DataOps and MLOps

DataOps (Data Operation) is an Agile strategy for building and delivering end-to-end data pipeline operations. Its major objective is to use big data to generate commercial value. Similar to the DevOps trend, the DataOps approach aims to accelerate the development of applications that use big data.

While DataOps started out as a collection of best practices, it has evolved into a fresh iteration of an autonomous approach to data analytics. DataOps understands the interrelated nature of the development of data analytics in alignment with business goals and applies to the full data lifecycle, from data display through reporting.

Why DataOps is Important?

In the present time, when the world of technology is dealing with data at every moment, DataOps in business matters a lot.

1. It enables quick experimentation and invention.
2. It helps in collaborating throughout the entire data life cycle of the organization.
3. It enables very excellent data quality and very low error rates.
4. It helps in establishing data transparency while maintaining security.

Processes are made simpler with DataOps, which also ensures continuous insight delivery.

Pros of DataOps:

Improves and emphasizes communication, collaboration, integration, automation, measurement, and cooperation between data scientists and quality assurance.

DataOps seeks and provides upgrade velocity, reliability, and quality of data analytics.

Improves better communication and collaboration between the teams and team members.

Provides real-time data insights.

It seeks to increase the velocity, reliability, and quality of data analytics.

Creates a unified, interoperable data hub.

Lower cycle time of data science applications.

Cons of DataOps:

Lack of cooperation between groups within the data organization.

Moves slowly and cautiously to avoid poor quality.

Waits for IT to dispose of or configure system resources.

Poor teamwork within the data.

Poor quality creates unplanned work.

Process bottlenecks.

Waits for access to data.

Tips for better DataOps:

While data operations are getting complicated in modern forms, which pose numerous challenges, in small teams. It keeps track of a lot of hidden ways for things to go wrong. In the DataOps approach, data pipelines are an essential component that is resilient, scalable, reliable and has high performance and throughput.

Create collaboration, Cross-functional teams.

Centralize your data sources.

Design data pipelines flexibility.

Log everything and store it.

Containerize your efforts.

Automates version control.

Learning to use DataOps for Advancement.

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.

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.

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.

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.