Brought to you by

CLOUDERA

Production Machine Learning

for dummies®
A Wiley Brand

Cloudera Special Edition

Prepare to run ML in production at scale

Apply an ML mindset to drive AI success

Apply a people, process, and tech approach

Ulrika Jägare

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises. Learn more at **https://www.cloudera.com.**

Production Machine Learning

for dummies®
A Wiley Brand

Production Machine Learning

Cloudera Special Edition

by Ulrika Jägare

for dummies®
A Wiley Brand

Production Machine Learning For Dummies®, Cloudera Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-73530-4 (pbk); ISBN 978-1-119-73534-2 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Development Editor: Rebecca Senninger

Editorial Manager: Rev Mengle

Business Development Representative: Molly Daugherty

Production Editor: Mohammed Zafar Ali

Contents at a Glance

Table of Contents

Introduction

Machine learning (ML) plays a critical role in optimizing the value of digital transformation. Across industries, organizations seek to leverage the digital revolution for more revenue or lower costs. Machine learning takes the digital transformation journey to another level and makes it possible for teams to work smarter, do things faster, and turn previously impossible tasks into routine.

However, it's not as easy as it may seem to effectively deploy machine learning in the enterprise. To be successful you can't only focus on the technical pieces, but you need to also address organizational aspects as well as enterprise processes and ways of working. This is crucial to break down barriers between ML in development and ML in production. The overall objective should be to quickly and seamlessly be able to move ML models and operate increasing numbers of models on a continuous basis. This is hard, and in this book you find out about the right production machine learning approaches, best practices, and MLOps technology that is critical for creating, sustaining, and scaling your business impact using ML.

About This Book

Production Machine Learning For Dummies, Cloudera Special Edition is about understanding and applying the right ML mindset across the enterprise from the beginning. By having a mindset focused on running ML in production in early ML exploration and development stages, it automatically makes your efforts focused on the right aspects. It helps you remove barriers towards moving ML to production and secure that your ML solutions are ready to run in real, live environments from the start.

This book explains how ML is shaping businesses today, and the concept of *production ML*. You learn what´s needed to succeed with production ML from a people, process, and technology perspective, as well as how to successfully apply a production ML approach at scale in your enterprise.

Icons Used in This Book

I occasionally use special icons to focus attention on important items. Here's what you find:

REMEMBER

This icon reminds you about information that's worth recalling.

TIP

Expect to find something useful or helpful by way of suggestions, advice, or observations here, leveraging experiences from other implementations.

WARNING

Warning icons are meant to get your attention to steer you clear of potholes, money pits, and other hazards. Paying extra attention to these parts in the book helps you avoid unnecessary roadblocks.

TECHNICAL STUFF

This icon may be taken in one of two ways: Techies zero in on the juicy and significant details that follow; others can happily skip ahead to the next paragraph.

Beyond the Book

This book helps you understand more about the importance of production ML as part of your digital transformation journey. However, because this is a relatively short, introductory book to production ML, I also recommend checking out the Cloudera Machine Learning product website at `https://www.cloudera.com/about/machine-learning.html`.

IN THIS CHAPTER

- Describing what machine learning (ML) is all about
- Pinpointing the business impact of ML
- Walking through the ML lifecycle
- Recognizing obstacles towards getting ML to production

Chapter 1
Approaching Production ML

Machine learning (ML) has many practical applications that drive real business results, such as time and money savings, which has the potential to dramatically impact the future of your organization. This chapter focuses on describing why it's important to consider ML production aspects early, how to approach ML lifecycle, and what the main challenges are related to moving ML from idea to production.

Explaining the Basics of Machine Learning (ML)

Machine learning (ML) is computer algorithms that improve automatically through experience based on patterns and deviations in data. It's seen as a subset of *artificial intelligence* (AI). Machine learning algorithms build a mathematical model based on sample data, known as "training data," to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications

where it's difficult or insufficient to develop conventional algorithms to perform the needed tasks.

These basic algorithms for teaching a machine to complete tasks and classify like a human date back several decades. But if machine learning isn't new, why is there so much interest today? Well, the fact is that machine learning algorithms need a lot of data and computing power to produce useful results. Today, we have more data than ever, and computing power is pervasive and cheap. The past few decades have seen massive scalability of data and information, allowing for much more accurate predictions than were ever possible in the long history of machine learning. Machine learning algorithms are therefore now better than ever and widely available in open source software. Some common usage scenarios for machine learning include:

- Predict a future value
- Estimate a probability
- Classify an object
- Group similar objects together
- Detect associations
- Identify outliers

There are many different types of machine learning algorithms, and each class works differently. In general, machine learning algorithms begin with an initial hypothetical model, determine how well this model fits a set of data, and improve the model iteratively. This training process continues until the algorithm can find no additional improvements, or the user stops the process.

So, how does deep learning relate to machine learning? Deep learning is a subfield of machine learning. Neural networks are a type of machine learning that represent knowledge as a set of mathematical functions organized in a directed graph and arranged in layers. Neural networks with multiple "hidden" layers are so-called "deep" neural networks. Deep learning is useful because it performs well on tasks such as image and speech recognition, where other machine learning techniques perform poorly.

Exploring How ML Is Shaping Businesses Today

The reality of today is that data science and machine learning are reshaping entire industries, making it possible to achieve previously impossible levels of scale through operational efficiencies and continuous learning and innovation. The reason for this is because machine learning automates the extraction of useful insight from data, detecting patterns in a way that would take humans weeks, months, or years to complete, if at all.

REMEMBER

You can use machine learning to automate business processes and enhance or invent new products and services. You can predict what a customer is likely to buy, and you can automatically detect manufacturing inefficiencies or fraudulent behavior. And you can do all of this while leveraging the new data and insight your machine learning capabilities generate, allowing for even further optimization and innovation opportunities.

Modern data storage, processing, and software capabilities have progressed far enough to allow any organization to capture and use its wealth of diverse data to train, test, and validate even the most complex predictive machine learning models. Many companies have successfully embedded predictive models in their core business capabilities to develop game-changing products and services that would have otherwise been unachievable. And in doing so, they've proven that machine learning has already changed the business landscape forever.

WARNING

Yet, while the most opportunity comes from adopting machine learning at enterprise scale, only 21 percent of enterprises embed AI across multiple business units. That equates to widespread AI investments across enterprises and industries with very little internal proliferation. One explanation could be that many business leaders are exploring the novelty of AI and don't fully understand the numerous ways in which machine learning and AI can create value across their business. AI can solve problems that were once unsolvable, and it can provide answers to questions enterprises didn't even know to ask. Because of this, achieving success requires experimentation and incremental approaches to

adoption, which might be a very transformational and challenging task for many large enterprises.

When investing in machine learning, organizations put the capabilities to work in numerous ways. For example, a retailer can use machine learning to predict the volume of traffic in a store on a given day and use that prediction to optimize staffing. A bank can use machine learning to infer the current market value of a home (based on its size, characteristics, and neighborhood); in turn, this lowers the cost of appraisals and expedites mortgage processing.

Autonomous vehicles offer an excellent example of applied AI. There are machine learning components built into an autonomous vehicle. But the vehicle also includes sensors that capture and encode data about the world and could be seen as a "brain" that reasons and makes decisions, and devices that instruct the wheels to turn, the engine to accelerate, and so forth.

In fact, ML is being used across various disciplines from healthcare to education and it is showing no sign of slowing down. What is clear from the advantages of using ML within businesses is that a majority of companies are actively working on a roadmap for handling data (68 percent), yet only 11 percent of these companies have completed this task, according to Forbes.

The models that are the most successful today are those that allow certain tasks to be taken over by AI whereby machine learning can acquire more information from and predict consumer behavior. Current ML models allow for rapid iteration of data and they deliver quick, reliable data sets that impact directly on the culture of work for businesses involved in any sort of real-time analytics, data integration and management, sales/revenue forecasting, and personal security and data processing.

REMEMBER

As machine learning has provoked worries in many quarters that jobs will be replaced by AI, the reality is that machine learning is already merely allowing humans to get on with the more interesting facets of their jobs as AI slogs away at the more mundane aspects of operations such as data mining. It's time to embrace machine learning for what it offers instead of worrying what it might take away.

Understanding the ML Lifecycle

To succeed with your ML investment, you need to rapidly implement and scale ML models across your entire organization. Usually these implementation scenarios span a large spectrum of ML use cases. The need for organizational speed in combination with growing regulatory scrutiny related to data and AI/ML, create new and unique challenges to move ML models from idea and experimentation, to production.

WARNING

In fact, currently only 35 percent of organizations indicate that their ML models are fully and successfully deployed in production. And on top of that, the journey doesn't end when models are deployed. On the contrary, it's vital to ensure that models continue to operate and perform as expected, or even better and more optimized, throughout their entire lifecycle.

A typical data science workflow goes through different phases but is complex and highly iterative. At the bottom of Figure 1-1 you can see the various steps of a data science workflow from *data* to *models* to *outcomes* to *business value*. The figure itself is mapped into three main phases: *data engineering*, *data science*, and *production*. This figure also has slightly more emphasis on the production phase and the governance aspects than a typical ML workflow.

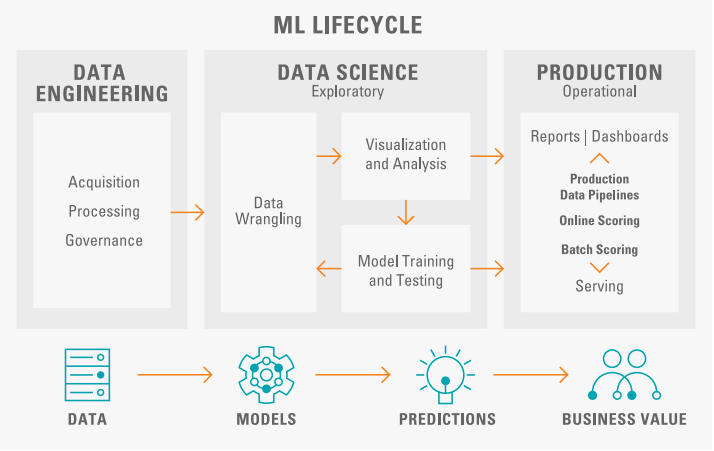

FIGURE 1-1: An overview of the key areas in the ML lifecycle.

