Informatica

Whitepaper

MLOps: How to Operationalize Machine Learning Models in 5 Steps

The Framework to Successfully Automate and Productize Machine Learning Algorithms

By Sumeet Agrawal and Anant Mittal

informatica.com

Contents

Data Science, Machine Learning and the Need for MLOps

Today, artificial intelligence (AI) and machine learning (ML) are powering the data-driven advances that are transforming industries around the world. Businesses race to leverage AI and ML to seize a competitive advantage and deliver game-changing innovation. This includes everything from new therapies in life sciences to reduced risk in financial services to personalized customer experiences.

But AI and ML are data-hungry processes. They require new expertise and new capabilities, including data science and a means of operationalizing the work to build AI and ML models. The world of software development has long been familiar with the concept of DevOps and Agile methodology. Now it's time for organizations to adopt the practice of machine learning operations (MLOps), which we will define in a section below, to succeed with their AI and ML initiatives.

The goal of this white paper is to provide a framework for data scientists and data engineers to operationalize data science by leveraging MLOps. Within this framework, we will address the different steps involved in MLOps and how Informatica can help. But first for context, let's briefly explore data science and ML.

What Is Data Science?

Although data science is a relatively new discipline, it's quickly emerged as a significant technology trend, with Harvard Business Review at one point calling a data scientist "The Sexiest Job of the 21st Century."[1]

Unlike traditional analytics, the aim of data science is predictive — answering questions about the future, as opposed to what has happened in the past. It uses AI and ML to produce transformative insights. As such, it needs large amounts of data in real time.

Data scientists use various tools to build and train ML models that can then detect patterns and make predictions from vast amounts of data. Some of their work is exploratory, involving raw data. But they also use ML algorithms to answer specific questions the business may have about future occurrences.

[1] Harvard Business Review, "Data Scientist: The Sexiest Job of the 21st Century," by Thomas H. Davenport and D.J. Patil, October 2012

ML is a subset of artificial intelligence in which an algorithm discovers patterns in data that include batch and streaming data. The ML models learn from past experiences and make predictions based on that experience. Let's take the example of creating personalized offers in retail. To predict a customer's propensity to buy a specific product, data scientists can build ML models using data from the customer's historical buying patterns, combined with data from real-time activities such as web searches, current geographical locations, activities in the mobile app, etc. This helps retailers create data-driven campaigns and increase the marketing campaign's ROI by providing the right offer to the right customer at the right time.

Data science and ML enable organizations to make decisions based on data and insights rather than gut feeling or instinct. Moreover, the ML models become more intelligent and more accurate with exposure to new data. For example, Google Maps can display faster routes when traffic conditions are added to its algorithm.

Data Science and Machine Learning Use Cases

Data science and ML are used in various industries, particularly:

- **Healthcare**: Providers, payers and other healthcare organizations have large amounts of data — such as medical records, diagnostic information and medical claims — that they can use to help reduce patient readmittance and detect fraudulent claims.
- **Banking and financial services**: Fraud detection, credit and loan approvals and blockchain are among the key applications of data science in banking.
- **Marketing**: From just-in-time offers and next-best-action to campaign effectiveness and churn analysis, data science can help companies provide the best customer experiences while increasing their top line.

Similarly, other data science use cases exist in pharma, automotive and transportation, insurance, energy, government, sales and supply chain management.

How Data Science Projects Differ From Traditional Data Warehousing

In the past, enterprises built data marts or data warehouses to support their reporting and analytics needs. Data integration capabilities in the form of on-premises extract, transform, load (ETL) tools were essential.

Today, approaches like cloud data lakes, schema-free storage, in-memory databases and lakehouses are transforming data warehousing radically.

As with traditional analytics, data science projects demand substantial data integration capabilities. But the data integration capabilities and patterns for data science are very different than those for data warehousing. Let's explore key differences below:

Data Warehousing Projects

Building a data warehouse on-premises is a staid, regimented and linear process. Much of it is designed to reduce chaos and provide order for what can sometimes be a confusing project with many stakeholders and participants. But today, the lengthy process of building a data warehouse is leading to its downfall.

Building a robust, enterprise-grade data warehouse can take months. And in many cases, what is finally delivered doesn't meet user expectations or the requirements, which may have changed over the course of the project. Sometimes, a data warehouse is obsolete the day it is delivered.

Data Science and Machine Learning (DSML) Projects

Designing, developing and deploying a data science project today (and its associated predictive model) is nowhere as mature as a data warehousing project. This space is much like the early days of data warehousing, with multiple tools targeting different parts of the problem and the maturity for a large-scale, end-to-end deployment still in the future. But compared to data warehousing, data science has two major differences:

1. DSML is inherently experimental and requires several iterations of model building, evaluation, adding additional data, rebuilding and so on. Since the outcomes are not known ahead of time, a perfect project plan cannot be put in place and data requirements are not finalized at the time the project starts.

2. Two different teams are involved in the development and testing of the end-to-end project. One deals with data provisioning and the final deployment of the model, while the other builds and evaluates the data science model.

