Mastering MLOps Architecture: From Code to Deployment

Manage the production cycle of continual learning ML models with MLOps

Raman Jhajj

www.bpbonline.com

First published: 2024

Published by BPB Online
WeWork
119 Marylebone Road
London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55519-498

www.bpbonline.com

Dedicated to

My family, that gave me the gift of dreams
and
Friends, who became family.

About the Author

Raman Jhajj is a passionate leader in the data and software engineering space with experience building high-performing teams and leading organizations to become data-driven. He has experience in leading the development of SaaS applications, modern data platforms and MLOps infrastructure. He brings technical expertise across the data stack including AWS, Python, Django, Java, PostgreSQL, Hadoop, Spark, Kafka, Docker, CI/CD, SQL, NoSQL, and more.

Raman holds a master's degree in applied computer science from Georg-August University, Germany as well as a bachelor's in computer science from ICFAI University, India. After living in India, Germany, Austria, and Malta, he now calls Canada home.

Over the course of his career, Raman has driven key initiatives around modernizing data infrastructure, establishing data engineering capabilities, and building MLOps platforms.

Raman thrives on bringing cross-functional teams together to ensure alignment between technology and business goals. He has a proven track record of mentoring engineers and nurturing their potential.

When he is not working, you can often find him reading, writing, or exploring new places and cultures. He is passionate about using technology for social good, driven by a mission to leverage data engineering and AI for positive change.

About the Reviewer

Ashish Patel, an accomplished author, data scientist and researcher with over 11 years of experience. He is a luminary in predictive modeling, data preprocessing, feature engineering, machine learning, and deep learning. Notably, Ashish has taken center stage as a keynote speaker at prestigious events like AWS Community Day, AWS AI ML Days, Faculty Development Programs (FDPs), and IIT Techfest, captivating audiences with his insights. Currently serving as the Sr. AWS AI ML Solution Architect at IBM India Pvt Ltd, he architects innovation by collaborating with IBM and AWS specialists to craft enterprise solutions on Red Hat OpenShift, AWS Infrastructure, and IBM Software technology, aligning seamlessly with the AWS Well-Architected Framework. Ashish is a five-time LinkedIn Top Voice and an AI Research Scientist, with expertise spanning MLOps and a multitude of LLMs and FM Models. Recognized on LinkedIn for his contributions in Statistics, Data Science, Data Analytics, AI, and Machine Learning, Ashish is also a GitHub sensation with over 5k+ followers, marking his profound impact in open-source communities. In the realm where data reigns supreme, Ashish Patel crafts, speaks, and influences the future. He a Quantum Machine Learning practitioner and researcher working with international research community.

Acknowledgement

Writing a book is harder than I thought and more rewarding than I could have ever imagined. None of this would have been possible without the support of my family and friends, whom I would like to acknowledge and thank.

I would like to start by thanking my awesome wife, Simran for being the constant support from those late-night writing sessions and frustration-filled days to my ramblings of how hard it is to put thoughts on paper.

I want to thank my parents - Dad for constantly guiding and showing me that writing a book is an achievable target and Mom for her unwavering belief in me.

I thank Kanwar, Kuljeet and Garima for their constant support throughout the ups and downs of life and for always being there for me. I thank Kaisha for those video calls and for filling the days with laughter.

To all those friends who have been a part of my getting here: Parminder, Kiran, Anmol, Harman, Jagvir and Sukhpreet, I thank you for your heartfelt support and ready smiles, shared meals, advice, perspectives, and friendships. I thank Baani and Ravtaj for the playtime and for reminding me of what it is like to be a child again.

To my mentors throughout this journey: Malaika, Dean Chen, Michiah, and Tovah, I thank you for being the leaders I trust, honour, and respect.

To everyone at BPB Publications who enabled me to write this book. Thank you for the guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

I would also like to acknowledge the valuable contributions of my colleagues and co-workers who have taught me so much and provided valuable feedback on my work, during many years working in the tech industry.

Finally, I want to thank you, my cherished readers, for taking an interest in my book. To have it received by you is an unexpected gift that keeps me grounded in the moment.

Preface

MLOps is the intersection of DevOps, data engineering and machine learning. Working in the field of machine learning is highly dependent on ever-changing data, whereas MLOps is needed to deliver excellent ML and AI results. This book provides a practical guide to MLOps for data scientists, data engineers, and other professionals involved in building and deploying machine learning systems. It introduces MLOps, explaining its core concepts like continuous integration and delivery for machine learning. It outlines MLOps components and architecture, providing an understanding of how MLOps supports robust ML systems that continuously improve.

By covering the end-to-end machine learning pipeline from data to deployment, the book helps readers implement MLOps workflows. It discusses techniques like feature engineering, model development, A/B testing, and canary deployments.

