Hakkōda

The Little Handbook of MLOps Engineering

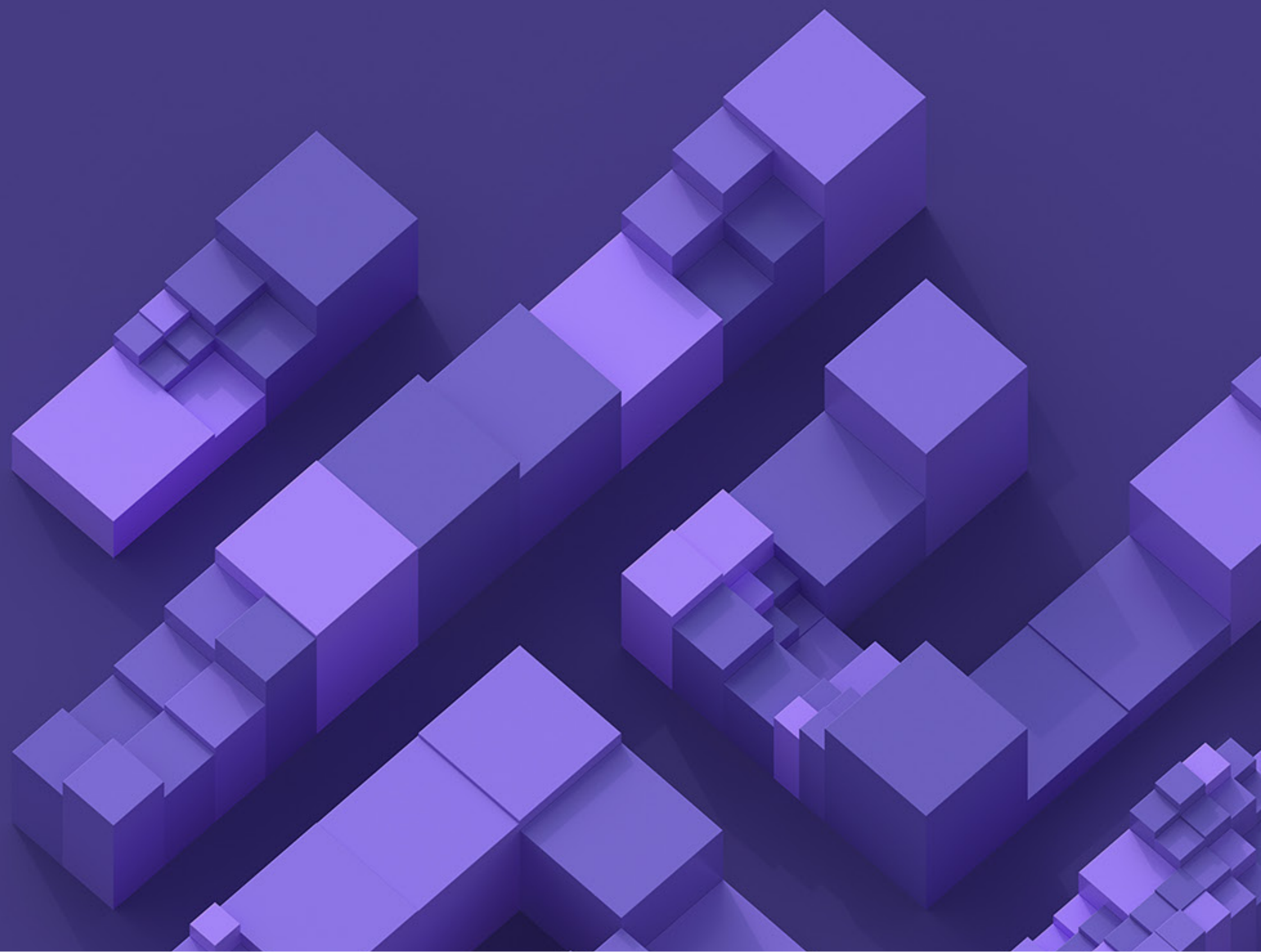
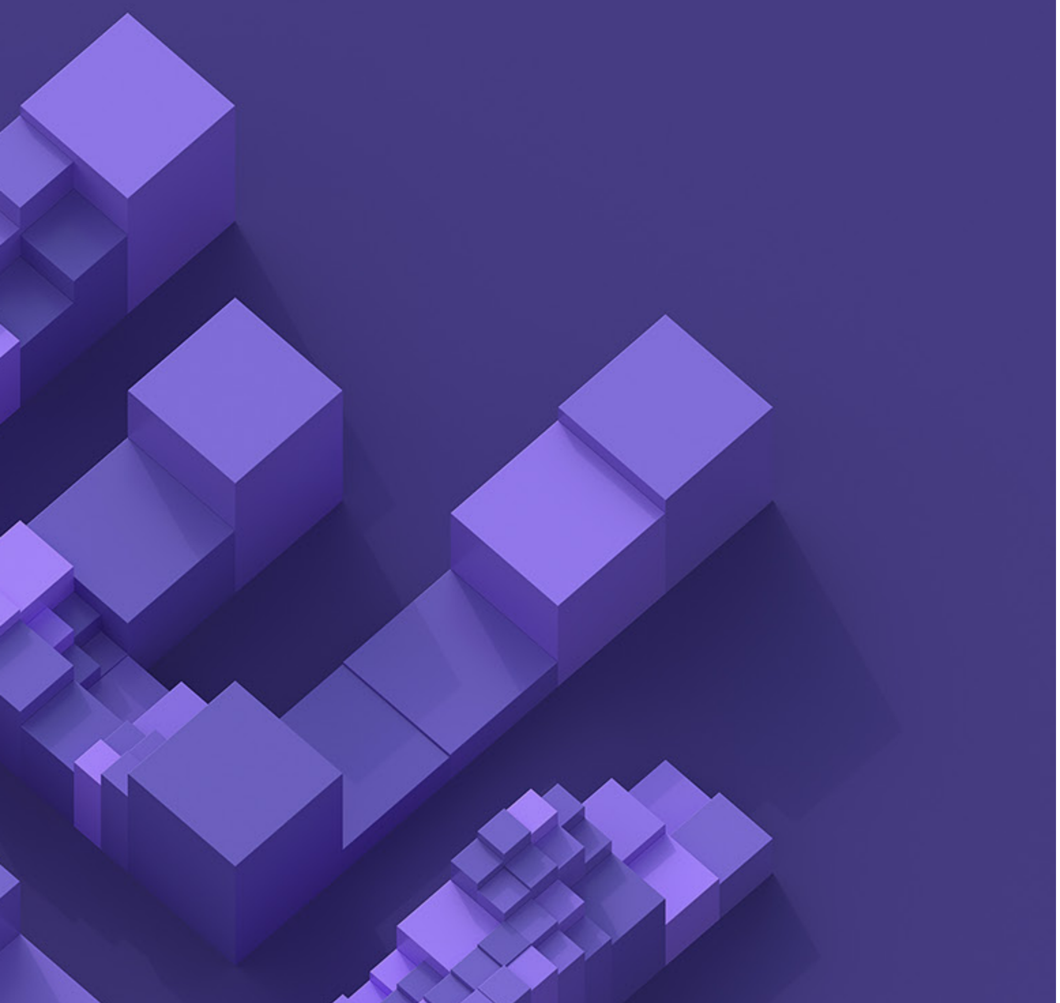