As we'll explain below in "The 5 Steps of MLOps," the steps involved in implementing a successful data science project are very different from those in data warehousing.

What Is MLOps and Why Is It Necessary for Data Science and Artificial Intelligence?

Nine out of ten leading businesses have investments in AI technologies, but less than 15% deploy AI capabilities in their work.[2]

Many AI/ML projects fail because they lack a framework and architecture to support model building, deployment and monitoring. We call such a framework MLOps, which is a new practice for collaboration and communication between data scientists and IT professionals for automating and productizing machine learning algorithms. MLOps enables the operationalization of the end-to-end pipeline that supports the continuous delivery and continuous integration of ML models in a production environment.

Most organizations engaged in data science have defined a process to build, train and test ML models. The challenge has been what to do once the model is built. Integration, deployment and monitoring are essential aspects for providing continuous feedback once the models are in production. This is where the entire process of building ML models aligns more closely with the software development lifecycle than with an analytics project.

Many organizations think data science and machine learning projects are limited to creating models. But once a model is developed and deployed, many other aspects are needed to operationalize it:

- Management and monitoring to ensure the model performs optimally within the thresholds defined by the business
- A feedback loop to monitor model drift or degradation
- Tweaking of the model and periodic retraining

Hence, you need MLOps to establish a continuous delivery cycle of models that form the basis for AI-based systems. MLOps is critical for delivering business value for data science projects.

[2] https://dataprot.net/statistics/ai-statistics/#:~:text=Nine%20out%20of%20ten%20leading,million%20new%20ones%20by%202025.

The 5 Steps of MLOps

There are five phases of an MLOps flow that are necessary for successful data science projects. This flow is inspired by a few project frameworks for data science, notably cross-industry standard process for data mining (CRISP-DM). Below, we step through the key stages that we've seen consistently emerge across many organizations' data science lifecycles.

The five steps are:

