fractal

Implementation roadmap for your MLOps journey

Microsoft
Solutions Partner

Content

Introducing MLOps

MLOPs in a nutshell

Machine Learning Operations (MLOps) is a **set of policies, practices, and governance** best practices that are put in place for **managing machine learning models solutions throughout their lifecycle**.

MLOps is collaborative in nature, enabling data science and IT teams to collaborate and boost model development and deployment pace by monitoring and validating machine learning models at scale.

MLOps allows data scientists to track and certify every asset in the ML lifecycle and provides integrated solutions to streamline models' lifecycle management.

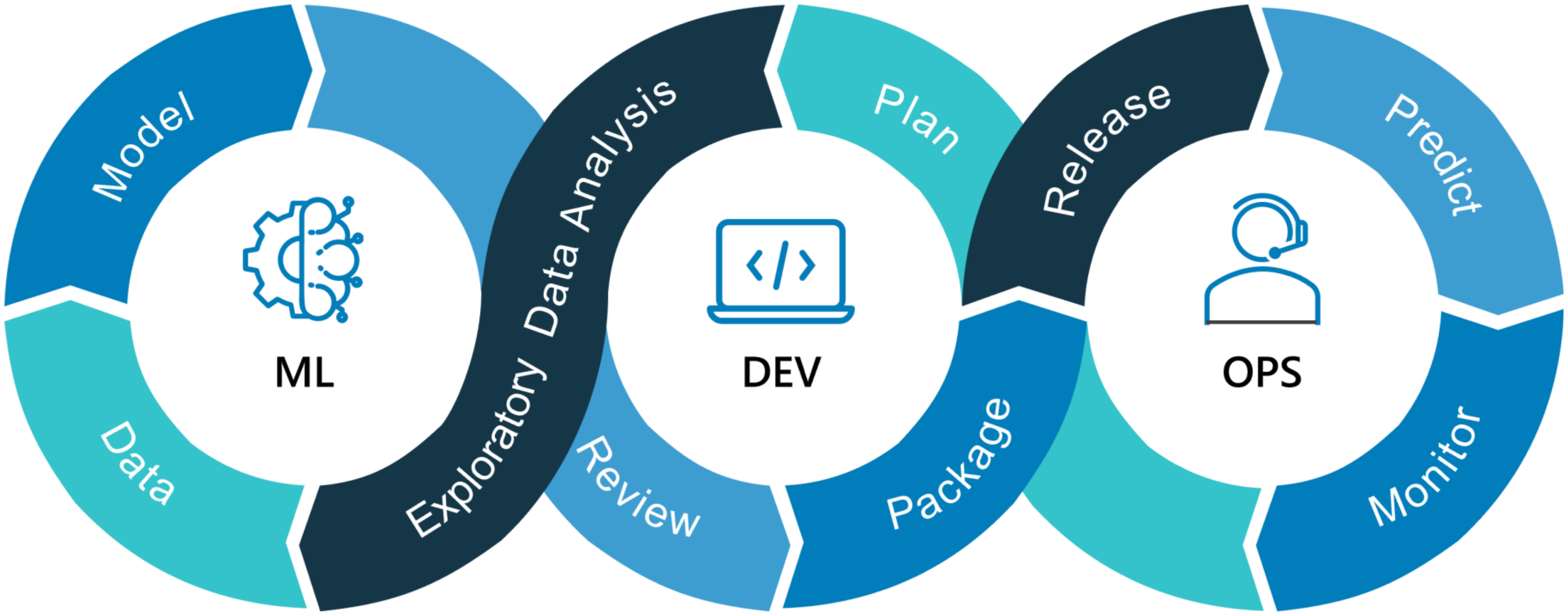
In addition, MLOps focuses on building a common set of practices which data scientists, ML engineers, app developers, and IT Operations can follow for systematically managing analytics initiatives.

These benefits put organizations with MLOps initiatives ahead of the competition as their counterparts continue to struggle packaging, deploying, and maintaining stable model versions.

MLOps can help mitigate these common challenges while providing additional value to the organization with improved ML model quality and better performance.