Table of Contents



1 Introduction

When it comes to managing machine learning models, businesses of all kinds are at a crossroads. Many of the machine learning (ML) models created exclusively for small scale situations are unmanageable at a company-wide level, and like business goals,ML models and their dependencies change over time.

Imagine a machine learning model that works efficiently to predict the most popular film genres. If you were to try to use the same model later (it really varies, depending on the type of model), you'd find that the model is inaccurate. Because it was calibrated to a specific data set, the model will be inaccurate until it has been calibrated again to work with a new data set.

These types of situations are manageable at a small scale, especially when a company only works with a handful of models used in a handful of departments. However, the moment an organization wishes to expand model use, the number of dependencies immediately increases. Managing business needs, processing changing data and making sure that "the reality of the model in production and on production data aligns with expectations"[1] are just a few elements that influence the machine learning model life cycle.

Enter the field of machine learning operations, or MLOps, a concept that pulls from development operations (DevOps) to streamline automation and trust between teams, creating a robust end-to-end machine learning model life cycle that prioritizes data quality.[2] MLOps offers a series of frameworks that help deploy machine learning models into production, a practice that couldn't be achieved on an enterprise level using only data scientists.

In this guide, we'll detail what MLOps is and offer concrete best practices for building transparency and accountability in your businesses with MLOps.

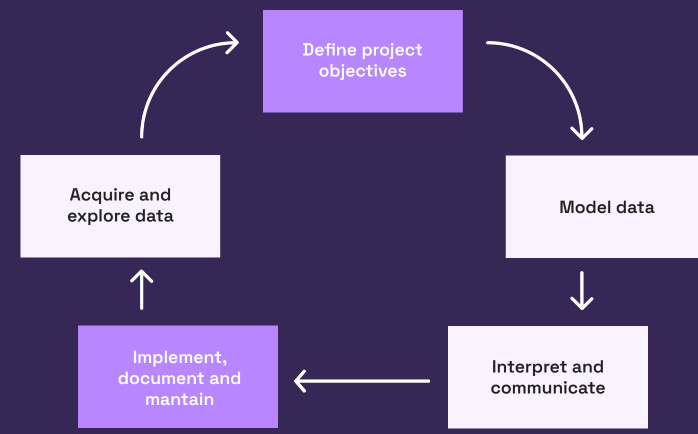
[1] Visengeriyeva, L. et al. MLOps Principles.
[2] Treveil M. and the Dataiku Team. Introducing MLOps.

2 Understanding the Machine Learning Cycle

To understand how MLOps works, it’s important to possess a basic understanding of the steps machine learning processes comprise. As Data Robot notes,[3] the machine learning cycle has five major steps, all of which have equal importance and can’t be skipped or shifted. Those five major steps are:

1. **Define project objectives**
2. **Acquire and explore data**
3. **Model data**
4. **Interpret and communicate**
5. **Implement, document and maintain**

The machine learning cycle always begins by first defining what the problem is and what is required for successful resolution. Once this first step is complete, developers and engineers can begin extracting data from different sources and then performing an exploratory data analysis. This analysis not only provides initial insight into the project, it helps experts evaluate whether their original plan is the correct strategy for tackling the problem.



[4] Figure 1: The machine learning life cycle.