1. Business understanding
2. Data acquisition
3. Model development
4. Model deployment
5. Model monitoring

The MLOps Flow

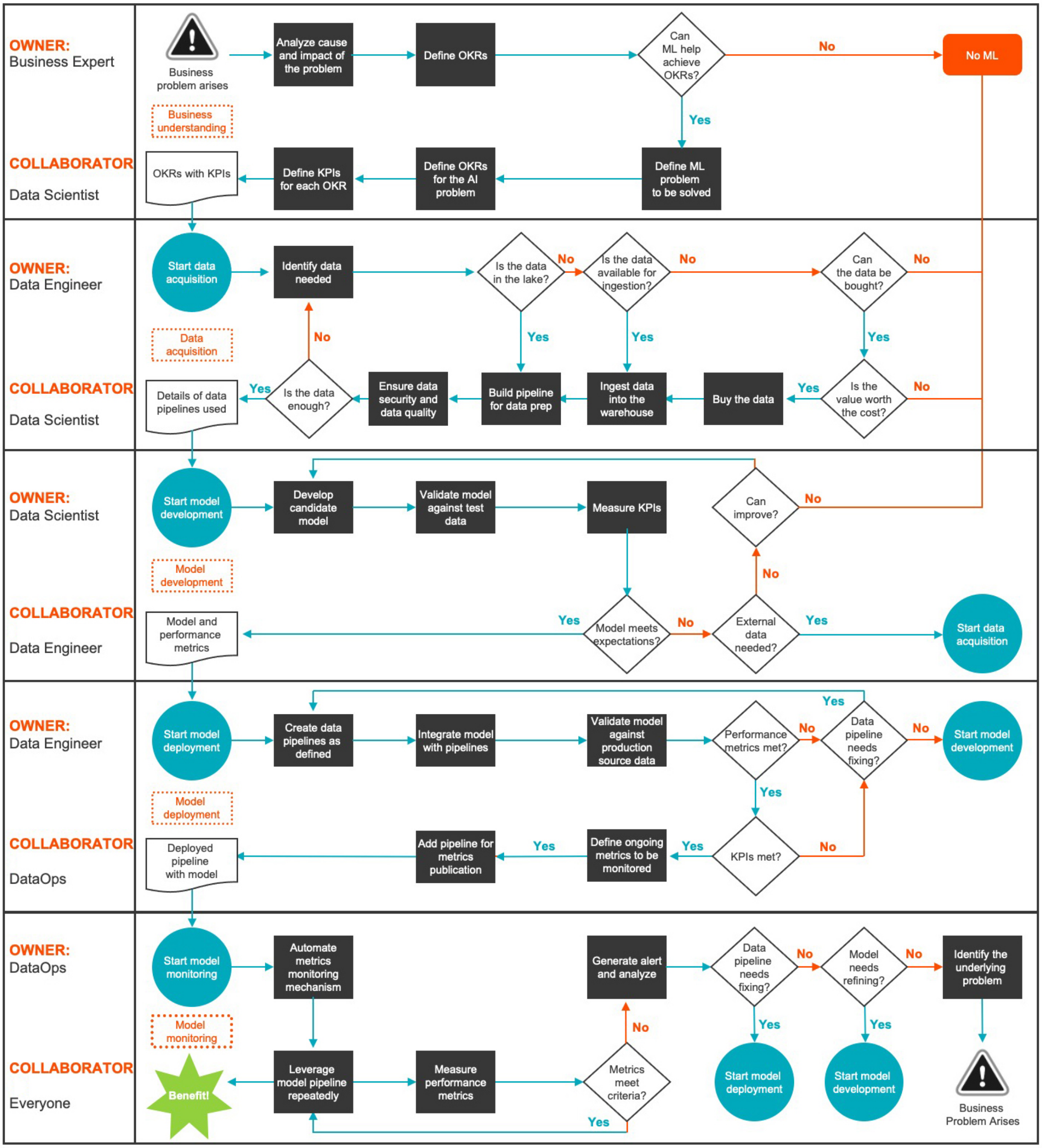

Figure 1. The five-step MLOps flow.

How the MLOps Steps Apply to a Retail Use Case

As we discuss each of the steps in detail below, we'll look at their application to a retail store that is leveraging ML to drive growth. The retail store has offers that are not performing to expectations. But the store has important untapped resources such as customer data, customer purchase histories and other metadata. By applying ML to this wealth of information, the store can potentially benefit. As we discuss the different phases of MLOps, we will see how this retail store could apply each of the phases to this problem and ultimately achieve the desired goal.

MLOps Step 1: Business Understanding

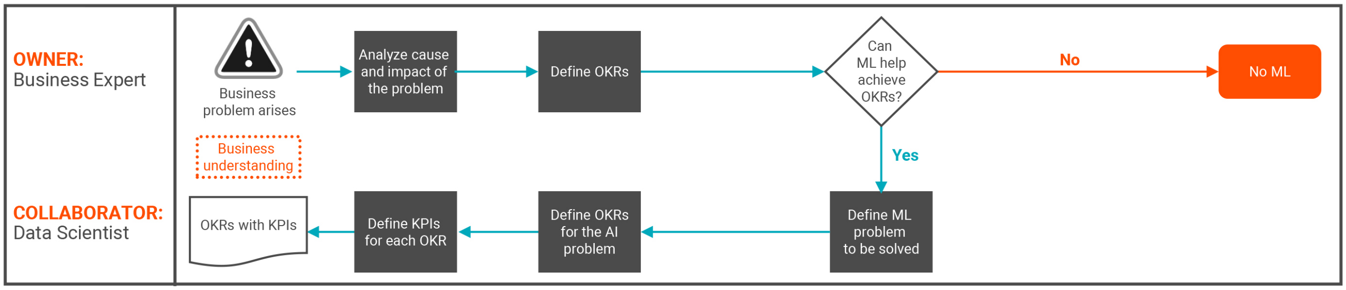

Figure 2. Step 1 of the MLOps flow: Business understanding.

Business understanding is the first and the most defining step in the process. Note that the flow starts with "business problem arises." This means that we should not be looking at the ML development process unless we have a problem. Taking a problem-first approach helps you clearly define the potential problems to be solved and the value gained from the project, setting you up for success.

The initial steps of the business understanding phase are similar to gathering requirements for any other data project. As with a traditional use case, the primary goal should be for business experts to analyze the problem and clearly define objectives and key results (OKRs). Only after OKRs are defined should a data scientist be involved in the process.

Keep in mind that not every problem is suitable for ML. In this phase, the data scientist should discuss OKRs with the business experts to determine if ML can indeed help. If so, they can define the ML problem that needs to be solved. The context gathered in the business understanding phase will help the data scientist identify business-side SMEs, as well as their relevant domains of expertise (process, system, semantics, policy and so on).

For each of the OKRs, one or more key performance indicators (KPIs) need to be defined. These KPIs and OKRs must be documented for future reference and will be critically useful in ensuring that the project delivers the expected value.

The definition and documentation of the business problems, as well as OKRs and KPIs that quantitatively address those problems, provide key context for subsequent phases. This helps to distinguish relevant data, defining how that data maps into the model (both during development and deployment phases) and identifying which dimensions of model performance should be monitored once the model is in production and against what criteria.

Business Understanding for the Retail Use Case

Business experts of a national retail store came up with a way in which special offers are created at the level of each local store every two weeks. These offers consider local goods' surplus as well as consumption patterns. The problem they have is that consumers are not aware of the offers at their local stores. As a result, sales of the special-offer product are lower than expected.

Based on this business problem, a simple objective would be to send consumers a phone message making them aware of the local store offers. This is a problem that, when discussed with a domain-expert data scientist, can be converted to its corresponding ML problem. A potential ML problem to solve here would be to identify any offers in the store that are likely to be of interest to a given consumer who is local to the store.

The other problem of identifying consumers local to a given store need not be an ML problem, as this would be part of the existing data. (In that case, the identified KPI could then be a metric measuring the effectiveness of the existing data.) A way to measure it could be the percentage of consumers who visited the store in two weeks after getting the message with identified offers.