REMEMBER

However, for enterprises, even before the data engineering workflow is started, it's vital to secure that you have a robust data management and IT practice in place to ensure enterprise governance, security, access control, and data lineage. This is a key part of the ML lifecycle because it eliminates things like silos and creates an observable, explainable, and transparent foundation for the execution of the ML workflow.

Data engineering

This phase consists of tasks such as data acquisition, processing, and governance. In this context data processing refers to transforming raw data by cleansing and preparing datasets to a more convenient format for a developer or a data engineer to run an analysis on.

REMEMBER

As organizations use data (and analytics) more, and for more important questions and user scenarios, the need be able to rely on the data, and therefore the need to govern those assets, increases. Every organization should be concerned about data quality in their source systems, but often these concerns are isolated and not visible across departments. Security, privacy, and regulatory compliance are also important elements of data governance.

Data science

REMEMBER

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Data science is also a concept to unify statistics, data analysis, machine learning, and other related methods to understand and analyze actual phenomena with data spanning from more traditional business intelligence (BI), through analytics, and more exploratory ML techniques.

More traditional analytics and BI includes areas such as data preparation, data visualization, and data modeling. Data preparation refers to the process of transforming and mapping data from its raw data form into another format with the intent of making it more appropriate and valuable for a variety of analytics use cases.

Data visualization (or data exploration as it´s also called) on the other hand helps identify interesting patterns and trends in the data that can be identified and analyzed through simple charts

such as line charts or bar charts. Data modeling is the process of producing a descriptive diagram of relationships between various types of information that are stored in a database.

Exploratory data science and ML includes probabilistic modeling and ML model development (model training and testing), where probabilistic modeling helps understand the probability of what could happen based on a variety of inputs and data.

ML model development can also be focused on automating processes or making ongoing predictions that learn/change based on new data. ML model development can be targeting a certain product, such as learning what a person regularly watches on Netflix and suggesting programming they will likely enjoy. Another objective can be a business segment, for example security and fraud prevention, where the ML model shall detect anomalies or patterns in incoming data to enable proactive detection of breaches.

Production

This part of the ML lifecycle is focused on the process of delivering the outcomes (better automation, predictions, innovations) to stakeholders (customers, internal business). There are several different ways to deploy a model and it´s key to understand the end user (customer) objective to determine the technology required.

REMEMBER

What the deployment phase actually means can differ a lot depending on what type of use case you're trying to realize. It could be as simple as generating a report, or as complex as implementing a repeatable data science process critical for business operations. Regardless of use case, however, successful production ML requires a streamlined, frictionless and predictable deployment, and ongoing governance of ML models in production, at scale.

Identifying Challenges with Getting ML to Production

Deploying and scaling AI/ML across the enterprise requires implementing complex, iterative workflows end-to-end from capturing data through developing ML models to achieving the

expected outcomes. This is not an easy task. In addition, as the number of AI/ML projects and models multiply, production ML can be slow, cumbersome, and fraught with "false starts" that make it even more difficult and expensive.

While the end-to-end ML lifecycle has always been pitched as an actual "cycle," to date there has been limited success in actually managing this end-to-end process at enterprise level scale. Some reasons for this are:

- Data scientists are usually not trained engineers and thus don't always follow good DevOps practices.
- Data engineers, data scientists, and the engineers responsible for delivery operate in silos that creates friction between teams.
- The myriad of machine learning tools and frameworks fosters a lack of standardization across the industry.

Machine learning realization from an enterprise perspective is slow and tough to scale. There is not much automated, collaboration is difficult, and the actual operationalized models delivering business value are few.

WARNING

Many projects don't make it into production because of model inefficiencies that slow down or halt the entire process. Or, in many cases, organizations fail to adequately adopt production models because of a lack of internal knowledge on how they work and other cultural/business impediments.

As organizations start to see artificial intelligence (AI) and machine learning (ML) as fundamental and vital pieces of the company, organizations are usually wrestling with growing organizational pain. Isolated projects exist in silos across the enterprise, putting quality, security, governance, and compliance at risk. When applying an "AI factory" approach to turning data into decisions, you can make the process of building, scaling, and deploying enterprise ML solutions automated, repeatable, and predictable, but lose the business value along the way.

REMEMBER

Only through the industrialization of AI can you shift focus from technology solutions to business outcomes, empower continuous optimization, and encourage a learning culture across the enterprise.

Truly excellent industrial-grade ML requires transformation in almost every part of an organization, and as a result, production ML hurdles are often actually organization wide hurdles. Some of these hurdles are described in the following sections.

Unsatisfactory model monitoring

Efficient monitoring of models in operation is an essential element of production deployment as it provides visibility into its various phases. Poor visibility into mathematical metrics and to the external tools used for monitoring is a major challenge. Custom tooling to monitor the technical health of models doesn't scale and mathematical monitoring is hard, customized, and little to no tooling exists.

WARNING

On the other hand, standard monitoring tools tend to be too generic for identifying model drift, such as identifying whether the ML model execution is deviating from its objective. The truth is that what really happened based on the prediction is usually only understood well after the fact. This means that determining whether a model is functioning as it should needs to be customized on a per model basis.

Inefficient deployment

Data scientists today use a variety of different tools to solve many critical business problems. This often results in models being recoded into different languages as the initial language may not be used in the production environment. This leads to longer cycle times and potential inconsistencies in the translated model.

Inadequate ML governance

As AI/ML moves to production, the need to govern all IT assets (data, models, infrastructure, for example) and ensure security, privacy, and regulatory compliance increases. Defining standards for ML Operations (MLOps) is essential for deploying and governing ML models at scale for enterprises. This includes visibility of models within teams and across organizations. It enables teams to understand how ML is being applied in their organizations and requires an official catalog of models.

WARNING

In the absence of such a model catalog, many organizations are unaware of their models and features, such as where they are deployed and what they are doing. This leads to substantial rework, model inconsistency, recomputing features, and other inefficiencies.

Insufficient security measures

There is a need for end-to-end enterprise security from data to the production environment. The chosen platform must be capable of delivering models into production with inherited security, unified authorization, and access tracking.

Inability to scale

As the model moves forward to production, it's typically exposed to larger volumes of data. The platform must have the ability to scale from small to large volumes of data and automate model creation. Applied machine learning at enterprise scale requires a particular combination of cutting-edge technology and enterprise expectations.

REMEMBER

Additionally, it's not only about scale of data, but the number of actual models in production. Operating and monitoring a few models may be fine, but when you scale up to hundreds of models it's very difficult to make sure they are not drifting and are maintaining their reliability. This is a common concern for customers in the banking sector, where they have entire teams of ML engineers dedicated to just keeping their models accurate in the long term.

WARNING

A major concern is also when brand new (and at times) untested or poorly supported technology, tools, and systems are rapidly deployed into enterprise application workflows aimed for the ML lifecycle and are expected to perform at least as well as existing software that has been in place (at times) for years.

IN THIS CHAPTER

- Describing the capabilities and process steps in production ML
- Identifying what is needed to succeed with production ML
- Sorting out the governance aspect
- Explaining how to approach scalability

Chapter 2
Applying the Right Process Approach

Machine learning (ML) has become one of the most critical capabilities for modern businesses to grow and stay competitive today. From automating internal processes to optimizing the design, creation, and marketing processes behind virtually every product consumed, ML models have flooded almost every aspect of our work and personal lives. And for businesses, the stakes have never been higher. Failing to adopt ML as a core competency will result in major competitive disadvantages that will define the next market leaders.

Because of this, business and technology leaders need to implement ML models across their entire organization, spanning a large spectrum of use cases. However, this sense of urgency, combined with growing regulatory scrutiny, creates new and unique governance challenges that are currently difficult to manage. Questions related to how ML models should run and stay in control after they're deployed are becoming important to understand and manage. This includes questions such as: How are my models impacting services provided to end customers? Am I still compliant with both governmental and internal regulations? How will my security rules translate to models in production?

This chapter helps to describe how to approach production ML in practice including detailing which capabilities you need to get in place and how to succeed with ML in operations. ML governance and scalability in production is also addressed.

Detailing Key Capabilities of Production ML

Machine learning holds great promise. But like most great things, it asks for a bit of patience, an open mind, and a willingness to persevere when small wins prove elusive.

If you aspire to bring machine learning into full-scale operational use, then start small and scale your efforts with a well thought-out and informed strategy. Remember, effective and innovative machine learning capabilities are cultivated over time.

The main capabilities needed for production ML is described in Figure 2-1. These production ML capabilities include the technology infrastructure and tooling necessary to deploy ML algorithms and data pipelines reliably so as not to destabilize other parts of the workflow.

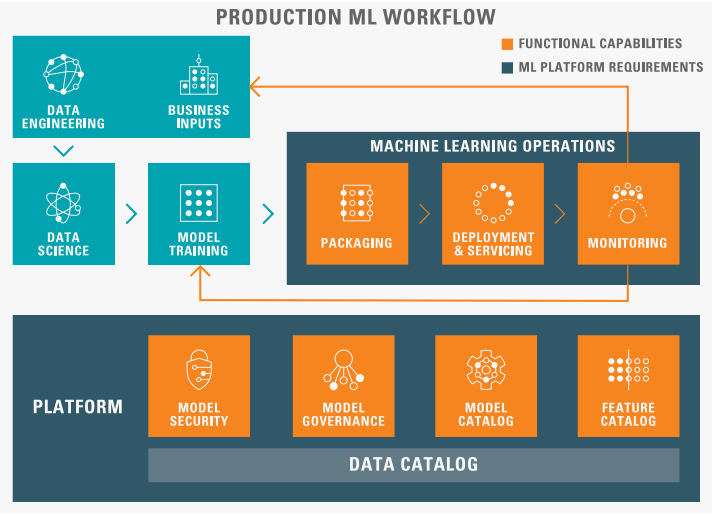

FIGURE 2-1: A production ML workflow with main capabilities.