Exploring the data lays the groundwork for data modeling, a stage that includes data preparation and cleaning, variable selection, and model validation and selection.[5] It is also in this step that experts define the target variable. The target variable is the factor about which they hope to gain a greater understanding. After the data has been prepared, data engineers and data scientists can move the model into production, where it can be evaluated using test data.

Interpreting and communicating results is a key step in the machine learning cycle because it provides a much-sought after level of insight. Once the results of a model have been translated to the public, stakeholders can visualize how the company should proceed. Finally, in the implementation and maintenance step, engineers in any data science project can move to production and improve upon models.

[3] Data Robot. Machine Learning Life Cycle.
[4] Data Robot. Machine Learning Life Cycle.
[5] Treveil M. and the Dataiku Team. Introducing MLOps.

The diagram described above is one of the most basic variations of the machine learning life cycle. However, as with any business, specific workflows and use cases can add additional nuance. Another way of visualizing the machine learning lifecycle is depicted below,[6] offering more advanced developers a framework to align to MLOps best practices by focusing on four distinct components:

1. **The Well-Architected Design Principles**
2. **The Well-Architected Machine Learning Lifecycle**
3. **The Cloud and Technology Agnostic Best Practices**
4. **The ML Lifecycle Architecture Diagrams**

The Machine Learning Lifecycle phases referenced within this framework are business goal identification, ML problem framing, data processing (which includes data collection, data pre-processing and feature engineering), model development (training, tuning and evaluation), model deployment and model monitoring. As seen in Figure 2, the Well-Architected Machine Learning Lifecycle introduces more steps and considerations, offering a more complete version of the cycle.

Nearly everyone, in every organization, in nearly every function, uses data. As cliché as it is to say, data is the lifeblood of business.

Futurists imagine a data-driven utopia in which that lifeblood flows effortlessly throughout organizations–driving frictionless customer experiences, more informed real-time decisions, and innovative new offerings that elevate both brand and revenue.

But realists see the current state of data today–"pockets of greatness" in organizations held down by data sprawl. Sprawl that keeps data fragmented, hard to use and even harder to manage. Sprawl that keeps driving costs up exponentially, keeping the promised data-driven utopia just out of reach.

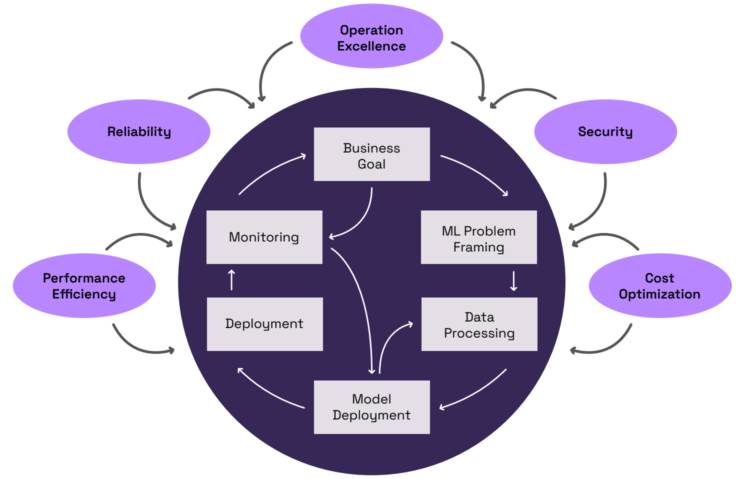


Figure 2: AWS' Well-Architected Machine Learning Lifecycle.

[6] AWS. Introducing the new AWS Well-Architected Machine Learning Lens.
[7] AWS. Introducing the new AWS Well-Architected Machine Learning Lens.

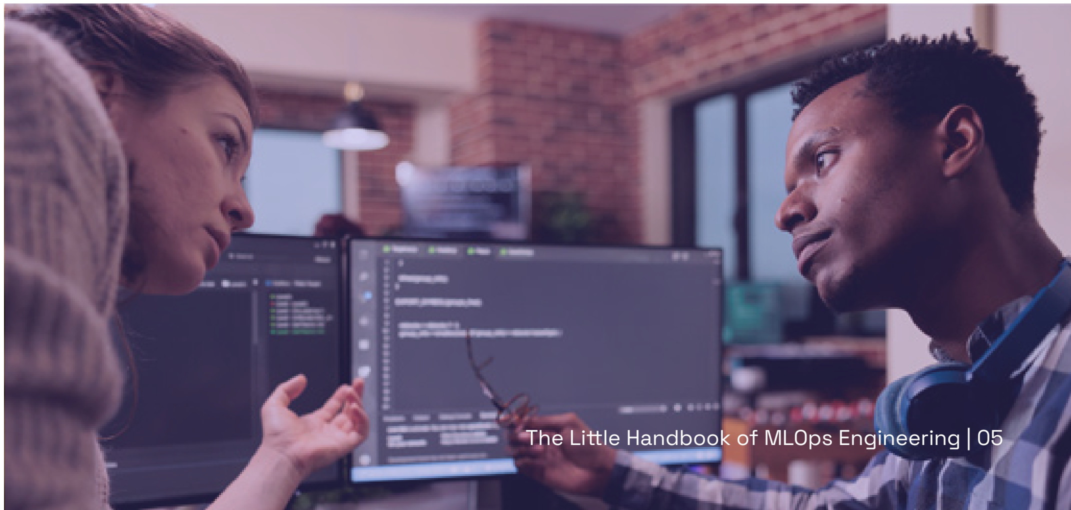
What difference does the Well-Architected Machine Learning Lens make in current projects? First off, it helps developers and engineers plan early and make informed decisions. Second, the framework introduces best practices that prove useful when navigating workloads. It also aids teams when evaluating their current workloads and identifies existing issues.

