CAPCO

HOW CAN WE OPERATIONALIZE MACHINE LEARNING?

a **wipro** company

Analytics at Scale is fast becoming a foundational element of operating models across financial services. Given the relatively untapped potential of Big Data, firms are looking for the best methods to generate deep and impactful insights, using techniques such as machine learning (ML).

In today's financial institutions, however, too many insightful data science and ML projects live in a Jupyter Notebook and die in a PowerPoint presentation. The insights fail to create value because the institution lacks the technical knowledge and infrastructure to properly integrate machine learning algorithms into day-to-day operations.

That hurdle can be overcome using Machine Learning Operations (MLOps), an approach that builds on Agile and Development Operations (DevOps) principles to help institutions move machine learning insights into a production environment. MLOps is already helping data science to create real value in areas such as card fraud protection, balance sheet forecasting and capital liquidity requirements. With MLOps, hyper-parameters and features can be varied easily using results discovered in the performance monitoring component of the MLOps workflow. Meanwhile, the modularity of the approach means that models can be more easily aligned to the business's needs whilst ensuring that each step is documented to ease compliance.

If firms do not rapidly adopt these MLOps structures as part of their business-as-usual processes, they will lose out on attractive opportunities, expose themselves to regulatory short-comings and realize mere fractions of each ML project's potential value.

Here we set out the seven key phases of MLOps to explore how the approach relates to DevOps, and how it generates important benefits. These include a greater quality of insight, controlled governance, and model scalability, as well as reproducibility and better management of both codebase and data.

MLOPS - SEVEN KEY PHASES

Figure 1 breaks down MLOps practice into seven phases that are all linked together. During the **Model Development Cycle**, a data scientist goes through a process of rapidly cleaning, wrangling, and preparing data to be passed through a series of orchestrated modeling experiments, where the model is trained and evaluated and finally validated offline. The data comes from a Feature Store, which is a central repository for standardized definitions that facilitates re-use and prevents skewness and duplication. The main output of this phase is source code, which is stored in a repository using version control systems.

The version control system stores and provides the code and configuration artifacts that are then used to build, test, and package the pipeline components. This process is called **Continuous Integration** (CI), a step that outputs artifacts, executables, and packages for the next phases of the MLOps flow, **Continuous Training** (CT), and **Continuous Deployment/ Delivery** (CD).

CT is a process that automatically re-trains a machine learning model, taking in data from the Feature Store and training code from the source code repository, and producing a new trained model artefact. CD is an automated process that takes deployment packages and updates the serving environment to ensure that the most recent source code and model are in use.

Once fully deployed, the model is ready for **Prediction Serving**. During prediction serving, extensive **Performance Monitoring** is required to provide maximum visibility over the operations of the production environment. An MLOps pipeline would also have an **Analytics and Triggering** layer which processes the monitored metrics and logs, applying statistical analysis to identify actionable triggers.