Production ML spans from the data science tools used to select and train ML algorithms down to the hardware that those algorithms use to process data.

Production ML capabilities include:

- **Packaging, deployment, and serving** are the three first steps in production ML. Right *packaging* is necessary for automated deployment of production models and to address multiple deployment patterns. Enterprise level *deployments* need high availability, autoscaling, and strong authentication features. *Serving* makes a trained model available to other software components. Models can be served in batch or real-time modes.
- **Monitoring** is an important element of operationalizing ML. After a model is deployed into production and providing utility to the business, it's necessary to monitor how well the model is performing. There are several aspects of performance to consider, and each has its own metrics and measurements that impacts the lifecycle of the model. However, *monitoring* is done at various stages of the lifecycle, for example to check input and output distribution, look for skew, model drift and accuracy change, add custom thresholds, send emails with results, and trigger alarms.
- **Model governance, cataloging, and lineage tracking** are a basic requirement for model *governance* and enable teams to understand how ML is being applied in their organizations. Governance requires a centralized *catalog* of models and features that facilitate tracking models and their features throughout their lifecycle to understand these features and their relationship with the data catalog. Catalogs also facilitate authorization and tracking access to models thereby maintaining end-to-end security of the environment.

  Data lineage regards the need to have visibility into the full ML lifecycle, starting with where the data originates and ending with the ongoing production environment. This includes every process along the way, from data ingest to data engineering, to model building, deployment, serving, and production, and even security and governance visibility — who touched what when, what data powered what models, and who had access to this data at every step.

Getting to Production ML Successfully

MLOps (ML Operations) is a relatively new term within production ML. In its purest form it can be seen as the true instantiation of the automated production ML lifecycle. MLOps is the logical reaction to the current difficulties' enterprises face putting machine learning into production.

MLOps refers to a practice for collaboration and communication between data scientists and operations professionals to help manage the production ML lifecycle. Similar to the development operations (DevOps) or data operations (DataOps) approaches, MLOps looks to increase automation and improve the quality of production ML while also focusing on business and regulatory requirements.

Figure 2-2 shows the different steps in the production ML workflow including the responsibility of different roles. As you can see it´s built on close collaboration between multiple data roles. There are multiple hand-offs between stakeholders also in the production phase. A typical setup (as shown in Figure 2-2) shows that IT or DevOps classically runs the model packaging, while it's MLOps or Engineering that manages the model deployment and model serving activities. This is usually the case because of the model expertise needed for the deployment related activities. Finally, the monitoring activities are usually handled by an ML engineer who can quickly respond to changes or needs, but this is highly tied to the business itself, or the business function the model has in production.

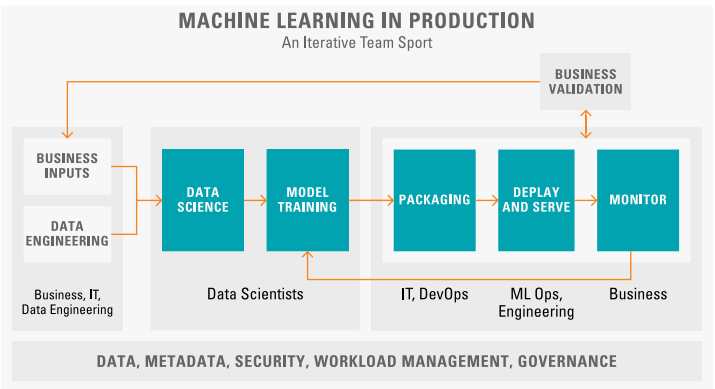

FIGURE 2-2: Roles involved in the production ML workflow.

However, after the model is up and running, its performance must be continuously monitored based on agreed upon model measurements and thresholds. If the model starts to behave strangely, model retraining is triggered as a request back to responsible data scientists.

Furthermore, underpinning the production ML workflow are fundamental aspects such as securing continuous and reliable data operations, security operations, workload management, and an operational governance model. From a governance aspect this is a real challenge, because there is little to no standardization in the production ML space.

WARNING

The fact is that the whole industry has begun to hit its breaking point, and technology is evolving rapidly to meet demand and alter the current standard for ML in production. Open source frameworks such as mlflow and kubeflow compete to become the standard of the open source landscape, while new startups slap user interfaces on these solutions in an attempt to bring "proprietary" MLOps products to market.

The same way as good DevOps ensure that the software development lifecycle is efficient, well documented, and easy to troubleshoot, a similar set of standards for machine learning addressing the operational aspects of ML is very much needed.

So, what does good MLOps look like? How can you ensure that you're getting your production ML right from start? Well, good MLOps looks a lot similar to good DevOps.

- Reduce the time and difficulty to push models into production.
- Decrease friction between teams and enhance collaboration.
- Improve model tracking, versioning, monitoring, and management.
- Create a truly cyclical lifecycle for the modern ML model.
- Standardize the machine learning process to prepare for increasing regulation and policy.

While this might seem like a good and comprehensive list, the question is how you put this into practice. You may have even dabbled with AI and machine learning (ML) models in a few pilot

projects. But has your organization actually delivered on the promise of AI with tangible business benefits? If not, you aren't alone; most of your peers are facing similar issues.

WARNING

Many enterprise organizations start their AI projects by setting up an innovation lab, only to realize later that they haven't been able to operationalize their ML models into real-world business processes. And it´s only operational ML models — models that have been integrated with business functions in production — that can deliver business value.

So, what do you need to get production ML right? Some important considerations are described in the following sections.

Identifying ML business objectives

Many AI/ML projects fail to deliver to expectations due to exaggerated expectations of what AI can do. As part of your initiative, it´s vital to identify what your organization is expecting from ML and what the main objectives are. You can start with the business perspective. For example, what metrics are you trying to improve, are you trying to reduce your customer churn rate, reduce cases of fraud, or reduce the time spent on processing customer applications?

TIP

Make sure to clearly identify the use cases you're targeting from the start, define measurable goals, benchmark current performance, and then realistically define agreed success criteria. All of this is not only valuable for getting your overall AI/ML initiative on track, but it helps you get a realistic picture of what it is that you actually have to manage in a production ML setting.

Securing commitment from key stakeholders

After you identify the main use cases, you can identify the different stakeholders who need to be involved in an operational setting. To figure this out, you need to detail your use cases all the way to the production setting. Include how the model output will be used and who will use it, because there's no point in having an ML system crunch numbers and make predictions if the output is unusable, inaccessible, or simply not planned to be a part of the decision-making process.

TIP

It's therefore essential to fully understand how the business expects the predictions to be made available to the downstream tools/processes/people and what expectations this puts on the production ML capabilities.

Finding or building the right competence

The shortage of available data science talent on the market for these types of roles remains a fundamental challenge for any company. But success in the area of AI/ML requires more than just data science skills. To succeed you need to understand and address the end-to-end view from data prep and model building, to model training, deployment, and production. Data science is truly a team sport requiring multiple different roles, including data engineers, ML engineers, and operational support engineers.

Organizing and scaling the team effectively is another challenge. You need to ask yourself whether you have the right people and skills in-house to take the project from idea through development, deployment, and finally running it in production. You also need to determine whether you want to build up the skills internally through hiring and retraining or whether you want to outsource the job. Building up the skillset yourself most probably helps you scale in the long run, whereas outsourcing or third-party services may help get the project up and running faster short term.

Establishing the right technology platform

Unfortunately, it's not uncommon to see data science initiatives fail due to a lack of planning on the technology front. And it's not just about having the right technology and tools for building and developing ML models, it's very much about the deployment and operationalization aspects of ML models. It's the operationalization that actually represents the toughest challenge. It's vital to consider the entire ML lifecycle from the start, even if the production phase feels far away. Some aspects to consider are:

- **Data:** Ensure that you have the right data for your use case and that you understand all aspects of that data such as data source, quality, volume, format, sensitivity, and so on. Take the time to think through how data will be acquired and

prepared in an operational setting including aspects such as processing capacity, security, and data privacy. These and more are all aspects that dictate the type of technology choices that are made for your production setting.

- **Applications:** Remember, this isn't about picking just one application or tool, there are a multitude of tools in the AI/ML and data science ecosystem, and there is a reason for that. Which tool to use really depends on what you're trying to do, and what type of use case you're addressing and what the production setting needs to serve. One tool that might be perfect for a certain use case or ML technique might not be the best one for another problem. The ML space is continuously evolving, and your technology stack needs to support multiple different and constantly changing needs from an operational perspective.
- **Infrastructure:** Public cloud services do offer some advantages, but the cloud is not always the answer for large-scale AI/ML initiatives in enterprise organizations. Instead, many enterprises are adopting a mixed hybrid cloud approach, using on-premises or cloud infrastructure depending on the use case and stage in the ML lifecycle and depending on where the data they need is located. This allows organizations to leverage what they've already built on-premises while taking advantage of the agility and elasticity offered by public cloud services. The use of cloud-native technology, such as containers, for AI/ML workloads has also greatly improved the speed of development while allowing the flexibility to "build anywhere and deploy everywhere."
- **Standardized workflows:** Machine learning workflows differ from typical software engineering workflows. Most enterprises lack standardized ML processes for model management, monitoring, and retraining. This often hinders collaboration and leads to delayed or lost business value. In addition, ML models are trained on historical data, so their accuracy tends to degrade over time as the underlying data changes. Detecting these deviations requires specialized debugging tools and processes to retrain the models once pre-defined thresholds have been crossed. Setting up the workflows, applications, and infrastructure to store multiple model versions, trigger retraining, and seamlessly update models in production is critical to ML operationalization.

More details on technical consideration in production ML is covered in Chapter 4.

Enabling Holistic Governance

WARNING

If anyone offers you an out-of-the-box machine learning "solution," take your business elsewhere. You can't purchase truly effective machine learning off the shelf, nail it onto an existing application or process, and reap the rewards. That's because machine learning isn't a single tool or platform or solution; it's a capability, one that can never really be mastered over time by taking a software-only approach.