3 Scalability with the Machine Learning Cycle

Managing a machine learning lifecycle at scale can be challenging depending on the number of dependencies involved. Any machine learning model a team creates is aligned with business goals, business expectations, production data and many other factors. A slight change in any of these factors will require making modifications on the model itself.[8] If you're not using MLOps in production, you may end up needing to perform additional, sometimes manual steps that create a high likelihood of error. For example, a locally trained model might have to be manually moved to something like an Airflow DAG.

This process is likely to result in occasional human errors, and with very little governance or ability to trace back artifacts, the risk to your MLOps lifecycle is great.

Multiple dependencies within a single model also requires sharing a common language. A single cycle in an MLOps project with multiple dependencies is likely to include stakeholders across a variety of disciplines, ranging from data engineers and data scientists to subject-matter experts (SME). Because none of these disciplines use the same tools or skills, finding a way to understand each other is fundamental. Before you undertake an MLOps project with multiple dependencies, you'll need to establish common metrics, success criteria, and tracking methodology that all of project stakeholders feel comfortable with and understand. This preparation phase will make or break your project at scale.



[8] Treveil, M. and the Dataiku Team. Introducing MLOps.

As models scale, either in size or scope, the difficulty of managing different disciplines and languages increases. As Mark Treveil, an experienced MLOps veteran points out, even employment changes represent a variable that must be accounted for in your MLOps projects: "The complexity becomes exponential when considering the turnover of staff on data teams."[9] A good example of just how many people are involved within a single machine learning cycle can be observed in Figure 3, where an average organization can require up to eight professionals.

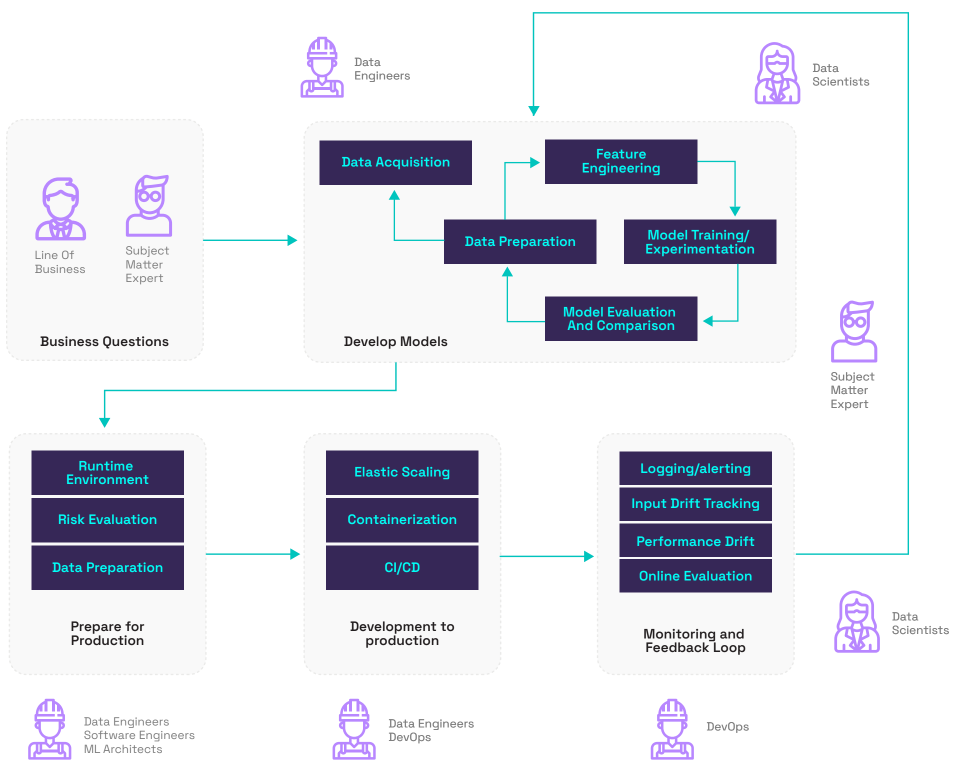


Figure 3: What a machine learning life cycle looks like, in terms of people involved.

[9] Treveil, M. and the Dataiku Team. Introducing MLOps.

Scalability can also be affected by the fact that machine learning models become stale with time. This phenomenon, known as data drift, occurs when the production data differs from the data used to test and validate the model. Data drift frequently occurs with old machine learning models, especially when the models haven't been retrained regularly on new data sets. And although there are many methods to continuously train a model, doing so may require tiring manual work.

An MLOps approach can not only help experts automate model retraining, it also helps administer the different dependencies involved and ensure the entire team works together seamlessly.

> **"While the tools to run an AI project have become more accessible, stakeholders today still face huge gaps in organizational knowledge that make it difficult for cross-functional teams to comprehend the scope of a data project."**

Patrick Buell, Vice President of Consulting at Hakkoda

4 The Fundamentals of MLOps

MLOps is not a methodology per se, it's a mindset that encompasses best practices and tools to monitor machine learning models in real-world production. An optimal MLOps experience is "one where Machine Learning assets are treated consistently with all other software assets within a CI/CD environment."[10] In other words, it's possible to deploy ML models alongside different types of services that accompany them and create a single, unified release process.