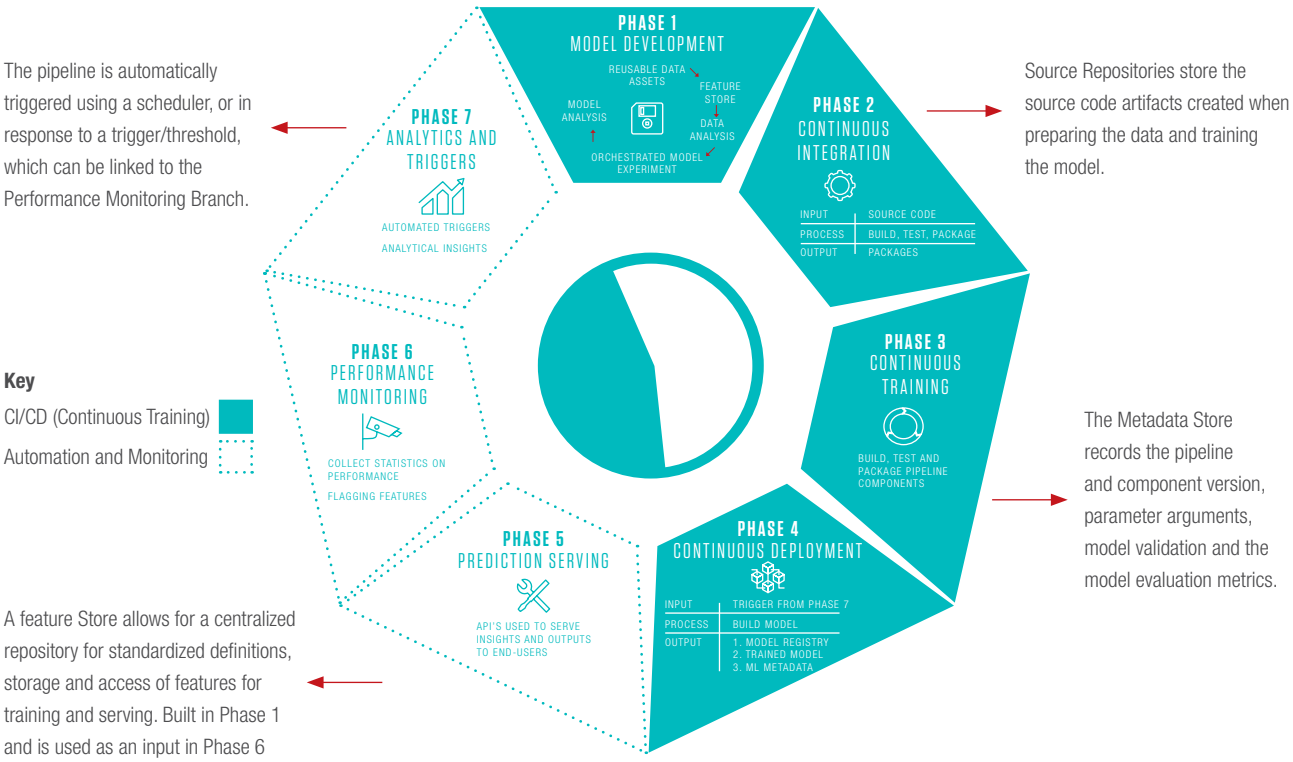

Figure 1: End to end MLOps project lifecycle

DEVOPS AND MLOPS – SIMILAR BUT DIFFERENT

These seven phases have direct parallels with DevOps, an approach already familiar to many financial services professionals and which, in turn, is the technical manifestation of Agile principles.

DevOps advocates the use of automated pipelines that enable continual updates to systems, allowing developers and engineers to focus on building application code without having to worry about breaking anything that's already there, or how their code is going to be deployed in production. DevOps implements continuous integration, unit testing, integration testing and continuous deployment as part of its fundamental processes.

As financial services embrace Big Data and AI as core components of their systems, similar functions will need to be incorporated into the ML pipeline – thus creating the need for MLOps. Figures 2 and 3 illustrate both the similarities, and the important differences, between typical DevOps and MLOps process flows.