Organizations with MLOps initiatives reap several benefits

- Improved confidence in their models
- Improved compliance with regulatory guidance
- Faster response times to changing environmental conditions
- Lower break-fix cost
- Increased trust and ability to drive valuable insights

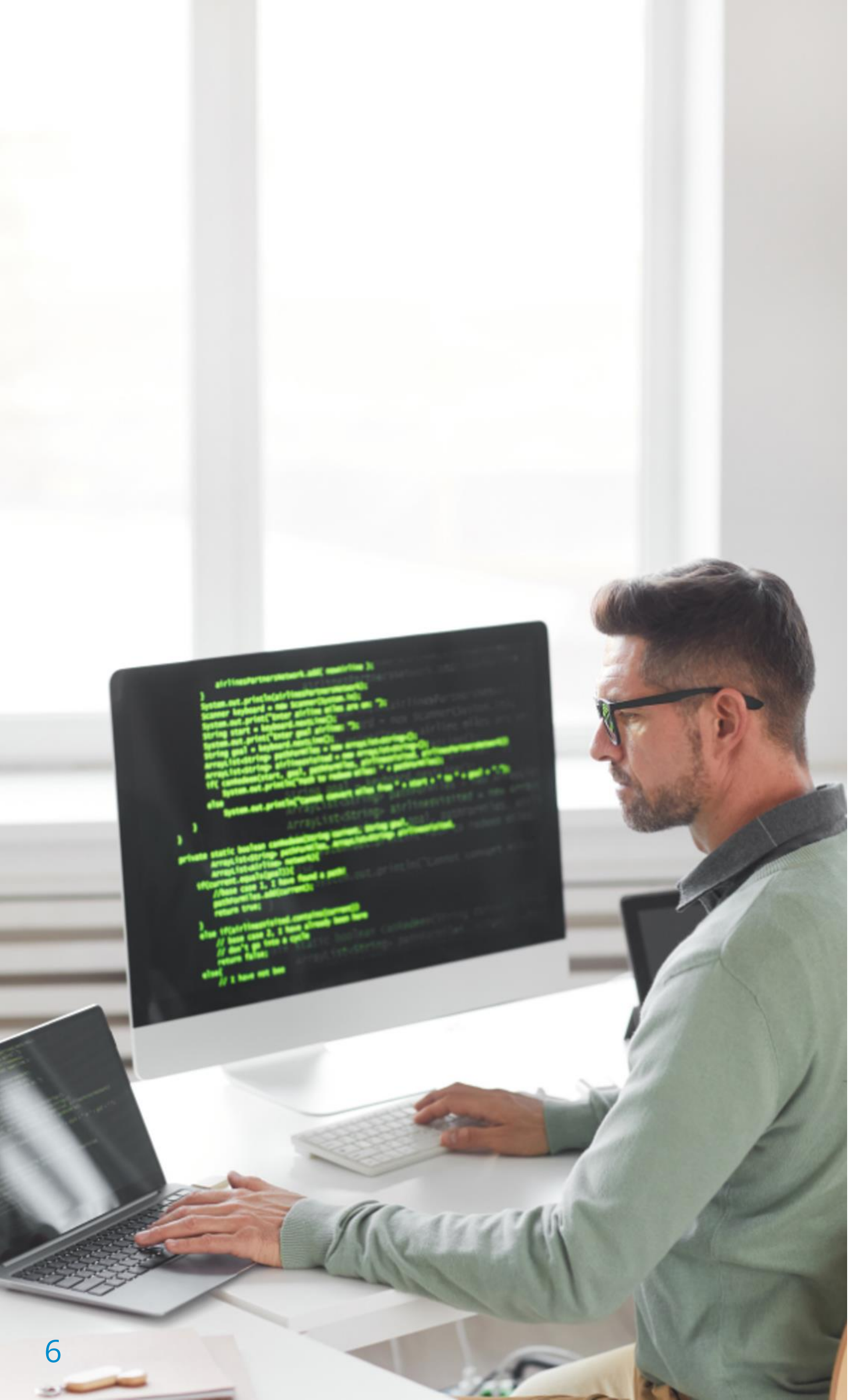


How does MLOps work?