How Informatica Helps Business Understanding

Business understanding is a non-technical phase. Communication is the key tool here, so that different stakeholders can identify the business problem(s) and document the OKRs and KPIs effectively.

During this phase, it's essential to map the processes, systems, key data elements and policies documentation for the key domains expressed in the business problem. This information is often created and maintained by the data governance team with an enterprise data governance tool like Informatica's data governance and catalog capabilities.

Our data governance and catalog capabilities facilitate collaboration between data governance and data stewardship practitioners and SMEs with line of sight into the relevant systems, data, processes and owners. It provides an entry point for exploring the context in which the business problem is first identified and to disseminate the MLOps team's understanding of the problem space, as well as their approach to addressing the project OKRs and KPIs.

One advantage to this approach is that the subsequent artifacts developed by the project (i.e., the model, deployment and production environments, the data pipelines feeding those environments and the measurements used to assess and monitor the performance of the model over time), can themselves be subject to governance control. This ensures the ongoing relevance, qualification and effective controls over the data being used by the project.

MLOps Step 2: Data Acquisition

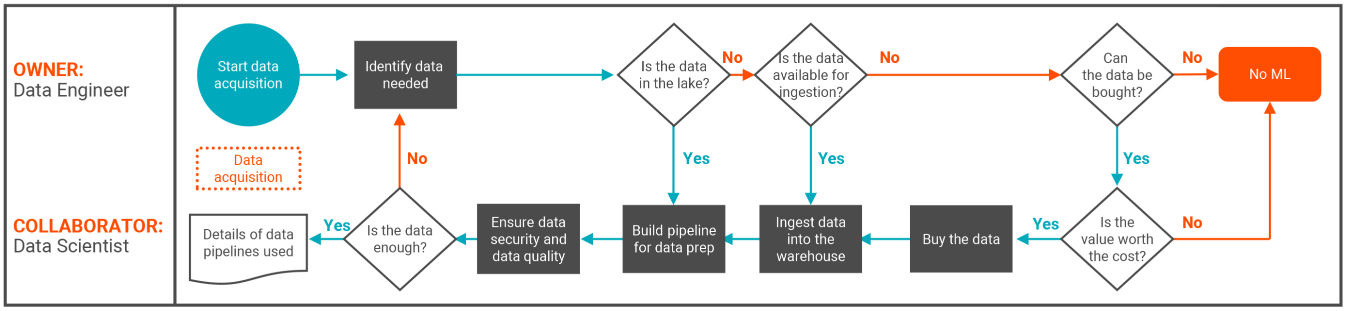

Figure 3: Step 2 of the MLOps flow: Data acquisition.

During the data acquisition phase, data is gathered for the solution. ML development is a highly iterative process, so it's not necessary to gather all data upfront. The goal here is simply to gather enough trusted data to take the first steps toward a solution.

For data acquisition, the data scientist first identifies the necessary information aspects. These aspects should be discussed with a field-expert data engineer to identify potential data sources. Enterprises targeting ML development likely already have a data lake or are in the process of building one. A data catalog of the data lake is the first place to look. Other options include ingestion from a new data source or purchasing data for ingestion into the lake.

Once data is identified, the data engineer builds the pipeline that makes sufficient data available for the data scientist. The data engineer performs preliminary cleansing steps and validates that there's sufficient volume of high-quality data to meet the data scientist's requirements.

These data requirements can be written to the data scientist's workspace. The data pipelines established to take the data from the lake, cleanse it, and take it to the workspace should be clearly defined and stored for future use to ensure the solution's future reproducibility and productization.

Semantic content for search and data acquisition includes quality and policy.

Data Acquisition for the Retail Use Case

To build a model identifying relevant offers for a given customer, the data scientist needs a large variety of data sources. Useful information could include past purchase histories, past offers and details about which customers engaged with previous offers. The data scientist and the domain-expert data engineer should discuss these and other potential sources of information.

Any useful data sources will be brought in by the data engineer from the data lake using appropriate pipelines. When bringing in data sources, the data engineer also must consider data security and data quality. Once the raw data is collected, the data engineer performs preliminary data preparation steps to ensure that enough high-quality data is available. For example, if there is not enough data on customer purchase history, then the data engineer needs to look for more data sources.

How Informatica Helps Data Acquisition

Data acquisition involves getting access to large amounts of data that are distributed. A data lake, along with a data catalog for search, are essential for efficient data acquisition. Moreover, data management tools greatly facilitate the creation and handling of complex data pipelines as compared to simply hand coding them. Data security and data quality are also important aspects and should ideally be done using appropriate tools. With products that enable data management for data science, Informatica can play a vital role in the data acquisition process.