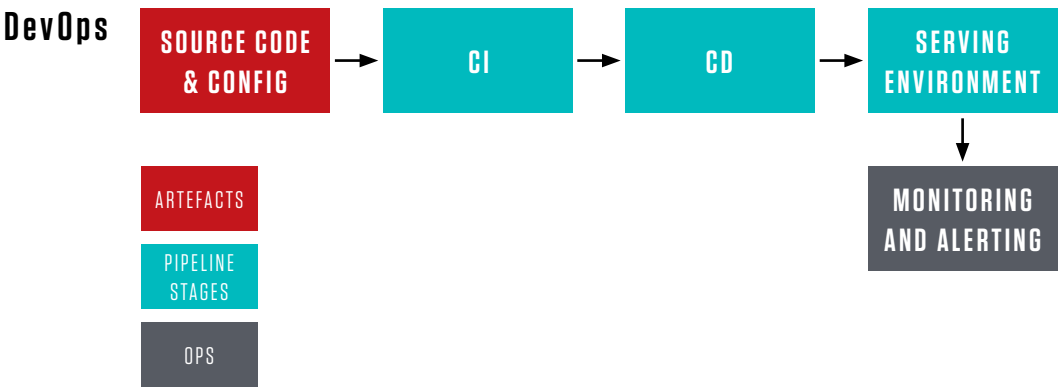

Figure 2: Typical DevOps process flow

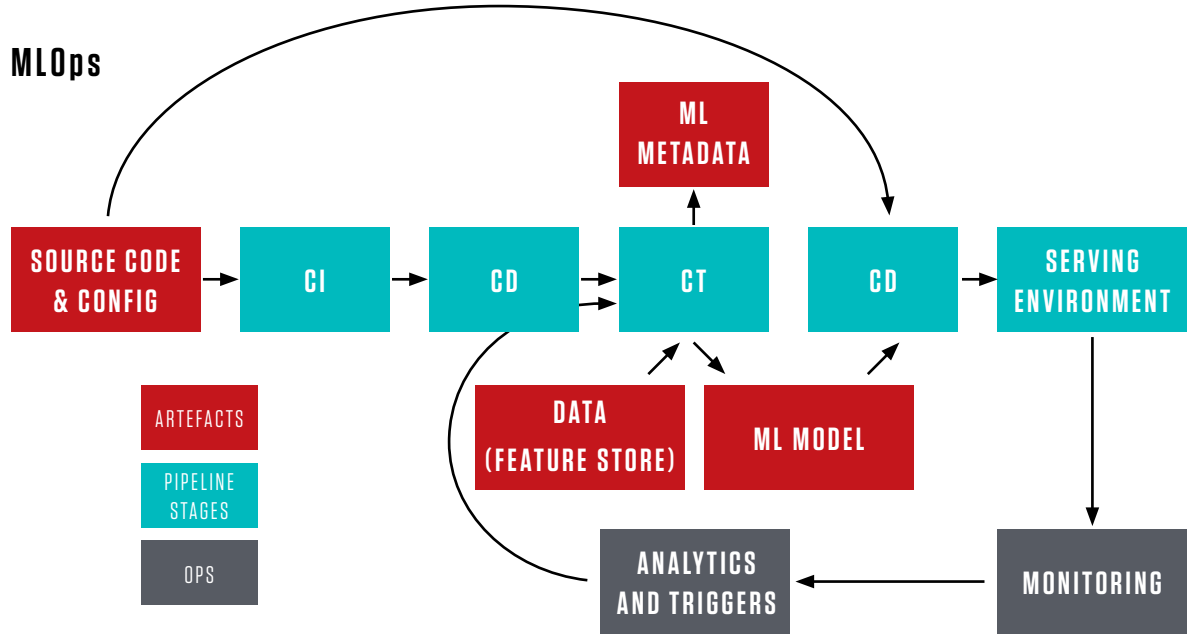

Figure 3: Typical MLOps process flow

Unlike DevOps, a successful MLOps pipeline must incorporate data, trained ML models, and ML metadata – as well as code and configurations – as first-order artefacts. This adds an additional layer of complexity.

The concepts of CI and CD have been extensively evaluated through DevOps. CI is the process of testing, linting, packaging, and compiling (for non-Python pipelines), while CD is the name given to the process of moving code, configurations and dependencies ('deployment packages') into a serving environment.

In MLOps the process is both less linear and less simple. In particular, a phase called Continuous Training must be introduced between CI and CD to automate the process of periodically re-training models.

In addition, ML code is divided roughly into training and inference. In CI training, code is packaged for deployment in the CT environment. A different packaging method is required for deployment packages used for inference serving. A serving deployment package requires the inference code to be brought together with a trained ML model, which is the result of the CT step.

There are two major serving patterns for ML models – batch and online. In batch, the goal is to use the ML model to calculate predictions using a large historical dataset all at once. It requires the ability to execute vectorized matrix operations in a scalable environment. As with other batch activities, the expectation is that that all calculations will be executed in bulk and that the output will become available in bulk as well. Depending on the amount of data and complexity of the model, such processes can take minutes, hours or more to run.

By contrast, an online deployment pattern aims to provision the ML model to provide near real-time calculations on demand via a REST API endpoint running on a webserver or a serverless platform. The goal is to perform a single prediction at a very fast speed. The expectation is that a response will be received by the caller of the API within milliseconds.

Monitoring is a key component of both DevOps and MLOps, providing an extensive layer of real-time metrics and logging to enable the proper administration of the systems in production. In DevOps the aim of this activity is to ensure a smooth, uninterrupted, reliable service.

This is also true for MLOps, but the monitoring function extends beyond that. The aim is to also collect metrics and log the data input into the systems for prediction so that it can be fed into an analytics layer where various anomalies can be detected.

An MLOps pipeline has an additional component around analytics and triggers. In that component extensive statistical analysis is performed on the input and output of production models to detect any biases, or interesting deviations, such as 'data drift'.

Data drift happens when we detect that the data fed into the models in the production system has a different distribution to the data that trained the model, to an extent that is statistically significant. This is important to know since it means the model is no longer making properly informed predictions, and we need to re-run the CT step to generate a fresh model that will perform better in production.

KEY BENEFITS OF MLOPS

MLOps is expanding as a fundamental practice in the financial services industry due to several key benefits:

Optimization and efficiency: The MLOps system offers faster model development, integration, and deployment into production, ensuring accelerated velocity to release. The Feature Store enables an ever-growing library of functionality that can be cut-and-pasted for new requirements. Combined with the ML Metadata Store and Model Registry, this helps to push high-quality ML models through the workflow quickly, maximizing business output.