In truth, machine learning thrives best when it's supported by an organizational ecosystem. Before you can deliver machine learning capabilities, you must first have the right data governance and data engineering tools and standards in place to develop your machine learning models at their core.

REMEMBER

Likewise, ongoing data governance, model sustainability, and the integration capabilities of your architecture greatly impact a machine learning capability as it moves forward into production. Machine learning must be viewed holistically as an integral part of your data science strategy. By putting it in context alongside your existing IT environments, processes, applications, and workflows, you'll better support business processes and drive greater results.

How well you can continuously govern data across the entire organization plays a major role in the success and sustainability of your machine learning initiatives. While automation, business predictions, and product innovations are the goals of machine learning, those goals are achieved only by creating and maintaining algorithms, and an algorithm can only be as accurate as the data that shapes and feeds it.

Data fuels the model, which in turn drives a given set of machine learning capabilities. Your data changes over time. New data sources come in, patterns evolve, and how well you govern your data may fluctuate. All these things impact a model.

Building enterprise capabilities with machine learning models at their core is different than traditional software application

development. And unlike most apps or microservices, models have the potential to shift in real time, depending on their function and the data they interact with.

If you want to see immediate and long-term success of your machine learning initiatives, you have to understand the dynamic, fluid nature of the models that drive machine learning capabilities. After models are deployed into production, they must be continuously monitored for updates because the data that feeds a model can change naturally over time. A model's data can also become corrupted and inaccurate. In either case, continuous updates to models ensure they continue to deliver the results and business outcomes they were designed for.


TIP

Some organizations try to force machine learning into a rigid structure where it doesn't fit. Others completely isolate it where it can't benefit anyone or anything. But solving the problem isn't really about trying to fit machine learning into your existing organizational scheme; it's about making the structure of your organization more flexible so that machine learning can be embraced.

There also comes a time when a model reaches its natural end-of-life. This could happen for any number of reasons. Maybe the original business problem that the model helped to solve is no longer an issue for the organization. Or, perhaps a model existed only to push recommendations to a customer-facing service that's being removed. By continuously monitoring your models, you can retire ones that are no longer needed and free up your resources.

Beyond regulatory or legal concerns, there are a number of reasons to have governance processes and procedures for machine learning. With proper ML governance in place it facilitates increased productivity. This is made possible through easier reuse of assets like data, models and features.

Another reason for investing in ML governance is because it assists you in controlling and maintaining models across many different business lines, hence ensuring that business-critical applications are doing what they're intended to do. Or finding those that aren't. Proper ML governance also allows you to access the organizational history of models and predictions including deprecated assets.

So, to explain how to approach ML governance, it's worth defining what models and features are conceptually, as shown in Figure 2-3. The term "model" is quite loosely defined and is also used outside of pure machine learning where it has similar but different meanings. In the context of production ML, it's defined as a combination of an algorithm and a set of configuration details that can be used to make a new prediction based on a new set of input data. The algorithm can be something such as a Random Forest, and the configuration details would be the coefficients calculated during model training. It's like a black box that can take in new raw input data and make a prediction.

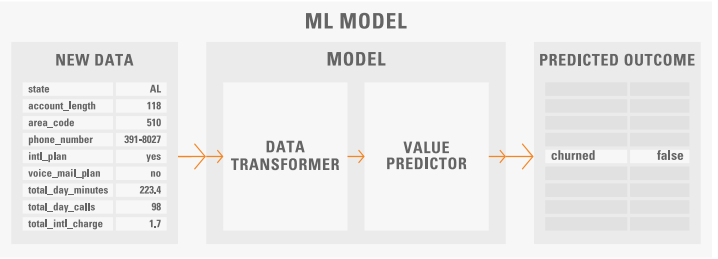

FIGURE 2-3: How a machine learning model works.

Many enterprises operate ML model infrastructure at different sizes and maturity that they require tools to help them govern their models. Ultimately, the need for ML governance can be distilled into the following key areas:.

- **Visibility:** A basic requirement for model governance is enabling teams to understand how machine learning is being applied in their organizations. This requires a canonical catalog of models and features. In the absence of such a catalog, many organizations are unaware of their models and features, where they're deployed, what they are doing, and so on. This leads to rework, model inconsistency, recomputing features, and other inefficiencies.
- **Explainability:** Models are often seen as a black box where data goes in, something happens, and a prediction comes out. Securing model explainability is key as it helps to increase trust in the model. It includes a description of the internal mechanics of an ML model in human terms.

- **Interpretability:** This refers to the ability to a) understand the relationship between model inputs and outputs, and b) to predict the response when the input changes.
- **Reproducibility.** Refers to the ability to reproduce the output of a model in a consistent fashion based on the same inputs.

To govern these four aspects efficiently requires a common framework and functionality across the organization, including a tie into the source data, a clear understanding of the internal workings of the models' like code and training data, and application support to monitor the models themselves. See Chapter 4 for more details on technology aspects to consider in production ML.

IN THIS CHAPTER

- Applying the right operational mindset in machine learning
- Establishing ML production focused culture
- Nailing down an ML efficient organization
- Finding the right competence
- Selecting a leadership strategy

Chapter 3
Sorting Out the People Perspective

Organizational and political issues are probably larger and more impactful to your ML outcomes than you may think.

Yes, ML projects are different from traditional projects whereas they require different skill sets, platforms, software, and workflows. However, they should still "project management wise" be managed as any other project. But because AI/ML projects tend to blend scientific methodology with business focused outcomes, traditional corporate decision support tools are often ineffective. On top of that, ML work is highly iterative and even well-defined criteria may need to shift throughout different phases of the project as data scientists and business owners alike explore what is possible, ethical, and profitable. This can cause confusion in the organization on when an ML project is actually "done" or still "under development."

When it comes to the production phase of machine learning it's also easy for organizations to take a "this will be somebody else's problem" approach. Vital areas such as data pipeline and data management, model monitoring and alerts, and other production

like activities may not be the most exciting areas to think about, but failure in any of these parts can crush even the most promising AI/ML efforts.

This chapter guides you in how to address the people perspective as part of successfully getting your ML investment in production.

Creating a Production Mindset

It's not an exaggeration to say that machine learning has the potential to transform your business. ML can automate processes, uncover new insights, make your products and services better, and make customers happier. However, integrating ML capability into your organization requires full operational transformation and a true experimentation-based approach. And to succeed with this it´s important that you *keep an open mind*.


TIP

Keeping an open mind is something you must set early on and reinforce often on the path to making machine learning operational for your business. To be successful with machine learning development, you must be incredibly intentional about the problems or opportunities you want to tackle. If you start out by solving something that is achievable and realistic, you can prove the value of ML fast and make ML real for your organization. Many times, ML becomes obscure, almost magic-like to people who do not fully understand the area. You can tackle this skepticism (or even fear) of AI/ML by delivering a tangible value early on.


REMEMBER

To shape your goal, think on a granular level — what incremental, positive change can machine learning make to your processes or applications? What opportunities can you go after with this capability? When you have your list, consider every option with an open mind. If you want to make machine learning operational across your organization, you need to start small. Build a model that you can test and iterate on without completely upending existing workflows.


TIP

A good way to think about your early applications of ML is that they should be of an assisting nature, rather than completely replacing a role or taking over a task and fully automating something right from the start. It's also important to make sure that your ML application has a feedback mechanism so you can verify that it's working after it's deployed. Deploying any technology

has a cost associated with it, so you need to know sooner rather than later whether it's providing the value that the organization expects.

After you identify the opportunities you're going to go after, it's time to experiment. To get that started you need to build a multi-disciplinary team with data scientists and subject matter experts (SME), as well as data management and governance resources. An SME can play a critical role in your team at the beginning of experimentation. They can effectively help you with the business context of your use cases, as well as access and interpret data (which is typically a big hurdle at the beginning of an ML project). However, once data is collected, data custodians can build information models and other data services that can be automated, to enable data scientists to be agile in building, testing, and iterating quickly.

REMEMBER

At the same time, you must prioritize governance, security, and transparency — but in a way that still allows teams to explore and experiment. To keep the cost of governance down and the rate of experimentation high, consider a platform that enables continuous and highly automated monitoring and security. This way you can keep your team focused on innovation, as opposed to the labor-intensive manual tracking and monitoring of model performance.

Building the Right Culture

To succeed with building the right data driven culture with a machine learning production mindset in your organization, it's important to be able to anticipate and identify unique barriers to change.

Some obstacles, such as employees fear of becoming obsolete, are common across organizations and are easy to anticipate. But a company's culture may also have distinctive characteristics that contribute to resistance. For example, if a company has relationship managers who pride themselves on being attuned to customer needs, they may reject the notion that a machine could have better ideas about what customers want and therefore ignore an AI tool's tailored product recommendations. And managers in large organizations who believe their status is based on the

number of people they oversee might object to the decentralized decision making or reduction in reports that AI capabilities could allow.

In other cases, siloed processes can prevent the broad adoption of AI that is wanted and expected. Organizations that assign budgets by function or by business unit may struggle to assemble interdisciplinary agile teams, for example.

Some solutions can be found by reviewing how past change initiatives overcame barriers. Others may involve aligning AI initiatives with the very cultural values that seem like obstacles. At one financial institution with a strong emphasis on relationship banking, for example, leaders in the organization used facts to show how the AI solutions they had actually enhanced the ties with customers. The bank created a booklet for relationship managers that showed how combining their expertise and with the AI-tailored product recommendations could improve customers' experiences and increase revenue and profit.

REMEMBER

Understanding and trusting that AI/ML capabilities are good for both the company and the people in it is fundamental for a cultural change to happen. The data science teams in your company play a very important role in this. From a data science team perspective, any team member must be able to understand and articulate what a certain model can do and why a model is producing the results that it is. Your data scientists need to take the time to explain the outcomes of machine learning models to your business teams, who in turn need to be able to explain the resulting predictions and business decisions to your customers or shareholders. Succeeding with this is key to getting the right culture in your organization.

TIP

For improved model explainability you could also consider using explainability frameworks such as LIME or SHAP to deliver self-explaining applications. This app functionality helps bring the business users closer to the data science.