DevOps solves the problems associated with developers handing off projects to IT operations for implementation and maintenance. MLOps introduces a similar set of benefits for data scientists.

With MLOps, data scientists, ML engineers, and app developers can focus on collaboratively working towards delivering value to their customers.

> *MLOps is positioned to solve many of the same issues that DevOps solves for software engineering.*



MLOps in action

Traditionally speaking, packaging and deploying machine learning solutions have been manual and error-prone processes.

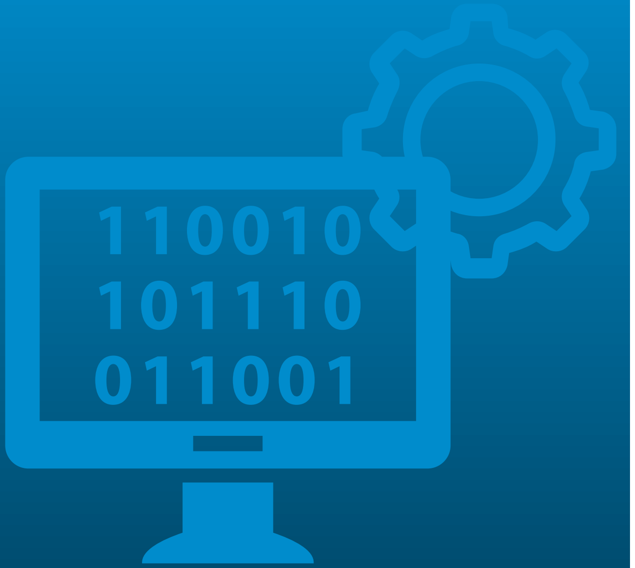
One likely scenario is that data scientists build models in their preferred environment and later hand off their completed model to a software engineer for implementation in another language like Java.

This is incredibly error prone, as the software engineer may not understand the nuances of the modeling approach, or the underlying packages used. Additionally, it requires a significant amount of work each time the underlying modeling framework needs to be updated.

A much better approach is to use automated tools and processes to **implement CI/CD for machine learning.**

> "Reproducible ML helps reduce the costs of packaging and maintaining model versions"

> MLOps supports ML models throughout their lifecycle by implementing a common set of practices



This is where MLOps comes in. The modeling code, dependencies, and any other runtime requirements can be packaged to implement reproducible ML.

Reproducible ML will help reduce the costs of packaging and maintaining model versions. This will, in turn, give you the power to answer the question about the state of any model in its history.

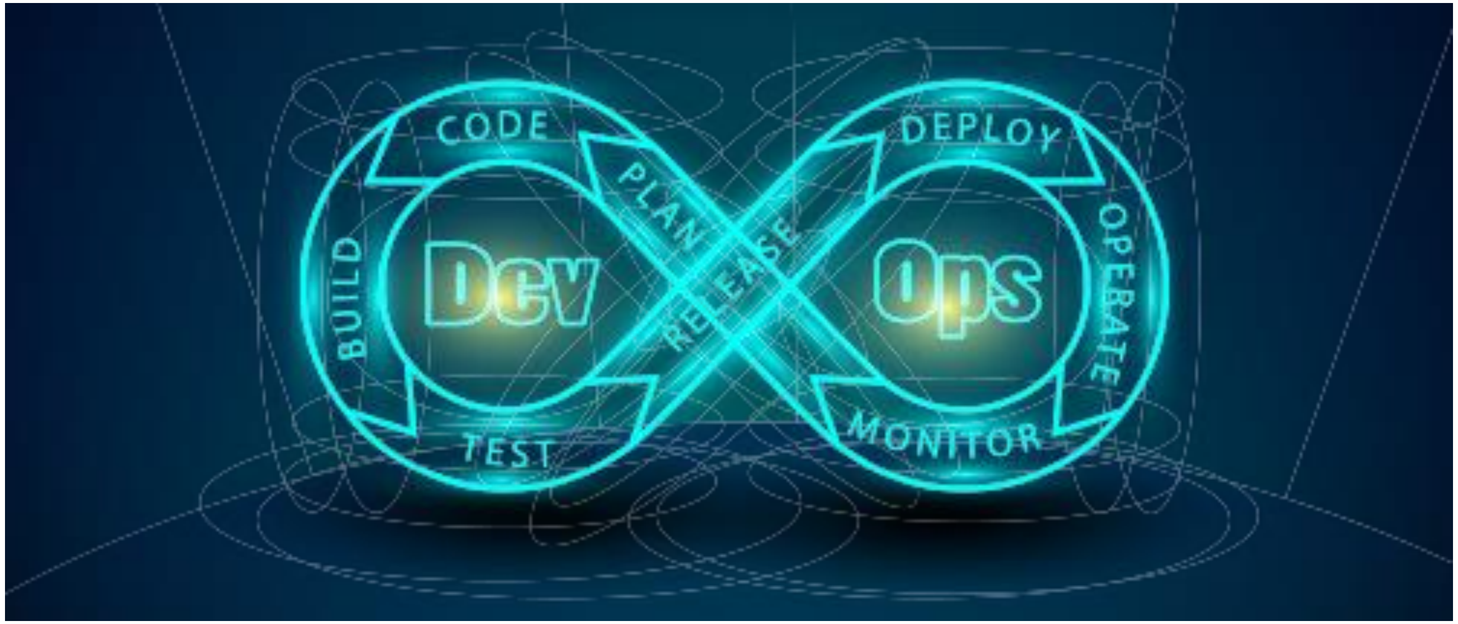
Additionally, since it has been packaged, it will be much easier to deploy at scale. This provides reproducibility and is one of several key steps in the MLOps journey.