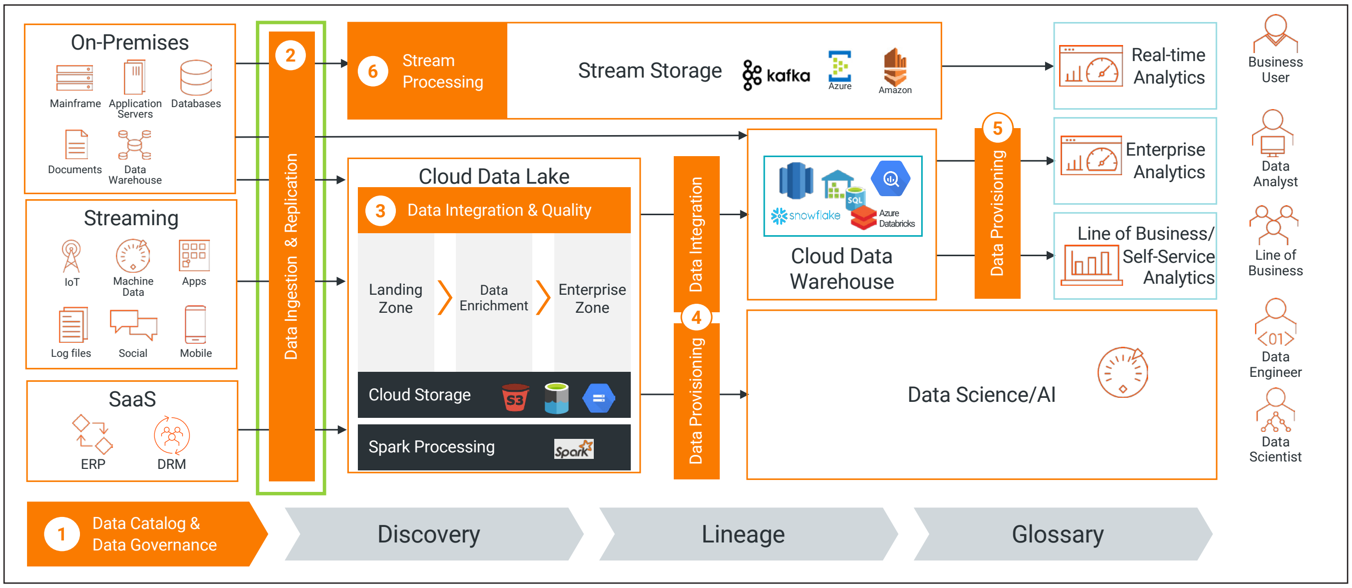

Figure 4: Reference architecture for data acquisition.

Successful data acquisition requires capabilities across four solution areas:

1. **Catalog**
   An intelligent, enterprise-class data catalog enables business and IT users to unleash the power of their enterprise data assets by providing a unified metadata view that includes technical metadata, business context, user annotations, relationships, data quality and usage. It helps users discover the right datasets for modeling.

   As with the business understanding phase, Informatica's data governance and catalog capabilities enable users to easily see definitional information (such as glossary terms) as well as key stakeholders directly in the catalog. What's more, by consulting data governance, users can easily see the business context and processes data is used in, providing them with a holistic view on usage, quality levels and applicable policies.

2. **Ingest**
   Data acquisition requires efficient ingestion of data into on-premises systems, cloud repositories and messaging hubs like Apache Kafka so it's quickly available for real-time processing. In addition, your solution should provide support for streaming IoT and log data, large file sizes and change data capture for databases.

Informatica offers four cloud-based data ingestion and replication services — databases, files, streaming and applications— to meet your specific data ingestion needs. Each managed and secure service includes an authoring wizard tool to help you easily create data ingestion pipelines and real-time monitoring with a comprehensive dashboard.

3. **Process**
Data engineers can help data scientists and data analysts by:

- Finding the right data and making it available in their environment
- Ensuring the data is trusted and sensitive data is masked
- Operationalizing data pipelines and helping everyone spend less time preparing data

The Informatica Intelligent Data Management Cloud™ (IDMC) provides comprehensive end-to-end cloud-native data engineering capabilities. This enables data engineers to process and prepare big workloads to fuel AI/ML and analytics: robust data integration, data quality, streaming and masking capabilities.

4. **Deliver**
Data scientists and data analysts need to rapidly discover, enrich, cleanse and govern data pipelines for faster insights. Informatica's data preparation is an AI-powered solution that simplifies self-service data preparation across cloud data lakes.