The idea behind creating a system of practices is to accelerate the adoption of artificial intelligence, and more specifically machine learning, in software. To do so, MLOps works as an iterative-incremental process divided into three components. The first component focuses on design, diving into business understanding, identifying the potential user and designing the machine learning model. Throughout this initial phase, it's important to inspect the data available and clearly define the requirements for the ML model.

Once the team understands the data set and has agreed on a model, they can move on to experimentation and development. Such a component is devoted to verifying the applicability of the model to make sure that it will run in production. The final component delivers the model to production, all while following established DevOps practices. It's important to know that all three components communicate and influence each other.

The decisions taken when designing, for instance, will affect development, experimentation and deployment later on.

Even though MLOps does share some similarities with DevOps, especially when following continuous integration and continuous delivery (CI/CD) principles, there are substantial differences. As Mark Treveil notes, "Within an MLOps cycle, teams may not necessarily deploy a model artifact —it's usually a lot more than that. Deployment of a machine learning system involves many processes, such as data extraction and model training.

There are concepts that are unique to the machine learning cycle (and, therefore, to the MLOps cycle) that don't exist within the DevOps sphere. Within a machine learning cycle, experts must focus on code testing and data quality. In other words, it's also important to make sure that the data being used fits for its intended use. Finally, MLOps handles a wide variety of unique concepts, such as continuous testing (CT)."[11] These concepts will be explained in depth in the next section.

[10] Visengeriyeva, L. et al. MLOps Principles.
[11] Treveil M. and the Dataiku Team. Introducing MLOps.

5 Other Relevant Concepts Within MLOps

The MLOps process relies on three core principles, two of which are present within the DevOps framework. These are continuous integration (CI), continuous delivery (CD) and continuous testing (CT).

To shorten development cycles and increase deployment velocity, DevOps rely on continuous integration and continuous delivery. Amazon Web Services defines the first term as a "development practice where developers regularly merge their code changes into a central repository, after which automated builds and tests are run."[12] With continuous integration, developers run several local unit tests before committing to a shared repository. When working with machine learning models, however, CI also includes testing and validating data, data schemas and models.

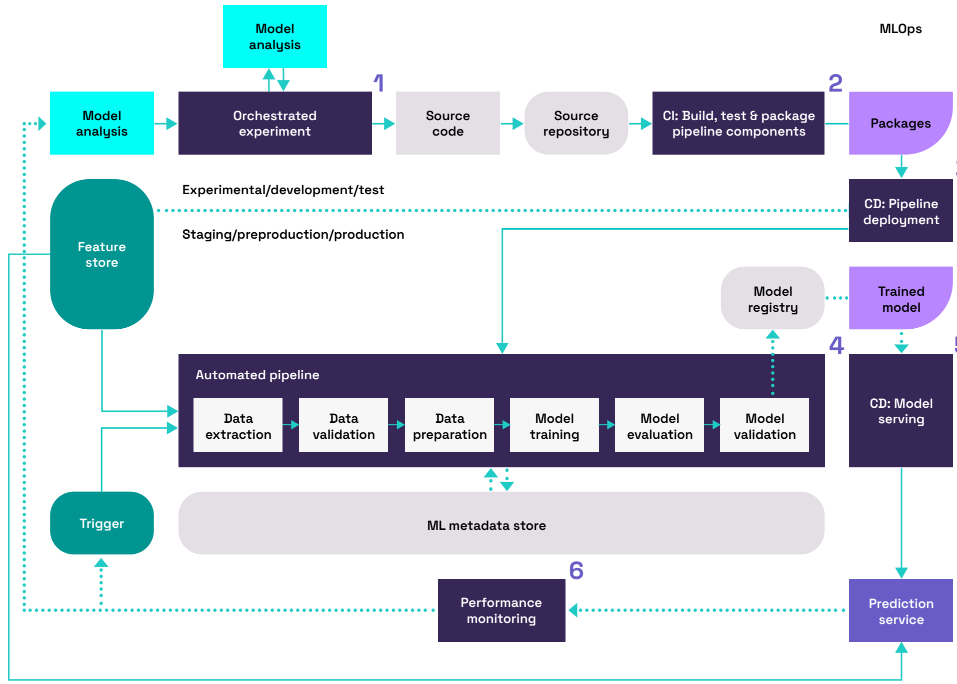


Figure 4: A CI/CD automated ML pipeline.[13] See "The Different Levels of ML Pipelines" for more on fully automated (or level 2) pipelines.

Continuous delivery, on the other hand, "is a software practice wherein code changes are automatically prepared for a release to production."[14] In a DevOps scenario, continuous delivery builds upon continuous integration and deploys all code changes to a testing environment. These tests go beyond simple unit tests, and usually include UI testing, API reliability testing and so on. If these concepts are properly implemented within a development process, developers will always have a deployment-ready build.[15] Finally, continuous testing (CT) helps teams incorporate automated retraining and serving machine learning models.