MLOps aims to support machine learning models throughout their lifecycle by implementing a common set of practices depicted below.

MLOps broad range of tasks

Though there are many similarities between DevOps projects and ML projects, it’s important not to take DevOps practices and techniques and apply them blindly to machine learning projects.

DevOps vs. MLOps



The IT team does not have deep expertise on modeling algorithms, and data scientists do not want to manage infrastructure. Therefore, to implement MLOps it’s important to bridge the gap between IT teams and ML engineers.

The ML Engineer role brings a specialized skillset with the mandate of collaborating with IT and the business function leveraging their models to ensure those models are well supported throughout their lifecycle.

Aside from skillset, there are also key differences in the activities required when implementing DevOps vs. implementing MLOps.

DevOps vs MLOps comparison table

	DevOps	MLOps
Team skills	Software development skills such as version control and object-oriented programming while working with delivery methodologies, such as Agile	Skills from roles such as data scientists and data engineers are also needed, which include modelling, data-wrangling, and experimentation
Development	Known target to develop towards with a mostly linear development curve to completion	Development experimental in nature where differing modelling techniques, features, and algorithms are used. Progress can be deemed slow until there is a breakthrough.
Testing	Typical tests include unit and integration tests	Also need, at the very least, data validation, trained model quality validation, and model validation tests
Deployment	Process typically starts with a build and then releases the software to staged environments (CI/CD)	In addition to the build and release, the deployment workflow must account for the continuous training of models from new data, based on sometimes complex conditions (such as data/concept drift)
Production	Solutions are monitored in production for bugs or degradation of service due to changing conditions such as an increase in traffic	Models have ties to the data from which it was created. As future data changes, model predictions may evolve. In addition to the typical metrics to monitor, you must monitor data profiles and raise alerts if these drift to unacceptable levels

Getting started with MLOps

"How do I get started with MLOps?"

This is perhaps the most common question!

One thing must be reinforced: **MLOps is not a product that you can buy**. It's a **combination of tools, processes, ways of working, and mindset**.

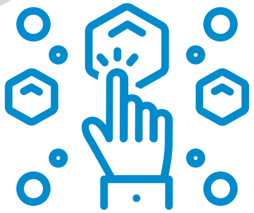
Implementing MLOps is **just as much about change management** and making sure that the right mix of personnel are involved throughout the ML lifecycle.

Additionally, the level of effort to implement MLOps practices can vary significantly depending on the maturity level of the organization.

Fractal can help you assess this maturity and get you started with your journey.

