MLOps or ML Ops is a paradigm that aims to deploy and maintain machine learning models in production reliably and efficiently. It bridges the gap between machine learning development and production operations, ensuring that models are robust, scalable, and aligned with business goals. The word is a compound of "machine learning" and the continuous delivery practice (CI/CD) of DevOps in the software field. Machine learning models are tested and developed in isolated experimental systems. When an algorithm is ready to be launched, MLOps is practiced between Data Scientists, DevOps, and Machine Learning engineers to transition the algorithm to production systems. [1] Similar to DevOps or DataOps approaches, MLOps seeks to increase automation and improve the quality of production models, while also focusing on business and regulatory requirements. While MLOps started as a set of best practices, it is slowly evolving into an independent approach to ML lifecycle management. MLOps applies to the entire lifecycle - from integrating with model generation (software development lifecycle , continuous integration / continuous delivery), orchestration , and deployment, to health, diagnostics, governance, and business metrics.

Definition

MLOps is a paradigm, including aspects like best practices, sets of concepts, as well as a development culture when it comes to the end-to-end conceptualization, implementation, monitoring, deployment, and scalability of machine learning products. Most of all, it is an engineering practice that leverages three contributing disciplines: machine learning, software engineering (especially DevOps), and data engineering . MLOps is aimed at productionizing machine learning systems by bridging the gap between development (Dev) and operations (Ops). Essentially, MLOps aims to facilitate the creation of machine learning products by leveraging these principles: CI/CD automation, workflow orchestration, reproducibility; versioning of data, model, and code; collaboration; continuous ML training and evaluation; ML metadata tracking and logging; continuous monitoring; and feedback loops. [2]

History

The challenges of the ongoing use of machine learning in applications were highlighted in a 2015 paper. [3] The predicted growth in machine learning included an estimated doubling of ML pilots and implementations from 2017 to 2018, and again from 2018 to 2020. [4] MLOps rapidly began to gain traction among AI/ML experts, companies, and technology journalists as a solution that can address the complexity and growth of machine learning in businesses. [5]

Reports show a majority (up to 88%) of corporate machine learning initiatives are struggling to move beyond test stages. [2] However, those organizations that actually put machine learning into production saw a 3–15% profit margin increases. [6] The MLOps market size was USD 2,191.8 Million in 2024, and is projected to be USD 16,613.4 Million in 2030. [7]

Architecture

Machine Learning systems can be categorized in eight different categories: data collection , data processing , feature engineering , data labeling , model design, model training and optimization, endpoint deployment, and endpoint monitoring. Each step in the machine learning lifecycle is built in its own system, but requires interconnection. These are the minimum systems that enterprises need to scale machine learning within their organization.

Goals

There are a number of goals enterprises want to achieve through MLOps systems successfully implementing ML across the enterprise, including: [8]

Deployment and automation [9]

Reproducibility of models and predictions [10]

Diagnostics [10]

Governance and regulatory compliance [11]

Scalability [12]

Collaboration [13]

Business uses [14]

Monitoring and management [15]

A standard practice, such as MLOps, takes into account each of the aforementioned areas, which can help enterprises optimize workflows and avoid issues during implementation.

A common architecture of an MLOps system would include data science platforms where models are constructed and the analytical engines where computations are performed, with the MLOps tool orchestrating the movement of machine learning models, data and outcomes between the systems. [8]

See also

ModelOps , according to Gartner , MLOps is a subset of ModelOps . MLOps is focused on the operationalization of ML models, while ModelOps covers the operationalization of all types of AI models.

AIOps , a similarly named, but different concept - using AI (ML) in IT and Operations.

References