The book equips readers with knowledge of MLOps tools and infrastructure for tasks like model tracking, model governance, metadata management, and pipeline orchestration. Monitoring and maintenance processes to detect model degradation are covered in depth. With its comprehensive coverage and practical focus, this book enables data scientists, data engineers, DevOps engineers, and technical leaders to effectively leverage MLOps. Readers can gain skills to build efficient CI/CD pipelines, deploy models faster, and make their ML systems more reliable and production-ready.

Overall, the book is an indispensable guide to MLOps and its applications for delivering business value through continuous machine learning and AI.

Chapter 1: Getting Started with MLOps - This chapter introduces MLOps, explaining how it combines machine learning, DevOps, and data engineering to enable continuous delivery of ML models. It covers the importance of MLOps, its principles like reproducibility and auditability, best practices, and strategies for implementation. The difference between MLOps and the traditional software engineering and the unique challenges of productionizing machine learning are also discussed. The chapter provides a foundation for understanding the MLOps methodology.

Chapter 2: MLOps Architecture and Components - This chapter covers the architecture and components of MLOps systems. It discusses the building blocks like data pipelines, model training, deployment, monitoring, and orchestration. The chapter outlines reference architectures for different maturity levels, from basic to enterprise-grade. It explains environment semantics and model deployment patterns. Finally, it walks through an end-to-end workflow integrating all components across development, staging, and production environments. The goal is to provide a foundation for designing and implementing MLOps solutions suitable for various use cases.

Chapter 3: MLOps Infrastructure and Tools - This chapter explores the infrastructure and tools needed for MLOps. It covers key components like storage, compute, containers, orchestration platforms, and ML platforms for deployment, model registries, and feature stores. The chapter discusses public cloud versus on-premises options, standardized development environments, and build versus buy decisions. It aims to provide guidance on setting up a robust, scalable infrastructure tailored to an organization's specific use cases and resources.

Chapter 4: What are Machine Learning Systems? - This chapter explains what machine learning systems are and how they differ from ML research. It covers an implementation roadmap with phases for initial development, transition to operations, and ongoing operations. The chapter discusses using standardized project structures like cookiecutter data science to facilitate eventual productionization. It aims to provide a foundation for taking a full systems approach to developing real-world ML applications, not just algorithms. The goal is to equip readers with an understanding of all components needed to build successful ML systems.

Chapter 5: Data Preparation and Model Development - This chapter covers data preparation and model development within the MLOps lifecycle. It discusses best practices for version control, preparing data, performing exploratory analysis, feature engineering, training models, and tracking experiments with MLflow. The chapter shows how these steps fit into a standardized project structure to enable collaboration and reproducibility. It aims to provide guidance on implementing key phases of the machine learning lifecycle in a way that facilitates eventual operationalization and automation.

Chapter 6: Model Deployment and Serving - This chapter covers model deployment and serving in the MLOps lifecycle. It explores strategies like static, dynamic, and streaming deployment, comparing deployment on devices versus servers using VMs, containers, or serverless technologies. The chapter discusses inference options like batch processing versus real-time APIs. It also looks at deployment patterns like canary releases and multi-armed bandits for controlled model rollout.

Chapter 7: Continuous Delivery of Machine Learning Models - This chapter explores methods for implementing continuous integration, continuous training, and continuous delivery in machine learning systems. It examines ML/AI pipelines and architectural maturity levels. Key topics include continuous integration tools like GitHub Actions, strategies for determining when and what to retrain models on, and considerations for rapidly deploying updated models into production through continuous delivery.

Chapter 8: Continual Learning - This chapter explores continual learning in machine learning systems, which involves models perpetually learning and adapting to new data without forgetting past knowledge. It covers principles like stateful training, challenges around obtaining fresh data and evaluating updates, and implementing continual learning in MLOps through triggers and robust monitoring. The goal is to enable frequent automated model updates while maintaining safety, transparency and control.

Chapter 9: Continuous Monitoring, Logging, and Maintenance - This chapter covers principles and best practices for monitoring machine learning models across environments. It examines why continuous monitoring matters, integrating it into MLOps workflows, logging model metadata and performance data, using frameworks like Evidently and Alibi Detect, and evaluating models with techniques like A/B testing.

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/mn9abap

The code bundle for the book is also hosted on GitHub at

https://github.com/bpbpublications/Mastering-MLOps-Architecture-From-Code-to-Deployment.

In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

Chapter 1
Getting Started with MLOps

Introduction

Being an emerging field, **Machine Learning Operations** (**MLOps**) is rapidly gaining momentum with data scientists, **Machine Learning** (**ML**) engineers, and **Artificial Intelligence** (**AI**) enthusiasts. In this chapter, we will go over the premise and background of the MLOps ecosystem. We will try to understand what it is, why it is useful, and what the principles and best practices are when it comes to MLOps. We will also go over what are the pillars of a successful MLOps strategy and how MLOps fits with the ROI requirements of a business.