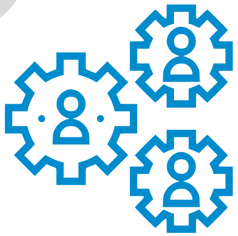
Three tips to get you started with your MLOps journey

1



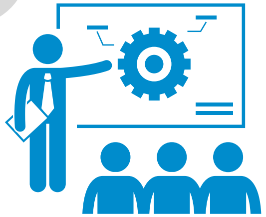
Choose up to 3 focus areas to kickstart your MLOps implementation

2



Leverage the right mix of skillsets

3



Provide adequate resources

1

Choose up to 3 focus areas to kickstart your MLOps implementation

If your organization is just beginning its MLOps journey, remember one thing: **don't try to do everything at once**.

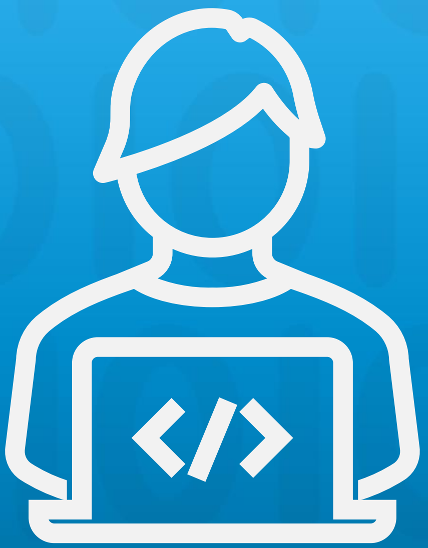
Trying to implement too many changes at once often results in an unorganized rollout with little evidence of progress at the end of the day.

When getting started, consider targeting a select few focus areas for implementing MLOps concepts. **For example, you may want to focus on model reproducibility**.

Achieving true reproducibility requires the thoughtful implementation of source control management processes, environment tracking, model portability, and model registries.

Rather than try to achieve all of these at once, pick one.

For example, a **good place to start would be to implement source control processes for the data science team**.



One common observation is that **data scientists tend to keep their work on their computer.** They may use their own file naming conventions to keep track of their code versions, but those are still stored locally and in an unstandardized way.

A better practice would require that they version their code and artifacts using a source code management system like GIT.

When implementing MLOps into your analytics organization, data scientists need to be **encouraged to become more like their software engineering counterparts**.

This change will need to be planned, tracked, and managed to ensure new processes are adequately supported.

A plan which provides training materials and review sessions will help ensure that changes are being adhered to.

2

Leverage the right mix of skillsets

Many organizations are eager to implement MLOps but do not stop to think if there is a role or skill gap in their existing organization. **One common misconception is that data scientists should be able to manage the models** throughout the entire model lifecycle.

This is not the way things should be approached, as it requires a highly specialized set of skills that is difficult to identify and hire.

Instead, **data scientists should be left to do the work they do best** (i.e., building accurate machine learning models) **while leaving the production machinery to another role: ML Engineers**.

ML Engineers should own the responsibility of taking the models built by a data scientist and building the machinery around them to be production ready. These ML Engineers will handle the packaging, deployment, and monitoring mechanisms for getting the model into a production state and ensuring it continues to perform appropriately.

Working together with data scientists, agreements need to be made about:

- How often to retrain a model,
- Defining adequate model performance metrics,
- When to require further investigations as conditions evolve over time.

3

Provide adequate resources

As new processes and tools are rolled out to the business, remember to make adequate training and resources available.

Whether incorporating a new tool to handle part of the MLOps model lifecycle, or introducing a new business process, you must give the data scientist and ML engineers the appropriate resources.

We recommend **an appointed team of practitioners** who can help answer any questions about new processes and how to best adhere to them.

On the technology side, many technology vendors can provide **training** on the various platforms and Fractal can provide advisory services on how to use the appropriate technology to achieve the desired outcomes.

Scaling MLOps

AI and machine learning projects are on the rise. According to Gartner, **48% of CIOs and tech executives deployed an AI/ML project in 2022**.

It's also estimated that 50% of IT leaders will struggle to drive their AI initiatives from Proof of Concept (PoC) to production through 2024.

Challenges moving AI/ML initiatives from PoC to production

What causes the gap between PoC and AI/ML model implementation?