TIP

As an exercise, see whether you or someone in the data science team can take an outcome determined by a model over the past few months and explain how it came to the decision it did. If it's not possible, it might be time for a meeting with your lead data scientist.

In that spirit, reporting to C-level executives is important for demonstrating success and keeping investments into your initiatives a priority. Early on, define what a "win" and "failure" looks like. Remember, this can go beyond the performance of a model and be based on what was learned through the model's recommendations. Then, establish a pace for reporting to executives and make sure to utilize data and analytical outcomes (ML or otherwise) in them.

When it comes to making machine learning operational for your business — and sustaining it — there are many factors to consider. One of the most important being the human touch.

Your ML models require consistent attentiveness and maintenance over the course of their lifecycles. Sometimes data expands out of a range and it confuses the model. Occasionally, a simple human error can result in complete disruption of the model performance. This is the point where your platform can make or, quite literally, break your efforts.

Your production environment should be built on a robust, intuitive architecture that can keep your models secure and alert you when they are not. Sufficient level of visibility into your ML production environment helps your teams navigate the inevitable errors (human or otherwise) that occur and can even empower them to find new solutions or opportunities going forward.

Setting Efficient Organizational Structures

Your team of data scientists have iterated on a machine learning model that shows promise. Maybe they discovered it is in fact possible to predict a target variable with a high degree of accuracy using a large data set. The breakthrough provides a morale boost, and other viable machine learning algorithms follow suit. You started your machine learning journey with a list of problems and opportunities. For at least a few of those items on your list, you're able to say, "yes, machine learning can help." Eureka!

WARNING

Now it's time to take the leap, and this is where many organizations can unfortunately fall short. There's a wall that seems to exist between experimentation and large-scale production. Many organizations hit this wall because they don't know how to weave

machine learning development, production, and maintenance into their existing processes, workflows, architecture, and culture. Some organizations try to force machine learning into a rigid structure where it doesn't fit. Others completely isolate it where it can't benefit anyone or anything.

REMEMBER

But solving the problem isn't really about trying to fit machine learning into your existing organizational scheme; it's about making the structure of your organization more flexible so that machine learning can be embraced.

Your enterprise undoubtedly has standards in place for code source control and integration. Have your team adopt and leverage the tools and best practices that already exist within your enterprise. Doing so makes integration much easier down the road.

WARNING

Imagine you want to try skiing. You went skiing a couple of times and then you told everybody that you can ski. Three years later you haven't skied again, but you still own skis and you're still telling people that you ski. That's exactly what's happening with AI. People are dabbling in it. They put it off to the side, they don't integrate it into the day-to-day life of their organization, and then three years later, they're doing artificial intelligence over here in the corner where it's not benefiting anybody.

From an organizational perspective, there are any number of ways to structure the business for optimized machine learning. It's all about identifying what works best for your company. One way is a centralized team strictly focused on data science. This team primarily builds machine learning models and then puts them into production across other parts of the enterprise where these capabilities can be used. This approach revolves around the idea of building a center of excellence.

Another organizational approach embeds one or more data scientists into business product teams across the enterprise. This lets your data scientists get close to a business problem so they can better understand it. However, they still need to collaborate together as a team of data scientists, which again underlines the importance of choosing a machine learning platform that lets them efficiently collaborate and share knowledge across departments.

REMEMBER

The best organizational approach for you depends on your particular business needs. Whatever structure you feel is best for your enterprise, what's most important is maintaining the ability for your data science teams to collaborate and share ideas and best practices with each other.

Hiring ML Talent

Building the right team structure up front is important, but it can be a challenging task because the very nature of machine learning tends to blur organizational lines and breaks down the barriers between traditional roles. It's said that data knows no organizational boundaries and it's probably true considering how different usage scenarios for a certain data type can spread all across a company.

During the model development phase your data scientists may grow more sophisticated with every iteration and experiment that is performed. However, when your data scientists are taking the model into production you might hit a wall. Especially if your team consists solely of data scientists. Often, data scientists have trouble making the leap from building models to putting those models into production and integrating them with other systems and applications. That is why you need to secure a cross-functional data science team from the start.

TIP

Try to build a team whose experience, talents, and capabilities — including data engineering, data science, software development, DevOps, product development, and domain expertise — overlap. You should look for candidates with the core skills that are necessary to accomplish your most important tasks and let them learn from one another. It's also important to search for individuals that have an interest and willingness to expand their skill sets and knowledge base over time. This is important because data science as a discipline is still new and constantly evolving and you can also benefit from cross-functional capabilities within the same role.

REMEMBER

However, the "do-it-all" data scientist is a rarity. That's because new roles around data science and machine learning are still being defined, and unique skill sets continue to merge into new job titles as time goes on. For now, it's not about finding the person who can tackle every aspect of machine learning; it's often about

cultivating the people you already have. For example, if you have a lot of data scientists in your team, perhaps you can encourage those who are interested to invest in their data engineering skills as well.

Adjusting recruitment strategies

As the uses of artificial intelligence technology expand into every industry, the available pool of machine learning talent are coming from a very small pool of available candidates. Even from an international perspective. But how can you find, tempt, recruit and *retain* ML talent when you're competing with acknowledged data companies such as Google, Amazon, and Microsoft? The answer lies in getting creative with your recruitment and hiring strategies.


TIP

The starting point when reevaluating recruitment strategies for high-end ML or other AI roles is to adapt strategies based on the experience level you're looking for. You can't assume that the recruitment strategy is the same for a junior versus senior data scientist. To access the talent you're looking to hire, you need to go where they're found.

For more junior-level roles, universities, hackathons, and specialized training programs are great sources of professionals that are versed in the latest tech that can help build your AI/ML department, and then over time they can transition into senior-level roles. For more senior or experienced roles, qualified applicants are most commonly found through network connections, academic papers, and academic conferences.

Understanding the need to adapt your recruitment and hiring strategies based on the level of experience you're looking for sets you up for greater success when it comes to attracting and retaining the professionals you need.

Motivating talent to switch jobs

Just how hard can it be to land the machine learning professionals you need? Pretty hard it seems — and expensive if you believe the average salary a data scientist is hired at these days. But the good news is you don't have to have a similar budget to get the experts you need. You just need to know what motivates the talent you're interested in, so you can convince this in-demand talent to switch jobs.

When it comes to in-demand ML talent, their motivation usually boils down to the following benefits:

- Intellectually challenging opportunities
- Competitive compensation and resources
- Location
- Brand recognition
- Diversity of problems
- Impact of their work
- Quality of the team
- Access to data
- Sufficient AI/ML infrastructure


TIP

If you find that you can't afford to pay a salary in the expected range for a data scientist, look into long-term incentives you might be able to offer as an alternative. For example, consider incorporating remote work flexibility if you're located in an area that has a hard time recruiting top ML engineers. Analyzing and providing these incentives help you recruit machine learning talent that would otherwise be out of reach.


REMEMBER

Not everyone thrives in environments at large tech companies. Expressing the differentiators between your business and leaders like Apple in what you're able to provide, whether that's with work environments or internal growth opportunities, could help sway talent to accept your offer instead.

Finding universities to partner with

When looking for more junior-level data science roles, you could consider partnering with a university and funding or supporting a university project that can open a machine learning talent pipeline that leads to paid internships and post-graduate employment. Given the short supply of in-demand AI-based talent, some companies have found this to be an incredibly effective tactic to recruit machine learning talent directly from the source.


TIP

If you decide to go down this route be sure to fully flesh out the project you have in mind to pitch. Given the success of this strategy for a variety of companies looking to recruit machine learning talent directly, these types of partnership programs have become

more popular. Meaning you need to develop an exciting project and clearly articulate the benefits of it to attract interested students.

Don´t give up! The machine learning talent you're looking for is out there. These in-demand experts simply require a little more creativity when it comes to recruiting.

Securing Sustainable Leadership Strategies

Artificial intelligence is reshaping business, but not at the pace that many assume. It's true, however, that AI is now guiding decisions across many different areas, from crop harvests to bank loans. The technologies that enable AI, like development platforms and vast processing power and data storage, are advancing rapidly and becoming increasingly affordable. The time seems ripe for companies to capitalize on AI. Indeed, it's estimated that AI will add $13 trillion to the global economy over the next decade.

WARNING

Yet, despite the promise of AI, many organizations' efforts with it are falling short and one of the biggest mistakes leaders make is to view AI as a plug-and-play technology with immediate returns. Deciding to get a few projects up and running, they begin investing millions in data infrastructure, AI software tools, data expertise, and model development. Some of the pilots manage to carve small gains in pockets of the organizations, but then months or years pass without bringing the big wins that executives expected. Firms struggle to move from the pilots to companywide programs — and from a focus on discrete business problems, such as improved customer segmentation, to big business challenges, such as optimizing the entire customer journey.

Leaders also have a tendency to think too narrow about AI requirements. While cutting-edge technology and talent are certainly needed, it's equally important to align a company's culture, structure, and ways of working to support broad AI adoption. But at most businesses that aren't born digital, traditional mindsets and ways of working run counter to those needed for AI.

To scale up and make AI operational, companies must transform the following:

- **Move from siloed work to cross-domain collaboration.** AI has the biggest impact when it's developed by cross-functional teams with a mix of skills and perspectives. Having business and operational people work side by side with data science experts ensures that initiatives address broad organizational priorities, not just isolated business issues.

  Diverse teams can also think through the operational changes new applications may require — they're likelier to recognize, say, that the introduction of an algorithm that predicts maintenance needs should be accompanied by an overhaul of maintenance workflows. And when development teams involve end users in the design of applications, the chances of adoption increase dramatically.

- **Move from experience-based decision making to data-driven decision making.** When AI is adopted broadly, employees up and down the hierarchy augment their own judgment and intuition with algorithms' recommendations to arrive at better answers than either humans or machines could reach on their own.

  But for this approach to work, people at all levels have to trust the algorithms' suggestions and feel empowered to make decisions — and that means abandoning the traditional top-down approach. Employees having to consult a higher-up before taking action inhibits the use of AI.