MLOps Step 3: Model Development

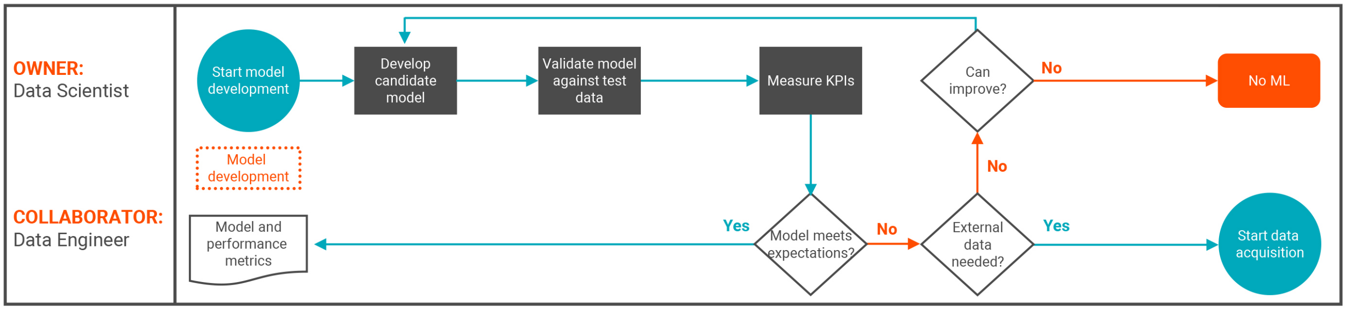

Figure 5: Step 3 of the MLOps flow: Model development.

Model development is the core of the MLOps flow. Up until now, the data scientist has been in an advisory and approver role. Now that the problem and KPIs are clearly defined and high-quality datasets are readily available, the data scientist can leverage their expertise.

During model development, the data scientist iterates through multiple candidate models, validating them against test data and measuring KPIs until expectations are met. If more data is needed, the data scientist can again coordinate with the data engineer on data acquisition and perform additional cleansing and standardization operations. As a result of this phase, the data scientist will be able to identify a model and provide performance metrics that can be used as benchmarks. These metrics may be quite different from the KPIs. Instead, they may be more like what would be published by a data scientist working on a standalone project.

Model Development for the Retail Use Case

To develop a model for the retail store, the data scientist may split the data into training and testing parts. There are multiple ways in which the problem can be solved. For example, a simple way might be to classify the customers into multiple categories and map the offers accordingly. Some offers might be specific to some categories while others might be more general. The extent of the offer — such as percentage discount — could also be a factor in determining the relevance of the offer.

With the model developed, the data scientist can also identify additional metrics. These metrics could, for example, indicate that an appropriate class for a given customer was found with high confidence.

How Informatica Helps Model Development

The tools needed for model development are primarily those that are already familiar to the data scientist, such as notebooks for model development using Python or RStudio for groups using R. Teams use a version control system such as Git for keeping the model code, code sharing, versioning and so on.

Informatica ModelServe is our cloud-native MLOps service in IDMC that caters to data science and ML use cases. It empowers data scientists and ML engineers to focus on solving their business problems rather than worrying about model provisioning infrastructure. ModelServe helps them operationalize high-quality and governed AI/ML models built using just about any tool, framework or data science platform, at scale, within minutes instead of days and months. Models can also be consumed by practically any application.

MLOps Step 4: Model Deployment

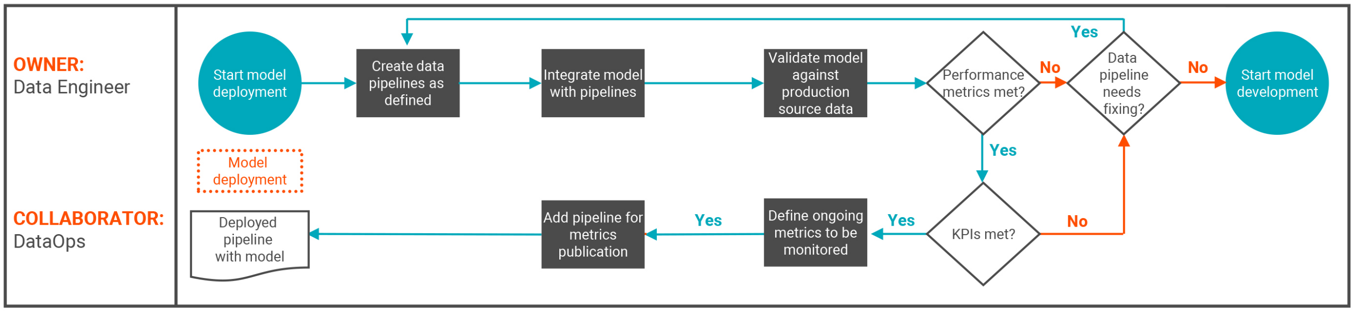

Figure 6: Step 4 of the MLOps flow: Model deployment.

In model deployment, the developed model is made ready for use in a production environment. The data engineer drives this phase, using the pipelines defined during the data acquisition phase as a starting point. The data engineer integrates the model developed by the data scientist and validates it against actual production data. Metrics and KPIs from previous phases are also validated. Invalidation means returning first to the pipeline and then to the model development process to determine the source of the error.

Once a validated pipeline is identified, a new pipeline that measures metrics for future monitoring has to be established. This will allow continuous validation of the metrics identified to ensure that the model remains correct with time. Changing upstream data models and data distributions make this a critical step for any models that are expected to be used over time. This final pipeline is deployed in production with the help of the DataOps team for continuous use and monitoring.