IT and business leaders often cite challenges relating to security, privacy, integration, and data complexity as the key barriers to deploying AI/ML models in production.

It is often due to governance frameworks not being shared across an organization to ensure compliance and maintainability – if a framework exists at all.

According to an O'Reilly report, this becomes "*even more dangerous when you're relying on AI applications in production.* ***Without formalizing AI governance, you are less likely to know when models are becoming stale, when results are biased, or when data has been collected improperly***."

AI models require constant attention in production to achieve scalability, maintainability, and governance. To do that, organizations need a strong MLOps foundation.

> "At some point, your proof-of-concept is likely to turn into an actual product, and then your governance efforts will be playing catch-up"
>
> Mike Loukides, O'Reilly report

Leveraging MLOps at scale

Organizations following an MLOps methodology also gain a clear advantage in time to deployment.

McKinsey found that **companies without a formalized MLOps process often took 9 months to implement a model.**

In comparison, companies **applying MLOps could deploy models** in 2 to 12 weeks, or **3 to 20 times faster!**

The secret? By applying MLOps practices, these companies were able to create a "factory" approach for repeatable and scalable AI/ML model implementation. Their engineers weren't building everything from scratch. They could pull from a library of reusable components, automate processes, and ensure compliance and governance throughout the organization.

Luckily, you can also take this approach with Fractal's **AI Factory Framework.**

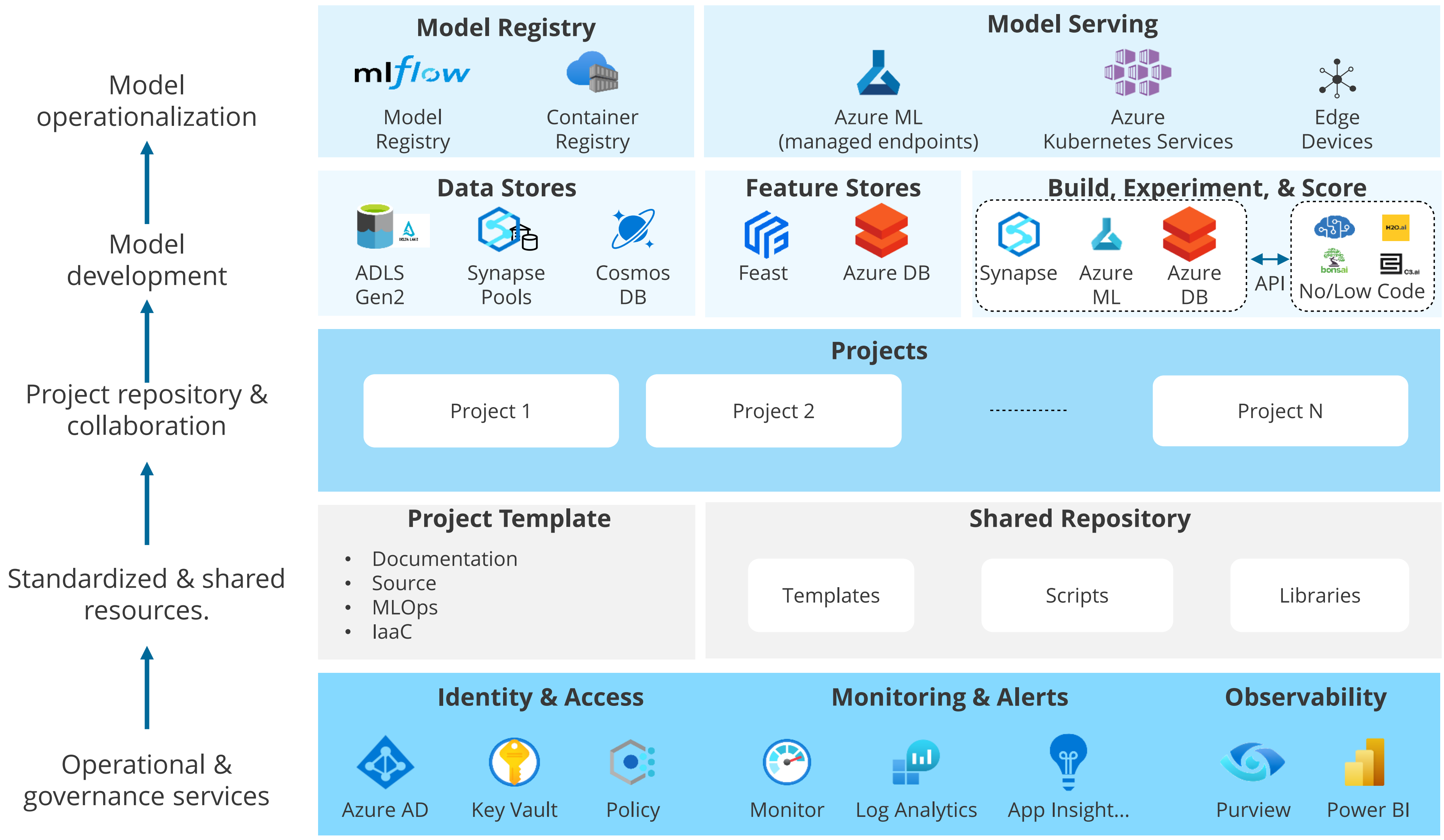
In one survey, Deloitte found that organizations that strongly followed an MLOps methodology were…