- **Move from rigid and risk-averse to agile, experimental, adaptable, and ready for operation.** Organizations must scrap the mindset that an idea needs to be fully baked or a business tool must have every bell and whistle before it's deployed. In the first iteration, AI applications rarely have all their desired functionality. A test-and-learn mentality reframes mistakes as a source of discoveries, reducing the fear of failure.

  Approaching your solutions from an operational perspective, getting early user feedback and incorporating it into the next version allows space for correcting minor issues before they become costly problems. Development can speed up, enabling small AI teams to create minimum viable products that are ready to deploy and operate in a matter of weeks rather than months.

WARNING

These fundamental transformations that are needed don't come easy. They require leaders to dare to invest in full transformation and to prepare, motivate, and equip the workforce to make the change happen. But leaders must first be prepared themselves. Many failures are caused by the lack of a foundational understanding of AI among senior executives. And to fully benefit your AI investment from development to a production setting, this needs to change.

The ways AI can be used keep expanding. New applications create fundamental and sometimes difficult changes in workflows, roles, and culture, which leaders need to guide their organizations through carefully. Companies that excel at implementing AI throughout the organization, including mastering operational aspects, will find themselves at a great advantage in a world where humans and machines working together will outperform either humans or machines working on their own.

IN THIS CHAPTER

- Walking through the technical aspects of production ML
- Picking the right technology platform
- Presenting the Cloudera Machine Learning Platform

Chapter 4
Understanding the Technology Perspective

Deploying and scaling AI/ML across the enterprise is a daunting task that requires implementing complex, iterative end-to-end workflows from data to models to outcomes. And as the number of AI/ML projects and models multiply in your organization, the road to production ML and realizing your ML investment can become slow and expensive to maintain.

REMEMBER

It's important to realize that if the area of machine learning is much more immature than software engineering when it comes to available tooling and standardized infrastructures, it's even more immature in the production part of the ML lifecycle. So even if it's apparent that what's needed is an open, unified, collaborative, secure and governed enterprise-grade environment to run and manage all AI/ML models with transparency, consistency, trust and high-performance, it's still not clear what that means in terms of platform selection.

This chapter therefore focuses on describing what type of platform characteristics you should be looking for to overcome key production AI/ML challenges such as model monitoring, deployment, security, governance, scalability, and infrastructure.

Describing Technical Considerations in Production ML

Production ML is one of the most difficult, but at the same time most important, processes to unlock ML value. It requires coordination between data scientists, IT teams, software developers, and business professionals to ensure the model works reliably in the organization's production environment. Sometimes, there could also be a discrepancy between the programming language in which a ML model is written and the programming language that components in the production system are written in. Obviously, recoding the model could cause further delay to the model implementation. Practitioners and platform architects must be wary of point solutions that appear to solve a particular part of the problem of taking a use case to production, but introduce their own security and governance issues, such as by requiring data to be moved between multiple platforms, or leading to "shadow IT."

However, this doesn't mean that you have to stay with one programming language throughout your data and AI/ML infrastructure. The important thing is to identify and understand these different dependencies and then take conscious design decisions along the way. For example, say you have a data science organization where the majority of your competence is actually software engineers, and only a small part of the team are data scientists. In that scenario it's likely that while the data scientists are skilled in programming languages such as Python and R, the software engineers are experts in Java. In this type of situation, an approach could be to write all the traditional software components and application programming interfaces in Java, while the models are written in Python. But again, make sure to think through the production aspects thoroughly, so you don't end up in a situation where the production environment for the models are hindered due to incompatible technology stacks.

REMEMBER

When implementing production MLOps at scale you need to think holistically about the technology infrastructure and applications necessary to deploy ML algorithms and data pipelines. The scope includes everything from the data science tools used to select and train ML algorithms, to the hardware that those algorithms use to process data. It also includes the databases and message queues used to store, move, monitor, and track technical and mathematical metrics.

With regards to metrics, it's important to consider how you want and need to measure ML success. What's important to measure in your business context? Is it on a per model level? Or is it more important to measure success of the business outcome? If you're focusing more on a per model level, you should consider how you need to prepare for and handle ML model re-training. What type of model updates will be needed in your line of business? For example, are your models running on live data streams that could change fast, and therefore make your models obsolete quicker? An example could be data from social media feeds. In that case you should consider a dynamic model monitoring workflow and even automated model re-training setup.

TIP

If the use cases are based on a more stable type of data, for example on what a bicycle looks like, used for an image recognition program for identifying bikes. It's not very likely that a ML model that is trained on images of bikes is likely to change any time soon. A bike will most probably look like a bike for the foreseeable future. Perhaps the stable data feed also runs through the model based on scheduled or ad hoc batch uploads? In these types of cases the manual re-training approach would be sufficient for your needs. Make sure not to make your technology choice based on an imaginary scenario where your platform needs to handle any kind of use case, at any time, anywhere, if it's not needed for your business situation. Yes, it's important to have an open, flexible, scalable, and reliable platform, but it should also match the need for your ML use cases and their business importance.

WARNING

There are different approaches to deploying production models — and identifying which ones work best depends a lot on the use case you're trying to realize. Hence, you can't assume that the same production ML approach will work optimally for all your use cases. So, make sure to analyze and compile a view of different types of use cases you have, or are aiming to have. You need to fully understand what technical requirements different types of use cases drive, including how operational requirements are different for different use case types.

The ML use case characteristics are more important than you might think. For example, the level of business criticality of the model outcome will set different requirements on production ML monitoring and governance. Even the sensitivity of the data used, or the use case sensitivity, impacts your technology choices and could mean that you have to prioritize differently.

REMEMBER

Another aspect that is vital to consider is packaging of models. The right packaging is necessary for automated deployment of production models and to efficiently address multiple different deployment designs such as Batch, Function-as-a-Service (FaaS), and Scoring at the Edge (for example, embedded in devices).

TECHNICAL STUFF

FaaS is an event-driven computing execution model that runs in stateless containers and those functions manage server-side logic and state through the use of services. Scoring at the edge means that, instead of being processed in algorithms located in the cloud, data is processed locally in algorithms stored on a hardware device, for example in a sensor.

This flexibility in model packaging allows developers to build, run, and manage those application packages as functions without having to maintain their own infrastructure. Don´t forget that enterprise level deployments need high availability, autoscaling, and strong authentication features.

Furthermore, you need to consider how your serving infrastructure shall be set up. Model serving makes a trained model available to other software components. Models can be built on different technology stacks for different purposes. Assuming you already have a ML environment set up, you might have deep learning stuff on TensorFlow, and some natural language processing (NLP) built on Keras. In that situation it's important to consider what setup you're adding to or moving from. Other considerations include whether you need to perform model serving in real-time or in batch. This includes model training approaches such as requested/manually, recurring/batch, or online.

Another important technical element of production ML is model monitoring. Monitoring needs to be done at various stages of the lifecycle. Some model monitoring aspects are:

- Check input data flows and quality
- Check output distribution of model results
- Look for model skew, drift, and accuracy change
- Manage and custom model thresholds
- Measure towards thresholds and trigger alarms (as needed)
- Summarize and communicate model performance results

Key considerations for deep learning

Many enterprise data science teams are using different machine learning applications for model exploration and training, including the creation of deep learning models using Tensorflow, PyTorch, and more. However, training a deep learning model is often a time-consuming process, thus GPU and distributed model training approaches are employed to accelerate the training speed. You also might have hybrid setups where you train in GPU and serve on CPU for lower cost.

Deep learning models are generally trained using the stochastic gradient descendent (SGD) algorithm. For each iteration of SGD, you need to sample a mini-batch from the training set, feed it into the training model, calculate the gradient of the loss function of the observed values and the real values, and update the model parameters (or weights). Because SGD iterations have to be executed sequentially, it's very challenging to speed up the training process by parallelizing iterations, making training of deep learning models very time consuming and thereby more expensive.

Detailing ML governance considerations

Achieving and maintaining regulatory compliance (for example, GDPR and CCPA) not only prevents potential financial penalties and damage to your reputation, but also increases trust as well as insight and value from your ML investment. You need a modern data architecture that decreases business and security risks from ever-changing regulatory requirements.

The most valuable and transformative business use cases require multiple analytics workloads to run against the same diverse data sets on fluid multi- and hybrid-cloud infrastructures. If you have the objective of having a powerful ML governance structure for complete security, data and model governance and control across infrastructures in your organization, that still provides the ultimate deployment choice and flexibility in your ML workflow, there are a couple of considerations to take into account.

- **Enterprise-strength security:** It´s important to be able to standardize and enforce granular, dynamic, role- and attribute-based security policies on your platform. You should be able to prevent and audit unauthorized access to

sensitive or restricted data across the platform. Encrypting data across the stack for data at rest, as well as for data in motion, and to be able to manage encryption keys.

- **Governance and compliance:** You should be able to identify and manage sensitive data and effectively address compliance requirements with unified data management operations. This includes metadata search, data lineage, and data chain of custody, but also capabilities for auditing and security as well as data classification, data profiling, and a business glossary.
- **Data migration and replication:** Make sure you can migrate data and associated metadata, complete with security and governance policies, between environments, including legacy clusters, to deliver it to where the enterprise needs to work. Replicate data and metadata to provide backup and disaster recovery services between any environment, from on-premises to hybrid- and multi-cloud.
- **Trusted data catalog:** You should enable self-service access to trusted data and analytics. Let users find and curate assets and collections from a shared catalog that puts all information in context, spanning on-premises, cloud object stores, structured, unstructured, and semi-structured.
- **Clear, functional interfaces:** You should strive for seamless interfaces in key platform services that drive productivity and simplicity. Look for a management console that provides a single pane of glass to create and maintain platform users, roles, and access. It's desirable to have data lake services for safe, secure, and governed data lakes wherever data is stored, from object stores to HDFS.
- **Enterprise schema registry:** You should be able to capture and store any and all schema as well as metadata definitions automatically as they're discovered and created by platform workloads. Governance and data log capabilities should turn metadata into information assets, improving its usability, trust, and value throughout its lifecycle.