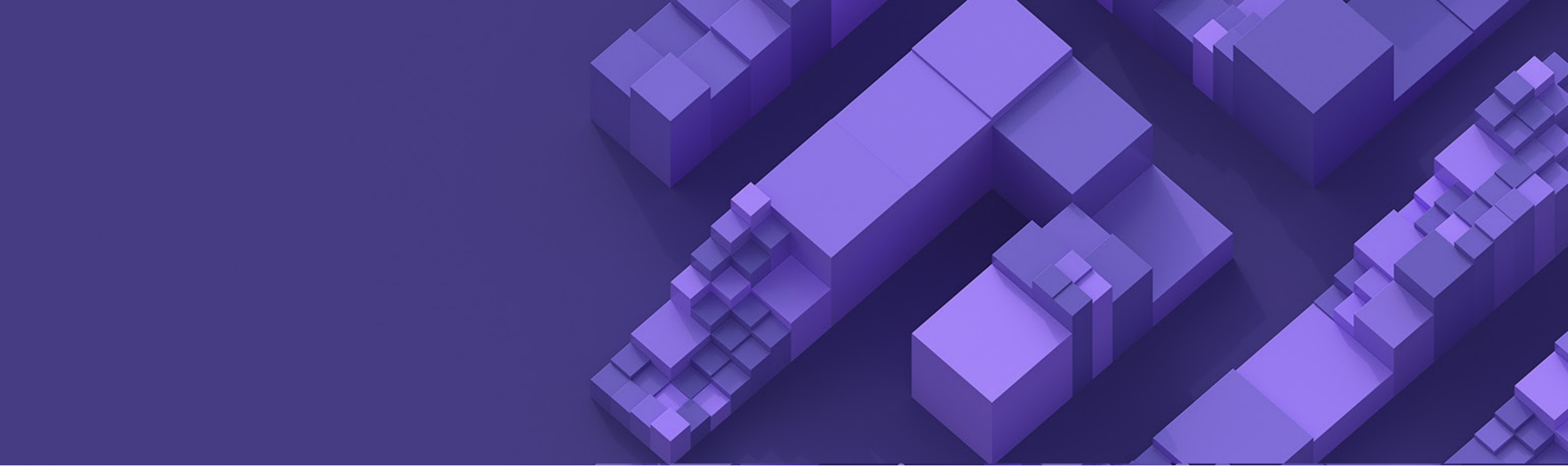
[12] Data Robot. Machine Learning Life Cycle.
[13] Data Robot. Machine Learning Life Cycle.
[14] Treveil M. and the Dataiku Team. Introducing MLOps.
[15] Susilo, M. MLOps Level 1.

6 The Different Levels of MLOps Pipelines

Although many data researchers and data scientists can build effective machine learning models, implementing an MLOps pipeline requires time, resources and knowledge. Google's Cloud Architecture Center[16] has created three levels of process implementation or maturity, which are defined below.

	CHARACTERISTICS	CHALLENGES
MLOps level 0	Every step is manual and there is a clear disconnection between ML and operations. Both CI and CD are ignored or not considered, and there is a lack of active performance monitoring. Finally, although the team might have set up their own set up for testing and deployment, all of these processes are done manually.	Models often break when deployed or fail to adapt to changes in the environment. Experts need to frequently retrain their models manually and continuously experiment to produce new model variations.
MLOps level 1	Allows teams to achieve the continuous delivery of the model and automate some parts of the process. Generally, the MLOps level 1 setup includes a rapid orchestration of the ML experiment steps, experimental-operational symmetry, modularized code for components and pipelines, and the CT of the model in production. For this setup, experts may consider adding a feature store to the pipeline automation.	New pipeline implementations, especially when managing a few pipelines, must be done manually. This setup requires submission of the tested source code to the IT team in order for deployment, which can be suitable for new data but not for new ML ideas.

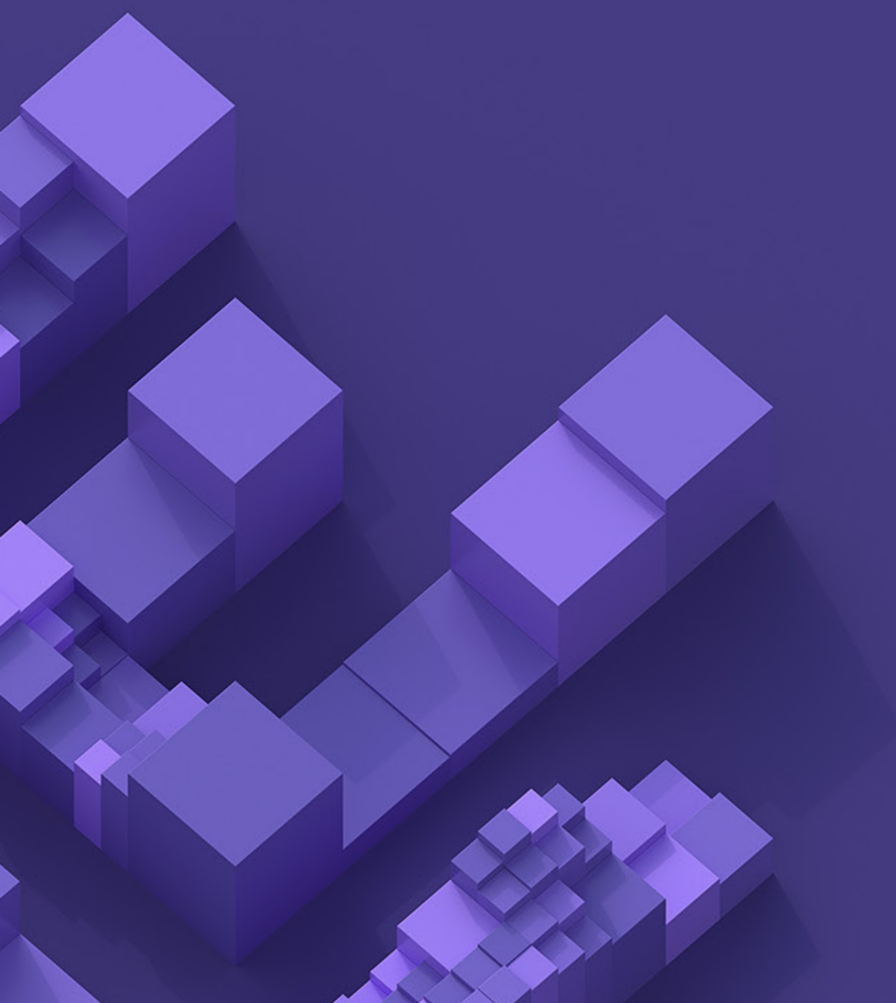
[16] Google Cloud Architecture Center. MLOps: Continuous delivery and automation pipelines in machine learning.