3x
More likely to achieve their goals

4x
More likely to feel prepared for AI related risks

3x
More confident in their ability to ethically deploy AI initiatives

AI Factory framework technical overview



Fractal AI Factory Framework is a cloud-based MLOps framework that provides organizations with the foundation to deliver Data Science, Machine Learning, and AI projects at scale.

It offers enterprise-level reusability, security, integration, and governance.

Simply put, **AI Factory helps customers scale MLOps, centralize governance, and accelerate time to deployment.**

Key benefits of the AI Factory

By leveraging reusable and standardized artifacts, automated pipelines, and governance solutions, our AI Factory framework reduces duplicate effort and upskilling needs between teams and projects.

AI Factory framework: Key outcomes and deliverables

Automated pipelines

- Environmental setup & configuration
- Data (batch & streaming) acquisition
- Feature engineering
- Model training, deployment, & retraining
- Batch inferencing

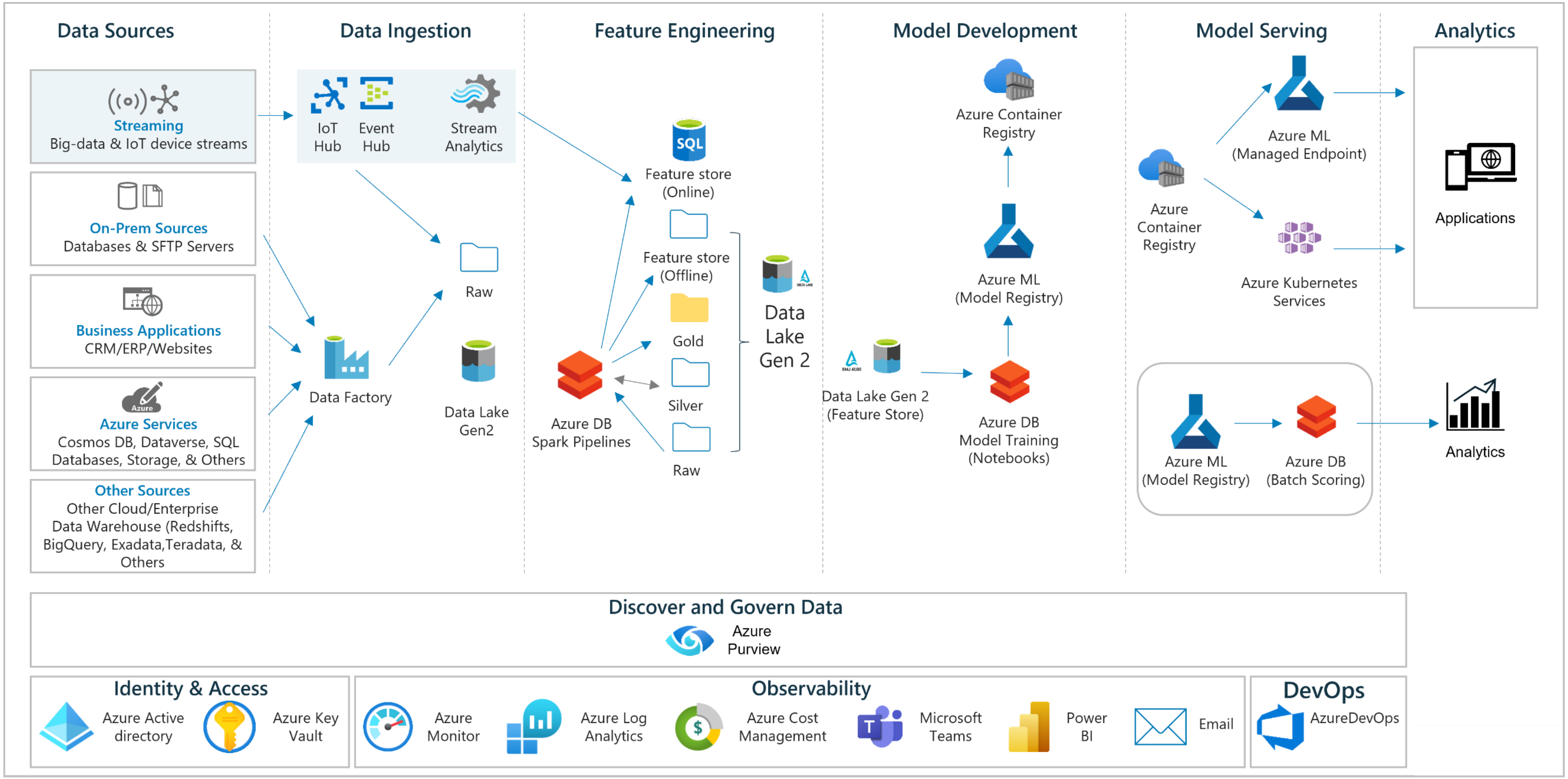
Governance solutions

- Project setup/approval workflow
- Project governance & insights
- Model governance & insights
- Vulnerability scanning & fixes
- Security risk intelligence

Reusable components

- Reference architectures
- Project templates
- Infrastructure as code (IaC) templates
- MLOps pipeline templates
- Custom libraries/scripts
- How-to documentations
- Detailed playbooks

AI Factory framework : Reference architecture



AI Factory benefits

Customers leveraging the Fractal AI Factory Framework can take advantage of **Fractal's AI engineering best practices** to accelerate deployment and ensure model governance at scale.