Choosing the Right Technology Platform

As has been stressed throughout this book, it's important to consider how you're planning to operationalize your machine learning solutions for you to make the best technology choice. For

example, are your machine learning models only going to run internally, or will they be part of your companies' commercial solutions? Will your models be part of real-time business critical processes? Will your models require live data feeds that are not broken, and with a reliable data quality? All these questions (and more) are examples of important aspects that need to be understood in the context of your organization and the business that you're conducting before making your technology choice.

REMEMBER

The technology platform choice is not only important to manage your models in production, but it's vital for your end-to-end approach and setup. Getting an efficient and iterative machine learning lifecycle up and running is vital from an ROI perspective, but also to manage ML model quality, security, scalability, and innovation.

Listing important features

This section gives you some important user scenarios that you want to make sure are possible to realize on your platform.

- **ML workspaces:** The ability to let administrators deploy new machine learning workspaces for teams in a few clicks, giving data science teams access to the project environments and resources they need for end-to-end ML without waiting is important for your ML efficiency.
- **Secure self-service data access:** You want to make sure that administrators can easily replicate governed data sets across hybrid and multi-cloud environments to give data science teams self-service access to the business data they need while maintaining enterprise data security and governance controls.
- **Running open source applications on the platform:** Beyond Python, R, and Scala for Spark, modern data science teams need the latest open source tools and libraries for innovation and to collaborate while working in their preferred integrated development environment (IDE). It's important to give data practitioners the freedom to use their favorite tools while preserving security, efficiency, and scalability without administrative overhead.
- **Elastic and auto-scaling resources:** You should strive for a platform that enables elastic applications and infrastructure that can be summoned on demand when traffic or workloads get high. Innovation can be unpredictable but should

be supported. By giving data science teams access to the scale-out, heterogeneous computing resources they need to get work done fast while maintaining adjustable guardrails that help IT easily manage and optimize infrastructure resources and costs.

- **Comprehensive user experience:** Machine learning can't begin until data is ready, and it doesn't end when a model is trained. ML for business requires data engineering, model training, and experiment tracking, and deploying and managing models in production. By offering your teams the tools for it all in one cohesive environment without switching or stitching, you ensure higher end-user satisfaction and increased data science team efficiency throughout.
- **A portable and consistent platform:** In a hybrid or even multi-cloud world, you should be able to demand that your ML platform is portable. The business should be able to move the data and infrastructure anywhere without creating disconnected silos and without changing the consistent user-experience that data science teams rely on for building robust workflows and processes for end-to-end ML.

Introducing the Cloudera Machine Learning Platform

Cloudera Machine Learning is focused on reducing time to value for production ML models. It enables data scientists, ML engineers, and operators to collaborate in a single unified platform that is purpose-built for agile experimentation and production ML workflows with enterprise-grade governance capabilities built in.


REMEMBER

As opposed to ML point solutions, a unified data platform doesn't compromise security or require complex, costly workflows for production models. Instead you get an end-to-end ML platform that enables standards-driven model and feature monitoring, cataloging, and ongoing governance at enterprise scale.

Figure 4-1 shows a high-level view of the architecture for the technology platform from Cloudera including the data platform and the machine learning layer. The following sections discuss this technology platform, starting with the Cloudera Data Platform (CDP).

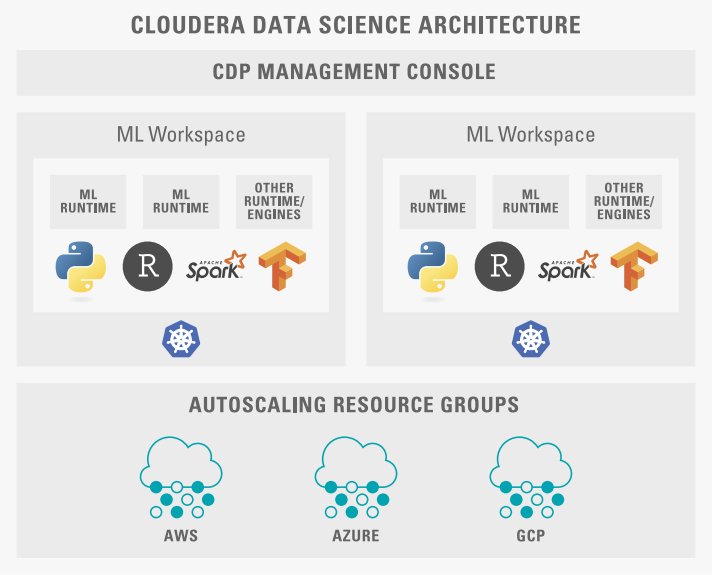

FIGURE 4-1: Architecture overview of Cloudera Data Science architecture.

Describing the Cloudera Data Platform

Cloudera Data Platform (CDP) is an enterprise data cloud, which functions as a platform for both IT and the business. It incorporates support for an environment running both on on-premises and in a public cloud setup. CDP also has multi-cloud and multi-function capabilities at the same time as it's both simple to use and secure by design. It supports both manual and automated functions and is open and extensible. It offers a common environment for both data engineers and data scientists, supporting data science team collaboration. Figure 4-2 shows a capability overview of CDP.

The data platform from Cloudera provides self-service access to integrated, multi-function analytics on centrally managed and secured business data while deploying a consistent experience anywhere — on-premises or in hybrid and multi-cloud. This includes consistent data security, governance, lineage, and control, while deploying the efficient, easy-to-use cloud analytics, eliminating end-user need for shadow IT solutions.

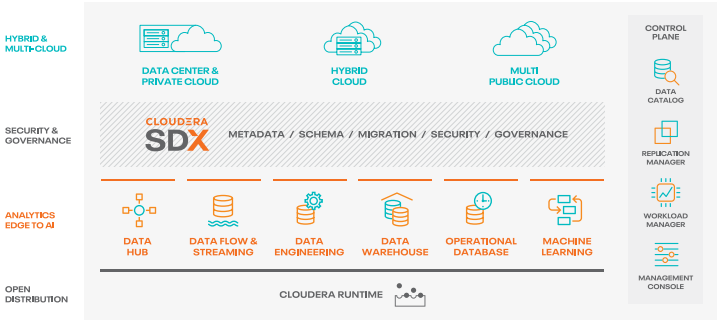

FIGURE 4-2: Cloudera Data Platform (CDP) overview.

To optimize the data lifecycle CDP has multi-function analytics capabilities that integrate data management and analytic experiences across the data lifecycle. You can collect, enrich, report, serve, and model enterprise data for any business use case in any cloud.

Because high value, data-driven business use cases require modern, streaming real-time data, CDP has integrated analytics and machine learning services that are both easy for IT to manage and deploy and easy for business users to consume and operationalize. CDP makes it easy to deploy modern, self-service analytics and machine learning services for any data, with shared security and governance and the flexibility to scale with the same experience. CDP also delivers security, compliance, migration, and metadata management across all environments.

TIP

CDP is an open platform where you can add and build your solutions using open source components including open integrations. It's also open to multiple data stores and compute architectures.

In addition to multi-cloud and hybrid cloud architectures, Cloudera has also integrated data center functionality. The CDP data center is an on-premises capability that is facilitating the realization of an effective enterprise data strategy through data-driven insights. This data center is also a foundational element of the data platform hybrid cloud deployment architectures, as visualized in Figure 4-3. The data center allows a complete platform overview through the data management console. This in turn enables the following capabilities:

- A holistic view of data and metadata.
- A common data catalog across all your deployments worldwide in various data centers and clouds.
- Synchronization of data sets and metadata policies between infrastructures as needed.
- Bursting on-premises workloads into the cloud when more capacity is needed.
- Analyzing and optimizing workloads regardless of where the workloads run.

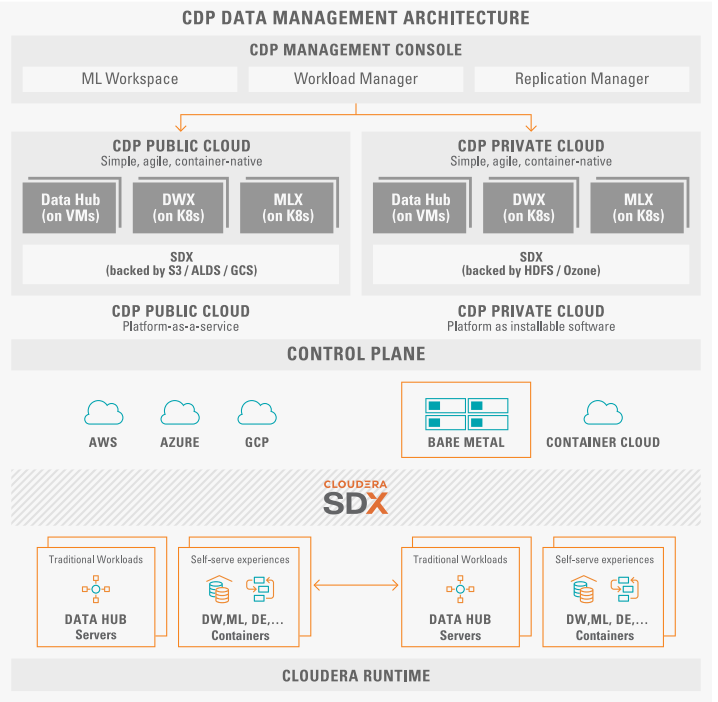

FIGURE 4-3: Data management architecture as part of Cloudera Data Platform.

Detailing Cloudera Machine Learning

On top of the data platform you can run Cloudera Machine Learning (CML); refer to Figure 4-1. CML has extensive MLOps features and a set of model and lifecycle management capabilities to enable

the repeatable, transparent, and governed approaches necessary for scaling model deployments and ML use cases. It´s built to support open source standards and is fully integrated with Cloudera Data Platform, enabling customers to integrate into existing and future tooling while not being locked into a single vendor.

