

DataOps

DataOps is a collection of technical practices, workflows, cultural norms, and architectural patterns that enable:

- Rapid innovation and experimentation delivering new insights to customers with increasing velocity
- Extremely high data quality and very low error rates
- Collaboration across complex arrays of people, technology, and environments
- Clear measurement, monitoring, and transparency of results

DataOps exerts control over your workflow and processes, eliminating the numerous obstacles that prevent your data organization from achieving high levels of productivity and quality. We call the elapsed time between the proposal of a new idea and the deployment of finished analytics “cycle time.” Many organizations require months of cycle time to deploy 20 lines of SQL. Lengthy cycle times discourage and disappoint users and hinder creativity.

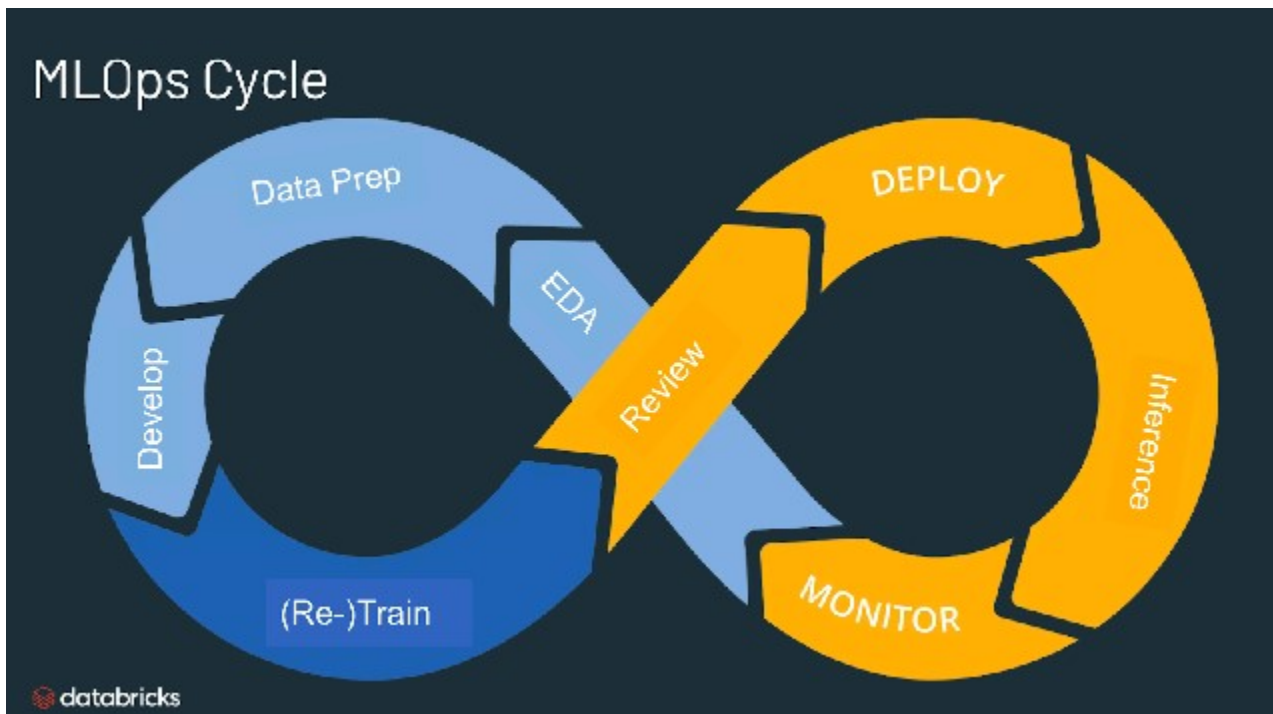
Ideally, data teams work hand-in-hand with their users like a well-oiled machine, fielding new idea proposals, implementing them rapidly and quickly iterating toward higher-quality models and analytics.

Unfortunately, our experience is the opposite. Data teams are constantly interrupted by data and analytics errors. Data scientists spend 75% of their time massaging data and executing manual steps. Slow and error-prone development disappoints and frustrates data team members and stakeholders. Lengthy analytics cycle time occurs for a variety of reasons:

- Poor teamwork within the data team
- Lack of collaboration between groups within the data organization
- Waiting for IT to disposition or configure system resources
- Waiting for access to data
- Moving slowly and cautiously to avoid poor quality
- Requiring approvals, such as from an Impact Review Board
- Inflexible data architectures
- Process bottlenecks
- Technical debt from previous deployments
- Poor quality creating unplanned work

What is MLOps?

MLOps stands for Machine Learning Operations. MLOps is a core function of Machine Learning engineering, focused on streamlining the process of taking machine learning models to production, and then maintaining and monitoring them. MLOps is a collaborative function, often comprising data scientists, devops engineers, and IT.



What is the use of MLOps?

MLOps is a useful approach for the creation and quality of machine learning and AI solutions. By adopting an MLOps approach, data scientists and machine learning engineers can collaborate and increase the pace of model development and production, by implementing continuous integration and deployment (CI/CD) practices with proper monitoring, validation, and governance of ML models.