Model Deployment for the Retail Use Case

To deploy the model, the data engineer starts with the same pipelines and quality rules (cleansing, standardization and so on) that were developed during data acquisition. The data engineer deploys new pipelines when new offers are added and when customer information is acquired to generate new offers. Once these pipelines are established, the data engineer executes the pipeline against production data to ensure correctness.

Also needed is an additional pipeline to calculate metrics and KPIs. To calculate if customers visited the store within two weeks after receiving an offer, you need a pipeline to get customer store visit data. This is correlated with the results of the predictions to measure the KPI and publish it. This whole set of pipelines will then be deployed for iterative use in coordination with the DataOps team.

How Informatica Helps Model Deployment

Informatica ModelServe enables data scientists and ML engineers to seamlessly deploy the models using the following steps:

1. **Model Registry**: Data scientists can build their models using AI/ML frameworks like Python, TensorFlow, Spark ML, Keras, etc., and register the models seamlessly.

2. **Model Deployment**: Once the model is registered, data scientists can deploy the AI/ML models in a serverless environment in minutes (instead of days or weeks) without worrying about model provisioning infrastructure.

3. **Model Monitoring**: With Informatica ModelServe, you can monitor the performance of the deployed model in a single pane of glass, detect anomalies and take remedial action like retraining the model. Post monitoring, the model can be consumed in just about any application.

MLOps Step 5: Model Monitoring

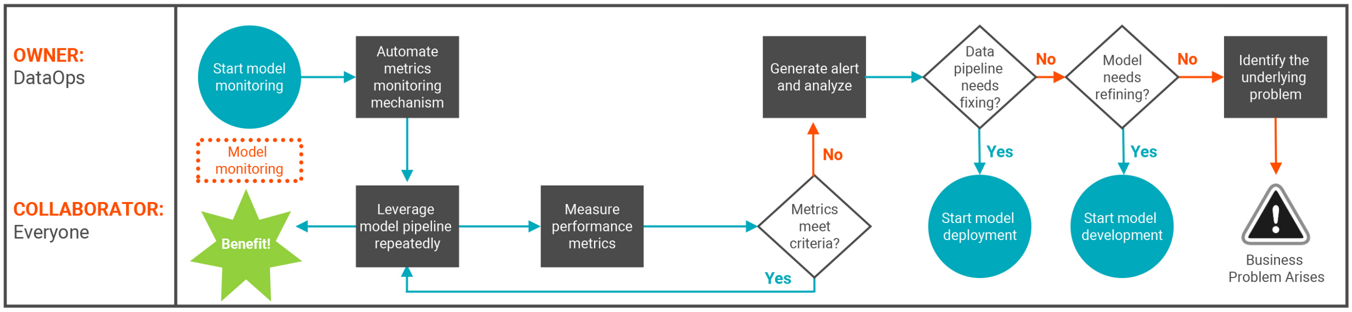

Figure 7: Step 5 of the MLOps flow: Model monitoring.

During the model monitoring phase, a deployed pipeline is integrated with a metrics monitoring mechanism. The DataOps team can then monitor the pipeline metrics, ensuring continued value and increased confidence in ML.

Alerts are generated when performance metrics aren't met. Often, a change in the data flow is the cause and a minor data pipeline change is all that is needed. In some cases, there is a general change in the underlying pattern of data and the model needs to be redeveloped. In extreme cases, a significant change in the business landscape requires going back to the business understanding phase to address a new business problem.

Continuous data profiling and quality scorecard evaluations help identify changes in data over time that require updated model training and evaluation.

Model Monitoring for the Retail Use Case

Now that the pipeline is in production, the store can iteratively send out messages with targeted offers to customers. The DataOps team monitors pipeline performance to ensure it continues delivering value. Over time, source data may change, prompting an alert. For example, the source generating deals might change schema. A simple pipeline change will take care of the issue.

Changing demographics may mean that the model needs redevelopment after a few years. In a more extreme example, shoppers may shift online — a much bigger problem that would need to be handled by starting at the business understanding phase.

How Informatica Helps Model Monitoring

This phase requires metrics monitoring and reporting. An appropriate alerting mechanism is also needed to generate alerts as per the defined rules. Informatica's data quality capabilities can be used to track profiles and changes to pattern and value frequencies over time as well as track business metrics using scorecards. For more advanced reporting requirements, these results can be integrated with an appropriate BI reporting tool.

Using Informatica's data integration capabilities, customers can establish processes across the software development lifecycle. Using the automated data engineering pipeline and the BI tool of their choice, customers can monitor the health of ML models.