REMEMBER

Cloudera Machine Learning (CML) enables enterprises to proactively monitor technical metrics such as service level agreements (SLA) adherence, uptime, and resource use as well as prediction metrics including model distribution, drift, and skew from a single governance interface. Users can set custom alerts and eliminate the model "black box" effect with native tools for visualizing model lineage, performance, and trends.

Some of the benefits with CML include:

- Model cataloging and lineage capabilities to allow visibility into the entire ML lifecycle, which eliminates silos and blind spots for full lifecycle transparency, explainability, and accountability.
- Full end-to-end machine learning lifecycle management that includes everything required to securely deploy machine learning models to production, ensure accuracy, and scale use cases.
- An extensive model monitoring service designed to track and monitor both technical aspects and accuracy of predictions in a repeatable, secure, and scalable way.
- New MLOps features for monitoring the functional and business performance of machine learning models such as detecting model performance and drift over time with native storage and access to custom and arbitrary model metrics; measuring and tracking individual prediction accuracy, ensuring models are compliant and performing optimally.
- The ability to track, manage, and understand large numbers of ML models deployed across the enterprise with model cataloging, full lifecycle lineage, and custom metadata in Apache Atlas.
- The ability to view the lineage of data tied to the models built and deployed in a single system to help manage and govern the ML lifecycle.

- Increased model security for Model REST endpoints, which allows models to be served in a CML production environment without compromising security.

Furthermore, all enterprise production ML workflows are securely contained in CDP, Cloudera's enterprise data cloud. This enables seamless workflows for governing and quickly customizing models in production while maintaining complete visibility into the end-to-end data and model lineage. Clients can effectively maintain hundreds or thousands of models in production with resources that auto-scale to business needs and set model governance rules that enable fast response to mission critical changes in their production environments.

TIP

Governing production ML workflows in CML enables enterprises to accelerate time to value and deliver ongoing results securely. CML is able to integrate data management with explainable, interoperable, and reproducible MLOps workflows in various execution environments, from edge to cloud.

IN THIS CHAPTER

- United Overseas Bank
- Santander
- Deutsche Telekom

Chapter 5
Production ML Case Studies

In this chapter, we outline how Cloudera has helped three companies with their production machine learning.

United Overseas Bank

United Overseas Bank (UOB) is a leading full-service bank in Asia with a network of more than 500 offices in 19 countries and territories in Asia Pacific, Western Europe, and North America. As UOB's big data team set its vision to build an enterprise-wide big data platform that spans all the bank's business units and regions, it recognized the immense implementation challenges it would likely face. Cloudera enables UOB to realize their artificial intelligence and data science roadmap to drive the adoption of AI initiatives across 12 business units and approximately 2 PB of business data. Using Cloudera Machine Learning for the datacenter as the standard operating platform for data scientists to collaborate, the delivery of AI and data science models is made faster and more efficient. Since adopting Cloudera, UOB has successfully implemented AI initiatives from using natural language

processing to report on market trends, to personalizing features and services to engage customers more meaningfully.

Challenge

UOB set up its Big Data Analytics Centre in 2017, Singapore's first centralised big data unit within a bank, to deepen the Bank's data analytics capabilities and to use data insights to enhance the Bank's performance.

Essential to this work was implementing a platform that could cost-effectively bring together data from dozens of separate systems and incorporate a range of unstructured data, including voice and text. "With legacy databases, you're restricted by the amount of data as well as the variety," said Richard Lowe, Chief Data Officer at United Overseas Bank. "As a result, you can miss key data attributes that are necessary for anti-money laundering (AML), transaction monitoring, and customer analytics engines to work effectively."

Outcomes

With new self-service analytics and machine-learning driven insights, UOB has realized improvements in digital banking, asset management, compliance, AML, and more. This includes a new recommendation engine for increased conversion rates with the ability to understand lifestyle preferences and deliver personalized offers and recommendations — for everything from dining to shopping — to millions of customers and merchants. Customer analytics insights enable corporate relationship managers to better understand global client networks and identify new revenue opportunities. "This project not only led to a very large uplift in leads, it also saved relationship managers more than 1,000 hours in manually reviewing documents."

In addition to this, advanced AML detection capabilities help analysts detect suspicious transactions earlier based on hidden relationships of shell companies and high-risk individuals.

"You don't always know what data will be valuable for a particular use case or what use cases lie ahead," says Lowe. "Cloudera enables us to innovate, pursue new capabilities, and achieve outcomes that wouldn't be possible otherwise."

Santander

Originating in Spain, Santander is a multinational commercial bank and financial services company serving thousands of customers across the world. The company maintains a presence in all global financial centers as the 16th-largest banking institution in the world. With growing needs for streamlined machine learning on increasing qualities of data, Santander used Cloudera to implement a single data platform for every stage of their data and ML lifecycle. Cloudera enabled Santander to support all its workloads, including self-service, operational, machine learning across their business quickly and effectively. And enabling previously unattainable scale by processing over 10 million transactions daily across nearly 2 PB of data.

Challenge

Santander was looking to consolidate and streamline operations for analytics and machine learning across a large number of legacy data warehouses spread across its many business units. Because of data silos, working silos, and slow legacy systems, Santander could not get comprehensive customer insights to support their broader operations and business teams.

Outcomes

By implementing Cloudera as their core solution for data lifecycle management and machine learning, Santander was able to reduce capital spending by $3.2M through real-time analysis of thousands of new corporate customer prospects. By providing comprehensive insights across the business, Santander was able to identify 7,000+ new client prospects while decreasing operating expenses by $650,000, and realizing $2.4M in annual savings for their marketing department.

Deutsche Telekom

Deutsche Telekom is one of the world's leading integrated telecommunications companies, with some 184 million mobile customers, 27.5 million fixed-network lines, and 21 million broadband lines. By using Cloudera to eliminate silos and enable

unprecedented scale when applying ML and AI, the company is able to identify network problems by detecting fraud patterns and real-time threats before the business is affected.

Challenge

Deutsche Telekom needed to effectively address inefficiencies and tackle fraud across their global business. At this scale, preventing network fraud is a major challenge. Deutsche Telekom could not manage the huge amounts of data captured in silos across the business, making machine learning at scale impossible.

Outcomes

Deutsche Telekom was able to realize a 20 percent reduction in revenue loss through improved fraud detection, CRM, network quality, and operational efficiency with Cloudera. Cloudera enabled the enterprise to streamline their data management operations and enable fast, visible, and secure machine learning operations across the business. In addition to the revenue loss reduction, Deutsche Telekom was able to better serve its customers, reducing churn by 5–10 percent and improved operational efficiencies by 50 percent overall.

IN THIS CHAPTER

- Taking a holistic and experimental approach
- Securing a multi-disciplined team
- Addressing the architecture aspects
- Embracing ML in the organization

Chapter 6
Ten Steps How to Make ML Operational

Each *For Dummies* book ends with a Part of Tens chapter. This book is no different, so here I give you ten steps towards making ML operational.

- **Take a holistic approach to machine learning.** Machine learning shall be viewed holistically as an integral part of your company strategy. By integrating it and running it alongside your existing IT environments, processes, applications, and workflows, you drive greater results.
- **Be willing to experiment and, yes, fail.** Machine learning models and the algorithms behind them are by nature about science, not business results. Only their application can drive business results, and you should approach the business problem you're trying to solve as an experiment.
- **Build a multi-disciplined team and don't box them in.** Does the platform you're considering give your team practical access to the data, compute resources, and libraries they need? Can your team collaborate efficiently across disciplines, and can the platform enable this access and collaboration in a way that's governed and secure?
- **Iterate quickly; optimize later.** Let your data scientists select and use the tools and frameworks they want. They should

have the freedom to iterate quickly and build models that can be optimized later. Don't worry about getting a model that's flawless the first time through; you may spend too much time trying to perfect a model only to learn that the solution or enhancement you were hoping for wasn't actually achievable through machine learning. Let your team experiment rapidly, fail early and often, continuously learn, and try new things.

- **Choose the right technology to optimize your ML lifecycle.** Pick a platform that prioritizes holistic collaboration and streamlines your ML workflows from data to production in a secure, interpretable, and scalable way. Be wary of point solutions or "black box" ML platforms that create silos and compromise security.
- **Embrace machine learning by evolving your organization.** There's a wall that seems to exist between experimentation and large-scale production. Many organizations hit this wall because they don't know how to weave machine learning development, production, and maintenance into their existing processes, workflows, architecture, and culture. To succeed you must start making the structure of your organization more flexible so that machine learning can be embraced.
- **Maintain the integrity of your models.** As your underlying data changes and shifts, and as your models themselves have an impact on the data, the models using that data have to be updated and improved upon. Maintaining the integrity of your models demands vigilance. Otherwise, your models may drift and become inaccurate and ultimately impact your business.
- **Close the skills gap.** Try to build a team whose experience, talents, and capabilities including data engineering, data science, software development, DevOps, product development, and domain expertise, overlap. Look for candidates with the core skills that are necessary to accomplish your most important tasks, and then get them together and let them learn from one another.
- **Treat models in production like living software.** To protect your models in production, it's important that you have the ability to keep them secure. This means having the visibility into model lineage and the control over who can access and make changes to your models.
- **Understand and abide by your ethical obligations.** Make sure you truly have consent from customers and other stakeholders to apply the necessary data against a machine learning model.

Drive enterprise machine learning success

Machine learning makes it possible for teams to work smarter, accomplish things faster, and turn previously impossible tasks into routine tasks. But it takes the right approach throughout your enterprise, taking into account the people, process, and technology perspectives necessary for successful AI projects. *Production Machine Learning For Dummies,* Cloudera Special Edition shows how you can successfully create, sustain, and apply a production ML approach at scale in your enterprise.

Inside...

- Adopt ML production models across your company
- Identify challenges with ML
- Choose the right ML platform
- Cultivate a ML production mindset
- Learn how to sustain revenue with ML

CLOUDERA

Ulrika Jägare is Head of AI at Ericsson North America. With a decade of experience in data, analytics, and AI/ML as well as 20 years in telecommunications, she has held numerous data-related leadership positions across Ericsson. She is the author of *Data Science Strategy For Dummies.*

ISBN: 978-1-119-73530-4
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.