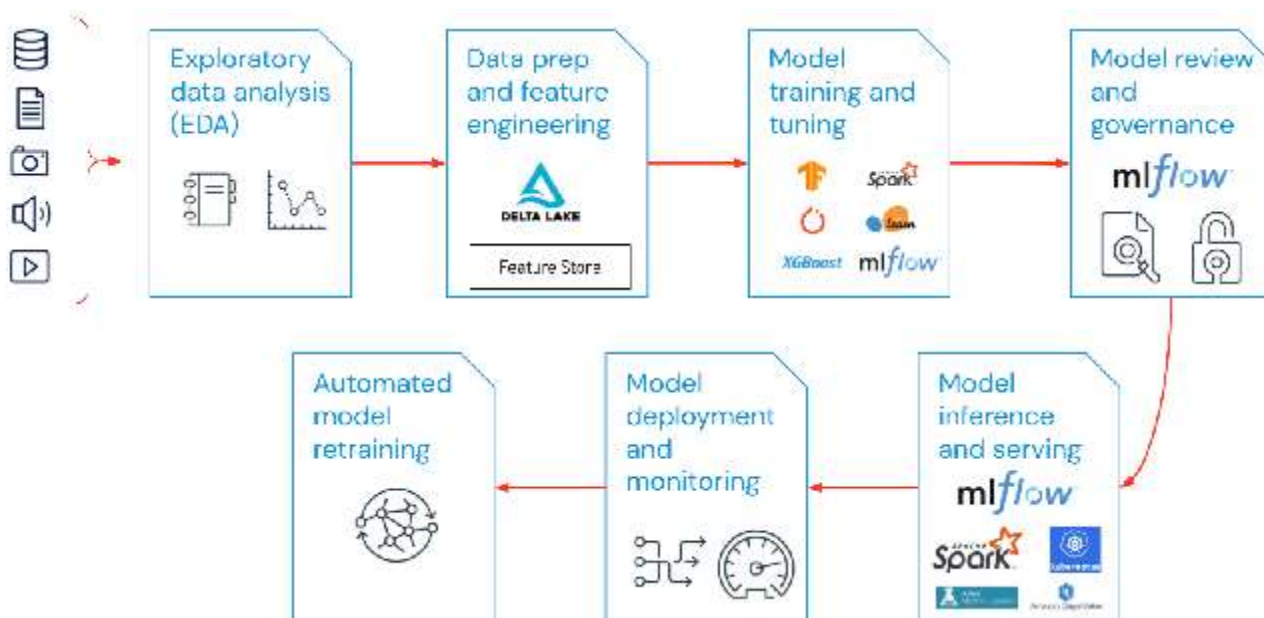
Why do we need MLOps?

Productionizing machine learning is difficult. The machine learning lifecycle consists of many complex components such as data ingest, data prep, model training, model tuning, model deployment, model monitoring, explainability, and much more. It also requires collaboration and hand-offs across teams, from Data Engineering to Data Science to ML Engineering. Naturally, it requires stringent operational rigor to keep all these processes synchronous and working in tandem. MLOps encompasses the experimentation, iteration, and continuous improvement of the machine learning lifecycle.

What are the benefits of MLOps?

The primary benefits of MLOps are efficiency, scalability, and risk reduction. Efficiency: MLOps allows data teams to achieve faster model development, deliver higher quality ML models, and faster deployment and production. Scalability: MLOps also enables vast scalability and management where thousands of models can be overseen, controlled, managed, and monitored for continuous integration, continuous delivery, and continuous deployment. Specifically, MLOps provides reproducibility of ML pipelines, enabling more tightly-coupled collaboration across data teams, reducing conflict with devops and IT, and accelerating release velocity. Risk reduction: Machine learning models often need regulatory scrutiny and drift-check, and MLOps enables greater transparency and faster response to such requests and ensures greater compliance with an organization's or industry's policies.

What are the components of MLOps?



The span of MLOps in machine learning projects can be as focused or expansive as the project demands. In certain cases, MLOps can encompass everything from the data pipeline to model production, while other projects may require MLOps implementation of only the model deployment process. A majority of enterprises deploy MLOps principles across the following:

- Exploratory data analysis (EDA)
- Data Prep and Feature Engineering
- Model training and tuning
- Model review and governance
- Model inference and serving
- Model monitoring
- Automated model retraining

What are the best practices for MLOps?

The best practices for MLOps can be delineated by the stage at which MLOps principles are being applied.

- **Exploratory data analysis (EDA)** - Iteratively explore, share, and prep data for the machine learning lifecycle by creating reproducible, editable, and shareable datasets, tables, and visualizations.
- **Data Prep and Feature Engineering**- Iteratively transform, aggregate, and de-duplicate data to create refined features. Most importantly, make the features visible and shareable across data teams, leveraging a feature store.
- **Model training and tuning** - Use popular open source libraries such as scikit-learn and hyperopt to train and improve model performance. As a simpler alternative, use automated machine learning tools such as AutoML to automatically perform trial runs and create reviewable and deployable code.
- **Model review and governance**- Track model lineage, model versions, and manage model artifacts and transitions through their lifecycle. Discover, share, and collaborate across ML models with the help of an open source MLOps platform such as MLflow.
- **Model inference and serving** - Manage the frequency of model refresh, inference request times and similar production-specifics in testing and QA. Use CI/CD tools such as repos and orchestrators (borrowing devops principles) to automate the pre-production pipeline.

- **Model deployment and monitoring** - Automate permissions and cluster creation to productionize registered models. Enable REST API model endpoints.
- **Automated model retraining** - Create alerts and automation to take corrective action In case of model drift due to differences in training and inference data.

What is the difference between MLOps and DevOps?

MLOps is a set of engineering practices specific to machine learning projects that borrow from the more widely-adopted DevOps principles in software engineering. While DevOps brings a rapid, continuously iterative approach to shipping applications, MLOps borrows the same principles to take machine learning models to production. In both cases, the outcome is higher software quality, faster patching and releases, and higher customer satisfaction.