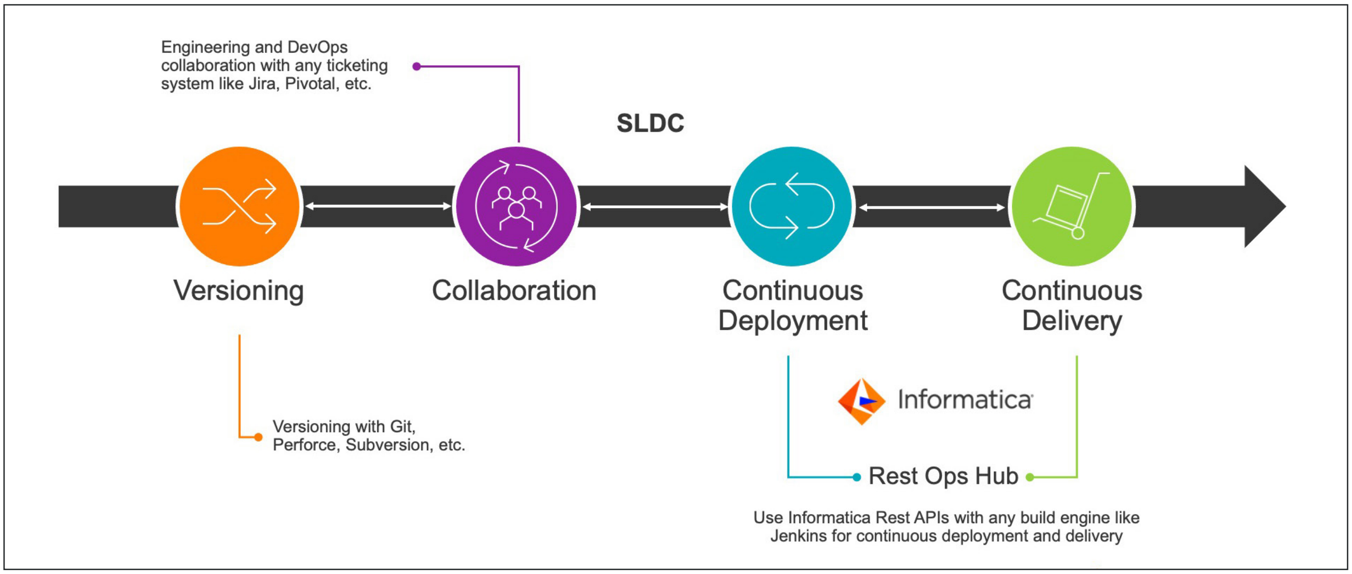

Figure 8: The software development lifecycle.

Conclusion

Data science is growing rapidly, transitioning from niche departmental or individual projects to impactful, enterprise-level initiatives managed or guided by IT. As with data warehousing, data science needs robust data integration and data management capabilities. But these requirements are considerably different than what organizations have implemented for data warehousing.

MLOps capabilities are essential to supporting data science use cases. To execute a successful data science strategy, implementing MLOps consistently is core to the entire initiative. Successful MLOps requires the orchestration of multiple, disparate tools and the skills to integrate them while managing it across a complex environment. A systematic approach is required.

IDMC services provide end-to-end functionality for MLOps. By using IDMC tools, customers can implement successful MLOps for data science projects.

Next Steps

Learn more about building end-to-end AI and ML pipelines at scale with our comprehensive **cloud data engineering capabilities of Informatica IDMC**.

About the Authors

Sumeet Agrawal

Sumeet Kumar Agrawal is a vice president of product management at Informatica. Based in the Bay Area, Sumeet has over 15 years of experience working on different Informatica technologies. He manages and leads the Informatica Data Integration product portfolio, which includes Cloud Mass Ingestion, Cloud Data Integration, Serverless, Elastic, ModelServe, InfaCore and cloud connectivities. He is responsible for defining product direction, roadmap, and key long-term strategy for cloud data warehouse and data lake products. His expertise includes the Hadoop ecosystem, Spark, AI/ML, streaming, IoT and cloud technologies like Amazon Web Services and Microsoft Azure, as well as development-oriented technologies such as Java. He works with cloud and analytics partners like Databricks, Snowflake, Google Cloud, Microsoft Azure and AWS. He is also responsible for evaluating CDW/DL partner technologies for Informatica.

Anant Mittal

Anant Mittal is a principal software engineer, machine learning, at Informatica. Based in the Bay Area, Anant joined Informatica as a computer science graduate from IIT-Delhi. With his deep understanding of on-premises as well as cloud-based data ecosystems, he has been contributing to Informatica Cloud Data Engineering offerings for more than eight years. In addition to his machine learning skills, he has hands-on expertise in Spark, Hadoop and microservices. As a constant learner with an eye for innovation, he is focused on building and productionizing solutions for machine learning at scale.

About Us

At Informatica (NYSE: INFA), we believe data is the soul of business transformation. That's why we help you transform it from simply binary information to extraordinary innovation with our Informatica Intelligent Data Management Cloud™. Powered by AI, it's the only cloud dedicated to managing data of any type, pattern, complexity, or workload across any location—all on a single platform. Whether you're driving next-gen analytics, delivering perfectly timed customer experiences, or ensuring governance and privacy, you can always know your data is accurate, your insights are actionable, and your possibilities are limitless.

Worldwide Headquarters
2100 Seaport Blvd,
Redwood City, CA 94063, USA
Phone: 650.385.5000
Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871

informatica.com
linkedin.com/company/informatica
twitter.com/Informatica

CONTACT US

IN09-0123-3819