the development process.
- **Repeatability** - Automating the MLDC process such as training, evaluation, versioning, and deployment of the model increases repeatability and accelerates ML development.
- **Reliability** - The speed of deployment, quality, and consistency increase by incorporating CI/CD practices.
- **Auditability** - Tracking and auditing of models, how they were built and deployed becomes easy due to the versioning of all inputs and outputs, be it data science experiments, source data, or trained models.
- **Data Governance** - Data governance and implementing policies to prevent model bias becomes easier with MLOps as it allows tracking changes to data statistical properties and model quality over time.

Integrating MLOps with DevOps

While there are many benefits to MLOps, it being a new field, integrating ML with DevOps can pose some challenges. These include:

Poor Integration Between Teams:

Converting business objectives into technical requirements can be one of the primary problems faced by ML projects. This is data scientists need to speak the same language as product owners and software engineers when collaborating, which they don't. Data scientists are also often not an integral part of cross-functional teams and a strong communication bridge required for greater effectiveness is often missing.

Building Data Models:

In DevOps projects, Infrastructure-as-Code (IaC) and Configuration-as-Code (CaC) are used to build environments and Pipelines-as-Code (PaC) for consistent CI/CD patterns. For MLOps, the traditional CI/CD tool has to also integrate with another workflow, the Big Data and ML training workflows. As production data is used for development activities, the building of models can take longer and requires unique metrics for performance evaluation. The use of versioned code and artifacts helps reproduce the entire end-to-end system and requires strict enforcement of the data policy to prevent biased input data that can lead to biased outcomes.

CI/CD:

Versioning of source data, a first-class input, the source code, and ML models is critical for MLOps. So when there is a change in the source or inference data, it triggers pipeline runs and enables traceability. The ML model needs to be properly validated during automated testing in the build as well as production phases.

The model training and retraining during the build phase can be a time-consuming, resource-intensive process. Therefore, the full training cycle must be performed only when there is a change in the source data or ML code, and not when related components change, requiring pipelines to be granular. Being a small part of an overall solution that can be consumed as an API by other applications and systems, the machine learning code deployment pipeline should also include steps to package it as a model.

Monitoring and Logging:

Both model training metrics and model experiments need to be captured In the feature engineering and model training phases. The form of the input data and algorithm hyperparameters need to be tuned for building an ML model, which needs to be captured. With experiment tracking, the effectiveness of data scientists improves and enables getting a repeatable snapshot of their work. In deployed ML models, data passed to the model for inference, the standard endpoint stability, performance metrics, and the quality of model output need to be monitored and assessed using an appropriate ML metric.

AWS Framework for MLOps -- An Ideal Solution

The AWS framework for MLOps facilitates the seamless integration of development and machine learning, enabling collaboration between the development teams and the data scientists and engineers. It can help with increasing the success rate of operationalizing data science and machine learning solutions for greater efficiency and speed in robust code development and overcoming the above-mentioned challenges.

The AWS MLOps Framework is extendable, providing a standard interface to manage ML pipelines for AWS as well as third-party services. It enables streamlining and enforcing architecture best practices for ML model productization. It also allows trained models (also called bring your own model) to be uploaded, pipeline orchestration to be configured, and pipeline's operations to be monitored using the solution template.

The team can become more agile and efficient as the solution allows repeating successful processes at scale, thereby accelerating development times.

AWS MLOps framework includes a pre-configured ML pipeline that can be initiated through an API or a GIT repository from the reference architecture. It also enables automating the model monitor pipeline, also known as the Amazon SageMaker BYOM pipeline. Model drift detection packaged as a serverless microservice delivers an inference endpoint.

For experimentation, development, and/or small-scale production workloads, the AWS MLOps Framework reference architecture allows a single account template that allows the deployment of all the solution's pipelines in the same AWS account.

Speeding up Time to Market with AWS MLOps

AWS's three-layered ML stack helps organizations with different levels of skills to leverage the solution for faster time-to-market. The three layers include:

AI services: This is a fully managed service layer where API calls enable you to add ML capabilities quickly to your workloads.

ML services: In this layer, managed services and resources such as the Amazon SageMaker suite enable the labeling of data for quickly building, training, deploying, and operationalizing the ML models.

ML Frameworks and Infrastructure: ML experts can use open-source frameworks to implement custom-developed tools and workflows for building, training, deploying, and operationalizing the ML models.

The framework also allows ML-based workloads to be implemented by integrating services and infrastructure from various levels of the AWS ML stack. Provisioning and configuration of cloud-based ML workloads and IT infrastructure resources can be automated using infrastructure as code (IaC).

Training/retraining and deploying of ML models can be automated using an ML (training/retraining) pipeline. With the orchestration tool, the automated ML workflow can be orchestrated and executed efficiently.

AWS offers a model monitoring solution that prevents model and data drift by monitoring the performance of production models. The feedback from the performance metrics can help with improving the development and training of future models.

Securing ML models and making them repeatable through a model governance framework that also helps with tracking and comparing is another advantage with the AWS MLOps framework.

Indium -- A Select Consulting Partner for AWS Implementation

Indium is an AWS partner with 700+ digital consultants who provide end-to-end cloud services such as consulting, system integration, and AWS-based vertical solutions, one of the key areas being MLOps.

Our AWS expertise includes:

- 150+ App Migration / Modernization
- 20+ PB Data Migration
- 70+ Cloud-native App development
- 250+ CI/CD pipelines

The Indium team also has expertise in DevOps and successfully integrates development with machine learning by facilitating collaboration between developers and data scientists on AWS MLOps platforms. Through containerization and microservices development using MLOps, we enable businesses to quickly meet customer needs and empower them to face competition.

To know more about Indium's AWS and MLOps capabilities, visit:
https://www.indiumsoftware.com/aws-partner/

INDIA

Chennai | Bengaluru | Mumbai
Toll-free: 1800 123 1191

USA

Cupertino | Princeton
Toll-free: 1 888 207 5969

UK

London

SINGAPORE

+65 9630 7959

Sales Inquiries
sales@indiumsoftware.com

General Inquiries
info@indiumsoftware.com