	CHARACTERISTICS	CHALLENGES
MLOps level 2	An automated CI/CD system includes a source control, test and build services, deployment services, feature store, ML metadata store and an ML pipeline orchestrator. An automated pipeline consists of five stages: development and experimentation, pipeline continuous integration, pipeline continuous delivery, automated triggering, model continuous delivery and monitoring.	At this level, there are many aspects that experts need to consider to enable reliable continuous delivery of pipelines and models. Reaching this stage requires a gradual implementation of MLOps practices in order to help improve automation.

7 Leveraging MLOps in the Modern Data Stack

When migrating to a modern data stack, companies and business leaders often struggle to create streamlined pipelines that ensure their models work consistently. Tools such as Dataiku, Snowpark and Data Robot help data engineers and data scientists systematize and manage their machine learning pipelines in a single location.



Within the Dataiku DSS, teams can prepare, build, deploy and monitor their MLOps life cycles at scale. In the preparation stage, data experts can clearly define model expectations and set data discovery parameters such as connector richness and feature store. Moving on to the model build, a data team can use Dataiku to specify model development and model readiness. This includes the ability to perform model testing, model and project versioning and native integration.

When it comes to deployment, Dataiku oversees every aspect related to model sign-off, performing self-service deployment using Dataiku Deployer, and deploying and running models in dedicated environments. This tool also provides great support in the final step of the MLOps life cycle. Dataiku not only tracks training and performance metadata for all runs, it also monitors external models using their prediction logs.

DataRobot, on the other hand, is an autoML platform that requires no engineering skills for its use. Its MLOps center provides a single space to deploy, monitor, manage, and govern all models. DataRobot requires a quick configuration and a few lines of code to help teams manage the AI production lifecycle.

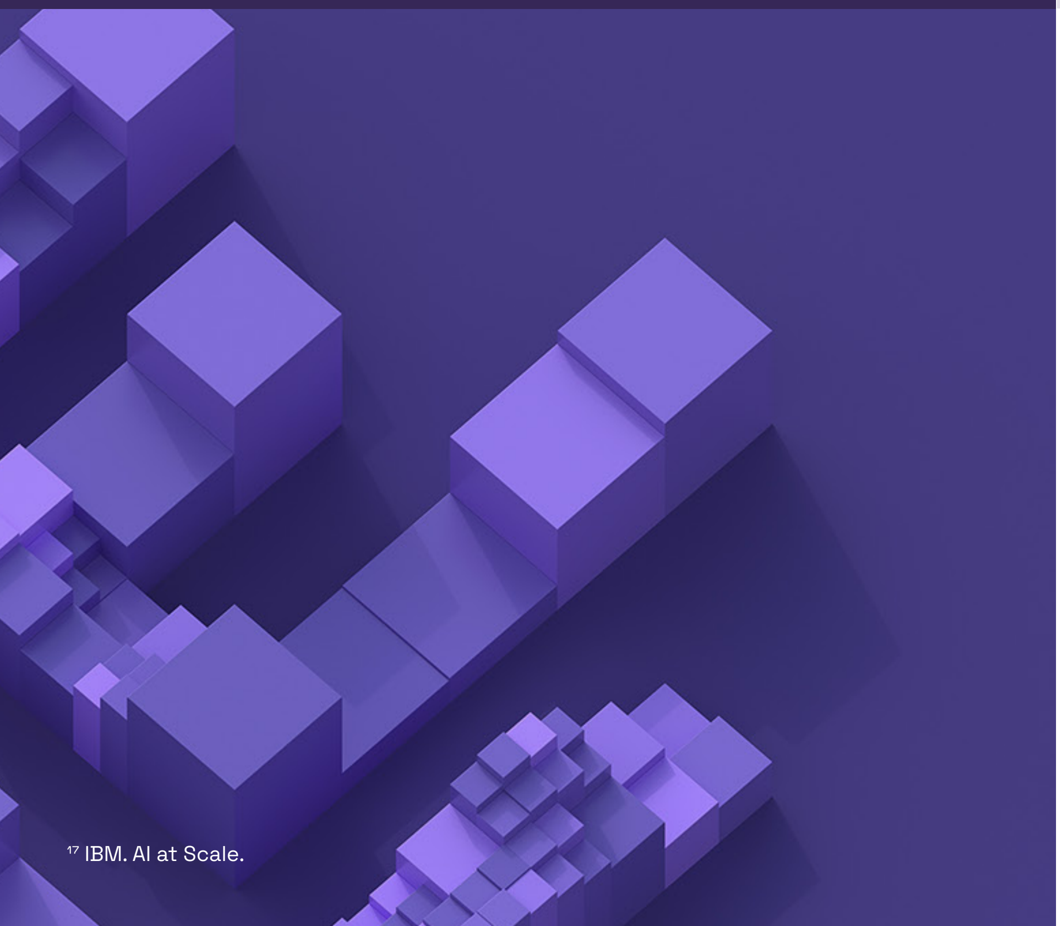
Aside from the partners mentioned above, whose platforms are continuously used by the Hakkoda team to manage machine learning pipelines, the Hakkoda Data Science Team has created a series of tools to aid data experts throughout the MLOps lifecycle. One of these is Model Buddy, an accelerator that "centralizes model monitoring so teams can overcome the challenge of maintaining and organizing models once they're implemented."