Process confidence and risk mitigation: Before MLOps implementation, ML workflow was mainly a manual process. Integration, deployment and monitoring were challenging due to different stacks of software and different hardware components. MLOps helps to ease integration and automation, shifting the focus back to model inputs, features, and outputs. It also provides for a greater level of transparency and a faster response time to compliance and regulatory checks.

Reproducibility and scalability: Components such as the ML Metadata Store and Model Registry enable reproducibility from one model run, to the next. They also provide clear lineage and versioning, which are fed back into the pipeline to help identify and monitor the most optimally performing model. MLOps provides a platform for multiple models to be run and managed in parallel, ensuring scalability through modularity.

Scientific decision-making: Advanced MLOps systems leverage the automation of model development and deployment to encourage the wider breadth of experimentation involved in the creation of an AI-enabled system. They also allow for a data-driven scientific approach to finding a winning hypothesis with the most accurate predictions. The champion/contender deployment pattern allows multiple models to be run in parallel against production data, while serving predictions from the current champion model to clients. As soon as a contender model begins to consistently outperform the champion, it becomes the new champion, and the system starts serving its predictions to clients.

USE CASE – REVOLUT'S SHERLOCK[2]

Fintech company Revolut made use of several MLOps principles when designing its card fraud prevention system known as Sherlock. When a customer uses their Revolut card to make a payment, this information is sent back to Sherlock which, in less than 50ms, determines whether the transaction is likely to be fraudulent.

Sherlock makes use of several Google Cloud Platform components in order to achieve a full MLOps pipeline. An offline nightly batch job loads the transaction deltas of the day into a BigQuery database, from which the training data is generated and loaded back to BigQuery. This data is then used to train the classification model which is saved to Google Cloud Storage. The model is then served on an API that runs on Google App Engine allowing for rapid real time predictions. Model and system monitoring is achieved through Kibana dashboards, and Google Cloud's operations suite (formerly called Stackdriver).

After one year in production, Sherlock had reportedly saved $3 million, with 96% of fraudulent transactions caught, dramatically improving the experience for customers.

CONCLUSION – WHAT NEXT?

Although this may seem like a lot to cover, it is important to note that MLOps is not a flip switch solution. You can't, and more importantly, shouldn't aim to go from nothing to everything in one big step change. Following in the footsteps of early advocates of Agile, MLOps can be thought of as "a process of ongoing improvement." [1]

If you're already implementing ML without the MLOps' bells and whistles, you should strive to implement the concepts mentioned above to unlock its full potential. If you're still finding your way around ML, then this is a great opportunity to take some informed steps in this nascent field, learning from the mistakes and successes of early adopters. If you're struggling to find nooks where ML can make an impact in your company, then you probably need to revisit your data architecture and data engineering fundamentals to build a solid foundation upon which ML can thrive. Whatever state you're currently in, looking ahead to MLOPs will help flag up what is next on your roadmap to improvement.

1. Eliyahu M. Goldratt, The Goal: A Process of Ongoing Improvement
2. Dmitri Lihhatsov, Building a State-of-the-Art Card Fraud Detection System in 9 Months, November 2019: https://medium.com/revolut/building-a-state-of-the-art-card-fraud-detection-system-in-9-months-96463d7f652d

AUTHORS

Georgi Kanchev, Senior Consultant
James Hawrych, Consultant
Jack Forrest, Associate

CONTRIBUTORS

Chris Probert, Partner
Zaheer Khaled, Managing Principal

CONTACTS

Chris Probert, Partner, chris.probert@capco.com
Zaheer Khaled, Managing Principal, zaheer.khaled@capco.com
Georgi Kanchev, Senior Consultant, georgi.kanchev@capco.com
James Hawrych, Consultant, james.hawrych@capco.com
Jack Forrest, Associate, jack.forrest@capco.com

ABOUT CAPCO

Capco, a Wipro company, is a global technology and management consultancy specializing in driving digital transformation in the financial services industry. With a growing client portfolio comprising of over 100 global organizations, Capco operates at the intersection of business and technology by combining innovative thinking with unrivalled industry knowledge to deliver end-to-end data-driven solutions and fast-track digital initiatives for banking and payments, capital markets, wealth and asset management, insurance, and the energy sector. Capco's cutting-edge ingenuity is brought to life through its Innovation Labs and award-winning Be Yourself At Work culture and diverse talent.

To learn more, visit www.capco.com or follow us on Twitter, Facebook, YouTube, LinkedIn Instagram, and Xing.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Gurgaon
Hong Kong
Kuala Lumpur
Mumbai
Pune
Singapore

EUROPE
Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Munich
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO
a wipro company