When looking at MLOps, we can easily relate it to DevOps. DevOps did to software engineering what MLOps is aiming to do to machine learning engineering. DevOps is a culture, philosophy, and set of practices that seek to break down the barriers between development and operations teams, improve collaboration, and deliver software continuously and reliably. It involves the use of various tools and techniques for developing, testing, deploying, monitoring, and operating software engineering systems. DevOps was able to achieve the following for software engineering:

- Shorter development cycles
- Increased deployment velocity
- Automated testing before each deployment

- Auditable system releases
- Continuous monitoring of the system for stability and scalability

This brings us to MLOps. It is similar to DevOps, but with a focus on the unique requirements of machine learning and data-specific workflows. It involves the use of practices and tools for developing, testing, deploying, monitoring, and operating machine learning systems, while incorporating many of the same principles and practices of DevOps. No single solution is going to either make or break a plan. Instead, it is essential to understand the unique requirements of what frameworks might fit into your workflow and have a comprehensive strategy to implement that. In this chapter and throughout the book, we will learn how that is achieved. Next, we will discuss the principles and fundamentals of MLOps and how to use them effectively to get models into production successfully.

Structure

In this chapter, we will discuss the following topics:

- Understanding MLOps
- Importance of MLOps
- The evolution of MLOps
- Software engineering projects versus machine learning projects
- DevOps versus MLOps
- Principles of MLOps
- MLOps best practices
- MLOps in an organization
- MLOps strategy
- Implementing MLOps
- Overcoming challenges of MLOps

Objectives

By the end of this chapter, you will have a solid understanding of MLOps and the reason behind its hype. We will also learn about the fundamental principles and best practices of MLOps, including reproducibility, transparency, auditability, and scalability.

You will understand the difference between software engineering projects and machine learning projects and how that impacts the need for MLOps versus traditional DevOps. We will also cover the evolution of MLOps over time.

We will discuss the role of MLOps in an organization and why having a good MLOps strategy matters for successful implementation, and how organizations can unlock business value from MLOps while overcoming inherent challenges in the machine learning system implementation.

The chapter will also provide an overview of implementing MLOps in different environments and how vendors and open-source solutions can accelerate implementation.

Understanding MLOps

MLOps is a set of practices designed for collaboration between data scientists, machine learning engineers, data engineers, and operations professionals. MLOps is the answer to the questions:

- Why is machine learning deployment not quick?
- How can we quickly productionize our machine learning models?
- Why can machine learning model deployment be ten times faster?

MLOps is a combination of **Machine Learning** (**ML**) and **Operations** (**Ops**). It refers to the processes and practices for designing, building, enabling, and supporting the efficient deployment of ML models in production and continuously iterating and improving upon these models.

Similar to DevOps, MLOps is heavily dependent on automation and integrations. MLOps aims to standardize the deployment and management of ML models alongside the operationalization of the ML pipeline. It supports the release, activation, monitoring, performance tracking, management, reuse, maintenance, and governance of ML artifacts.

Following and applying this set of practices simplifies the management of models and artifacts, automates the deployment of machine learning models, allows us to maintain data and artifact lineage, and improves the quality and speed of deployment. Implementation of these practices makes it easier to iterate over model development quickly and better align models with business needs/requirements.

MLOps combines and is at the intersection of **Machine Learning**, **DevOps**, and **Data Engineering**, as shown in *Figure 1.1*, with the goal of reliably and efficiently building, deploying, and maintaining ML systems in production. It is at the intersection of **DevOps**, **Data Engineering**, and **Machine Learning**. Machine learning projects and overall systems are experimental in nature. It consists of components that are comparatively more complex to build and operate than DevOps components. Other than the building and deployment, MLOps also needs to account for new components like data drift, the delta between changes in the data from the last model training and current model training, and so on. Refer to *Figure 1.1:*