8 Why Investing in an MLOps Strategy Is the Future

The machine learning cycle relies on data talent capable of programming in Python, SQL and experienced in libraries such as Scikit-learn, TensorFlow, PyTorch, and Pandas. All of the technologies we've mentioned enable data scientists to develop machine learning models, but creating a machine learning model is only part of the cycle, and it isn't enough.

90%

of corporate AI initiatives struggle to move past test stages.



According to IBM[17], almost 90% of corporate AI initiatives struggle to move past test stages. Although many are making great advancements within data science, organizations need to integrate data science teams with experts capable of developing and managing the infrastructure required to make the models work. MLOps as a practice offers a unique solution, binding together data science, business teams, data engineering and even DevOps to support machine learning success. For teams lacking robust internal resources, a modern data services provider like Hakkoda, with teams of experienced SMEs, data scientists, architects, and business analysts can be a way to immediately staff and ensure the success of your AI project by tapping into experts who have implemented similar projects.

[17] IBM. AI at Scale.

“if you discovered an algorithm that will save you a million dollars per month, every month this model isn’t in production or deployment costs you $1 million.”

Randy LeBlanc, VP of Customer Success at RapidMiner

Ensuring machine learning models work and scale through a well-structured MLOps practice can unlock incredible business potential for any organization. Randy LeBlanc, VP of Customer Success at RapidMiner mentions that “if you discovered an algorithm that will save you a million dollars per month, every month this model isn’t in production or deployment costs you $1 million.”[18] In other words, implementing an MLOps strategy allows one to take faster and more effective action, quickly capitalizing over business opportunities.

One of the key points of using technology within many industries is leveraging advancements to improve or streamline processes that would otherwise require time and resources. Patrick Buell, Vice President of Consulting at Hakkoda, states that “while the tools to run an AI project have become more accessible, stakeholders today still face huge gaps in organizational knowledge that make it difficult for cross-functional teams to comprehend the scope of a data project.” MLOps is gathering great traction because it ensures novel machine learning models work and improve upon real problems. Because of this, the MLOps solutions market is expected to reach $4 billion by 2025.[19]

[18] IBM. AI at Scale.
[19] Taulli, T. Forbes Magazine.

9 References

1. Amazon AWS. (2023) What is Continuous Integration? Retrieved from: https://aws.amazon.com/devops/continuous-integration/

2. Amazon AWS. (2023) What is Continuous Delivery? Retrieved from: https://aws.amazon.com/devops/continuous-delivery/

3. AWS. (2021) Introducing the Well-Architected Machine Learning Lens. Retrieved from: https://aws.amazon.com/blogs/architecture/introducing-the-new-aws-well-architected-machine-learning-lens/

4. Dataiku. (2022) MLOps with Dataiku. Dataiku Product Days. Retrieved from: https://content.dataiku.com/oreilly-mlops

5. Data Robot. (2023) Machine Learning Life Cycle: What is the Machine Learning Life Cycle? Retrieved from: https://www.datarobot.com/wiki/machine-learning-life-cycle/

6. Google Cloud Architecture Center. (2023) MLOps: Continuous delivery and automation pipelines in machine learning. Retrieved from: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#top_of_page

7. Google Cloud Architecture Center. (2023) DevOps tech: Continuous testing. Retrieved from: https://cloud.google.com/architecture/devops/devops-tech-test-automation

8. IBM. (2023) What is continuous testing? Retrieved from: https://www.ibm.com/topics/continuous-testing

9. IBM. AI at Scale. Retrieved from: https://www.ibm.com/blog/ai-at-scale/?c=Artificial%20Intelligence

10. Susilo, M (2021) MLOps Level 1: Continuous Training. Towards Data Science. Retrieved from: https://towardsdatascience.com/mlops-level-1-continuous-training-b2a633e27d47

11. Susilo, M (2021) MLOps vs DevOps. Towards Data Science. Retrieved from: https://towardsdatascience.com/mlops-vs-devops-5c9a4d5a60ba

12. Visengeriyeva, L. et al. (2023) MLOps Principles. MLOps.org. Retrieved from: https://ml-ops.org/content/mlops-principles#summary-of-mlops-principles-and-best-practices

13. Taulli, T. (2020) MLOps: What You Need to Know. Forbes Magazine. Retrieved from: https://www.forbes.com/sites/tomtaulli/2020/08/01/mlops-what-you-need-to-know/?sh=20ca39291214

14. Travel, M. and the Dataiku Team. (2021) Introducing MLOps. O'Reilly Media. United States.