AI Factory also helps businesses

- Make the entire end-to-end lifecycle more repeatable, governable, safer, and faster
- Reduce development risks and improve model quality
- Shorten planning and development with accelerated time to deployment
- Reduce team’s upskilling needs
- Streamline operational, security, and governance processes
- Achieve higher success rates and ROI

What next ?

Over the last decade, Fractal has helped many customers build and execute their AI governance strategy.

We distilled this experience and the derived best practices in this framework to help deliver customers’ AI/ML initiatives at scale.

If you are not sure where to start, feel free to contact us for a **first exploratory discussion with one of our MLOps experts.**

MLOps resources

Fractal

- MLOps solution page
- Optimizing ML Operations on Edge: Real-Time Deep Learning Solutions on Edge Devices

Microsoft

- Drive efficiency and productivity with machine learning operations
- Webinar recording: Accelerate the Lifecycle with MLOps
- Developing MLOPs maturity, GigaOm customer story
- MLOps: Model management, deployment, lineage, and monitoring with Azure Machine Learning,

Databricks:

- What is MLOps by Databricks
- The big book of MLOps v2 (with LLMOps) by Databricks

Other:

- MLOps series: AI Engineering with Opensource tools & AKS
- MLOps: The AI life cycle for IT production by Nvidia

fractal.ai/mlops

One World Trade Center Suite 76J, New York, NY 10007 | +1 (646) 547 1600

info@fractal.ai @fractalai linkedin.com/showcase/fractal-microsoft-partnership

Microsoft Solutions Partner
Infrastructure
Azure

Microsoft Solutions Partner
Data & AI
Azure

Microsoft Solutions Partner
Digital & App Innovation
Azure