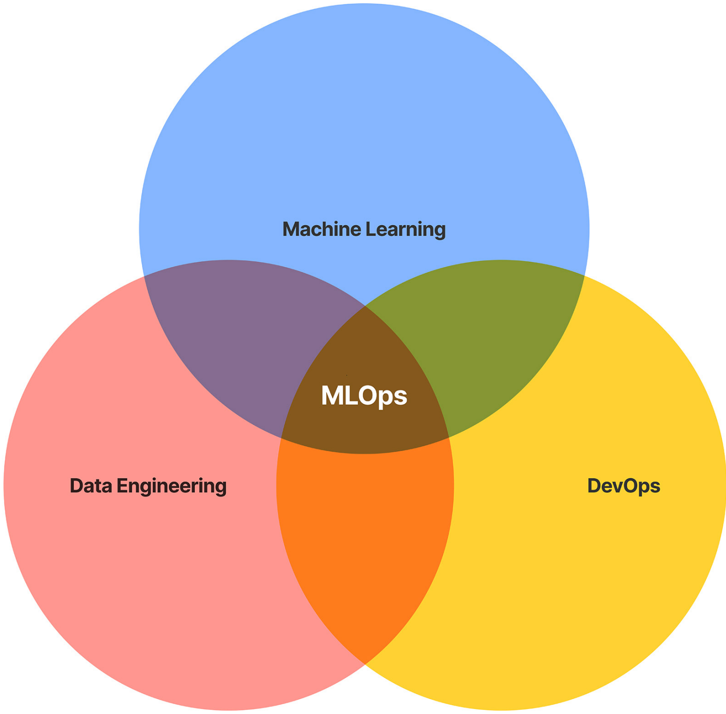

Figure 1.1: *MLOps as an intersection of three domains*

With the base driven by DevOps, MLOps is now slowly evolving into an independent approach to machine learning lifecycle management. It applies to the entire lifecycle and key phases being:

- Data gathering, collecting, and processing raw data
- Data analysis
- Data preparation
- Model training and development
- Model evaluation and validation
- Model serving
- Model health monitoring
- Model re-training and iterations
- Orchestration
- Governance

These key phases indicate how work-intensive the entire process can get, especially since it will most likely need to be repeated multiple times. While it is possibly easier the second time around since we only must update the model on new data patterns and trends, it is still a problem that can take up hours of manual labor. After all, the maintenance of applications in the software development process is usually where most of the money and resources go, not the initial development and release of the application. The same can apply to machine learning models and processes, worsening the overall maintenance costs.

Figure 1.2 shows the relationship between all the key phases of a machine learning pipeline and how these phases fit together to allow us to build a complete pipeline. Refer to the following figure:

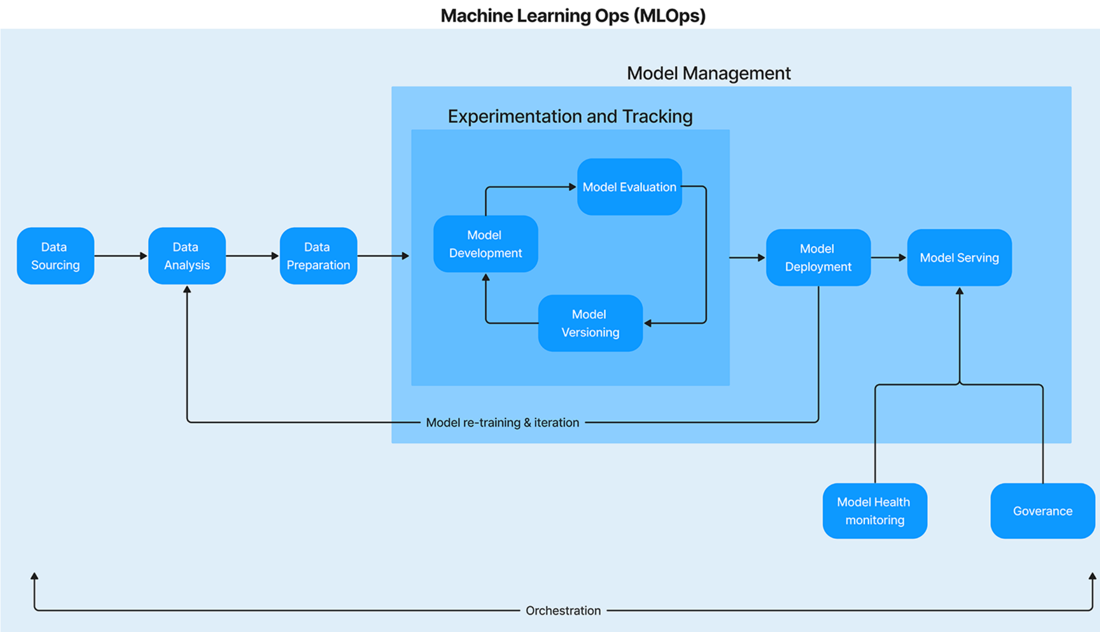

***Figure 1.2**: Machine learning project lifecycle*

Imagine if we could simply automate this entire process away, allowing us to take full advantage of high-performance machine learning models without all the hassle. This is where MLOps comes in.

In *Figure 1.2*, you will notice there are two more components that are part of MLOps: **Experimentation and Tracking** and **Model Management**. What are those, and how can we define them?

Experimentation and tracking

Experimentation and tracking are parts of MLOps, which focus on collecting, organizing, and tracking model training information and artifacts across multiple runs, and using multiple configurations. As machine learning is experimental in nature, using experiment-tracking tools to track and benchmark different models and configurations becomes important.