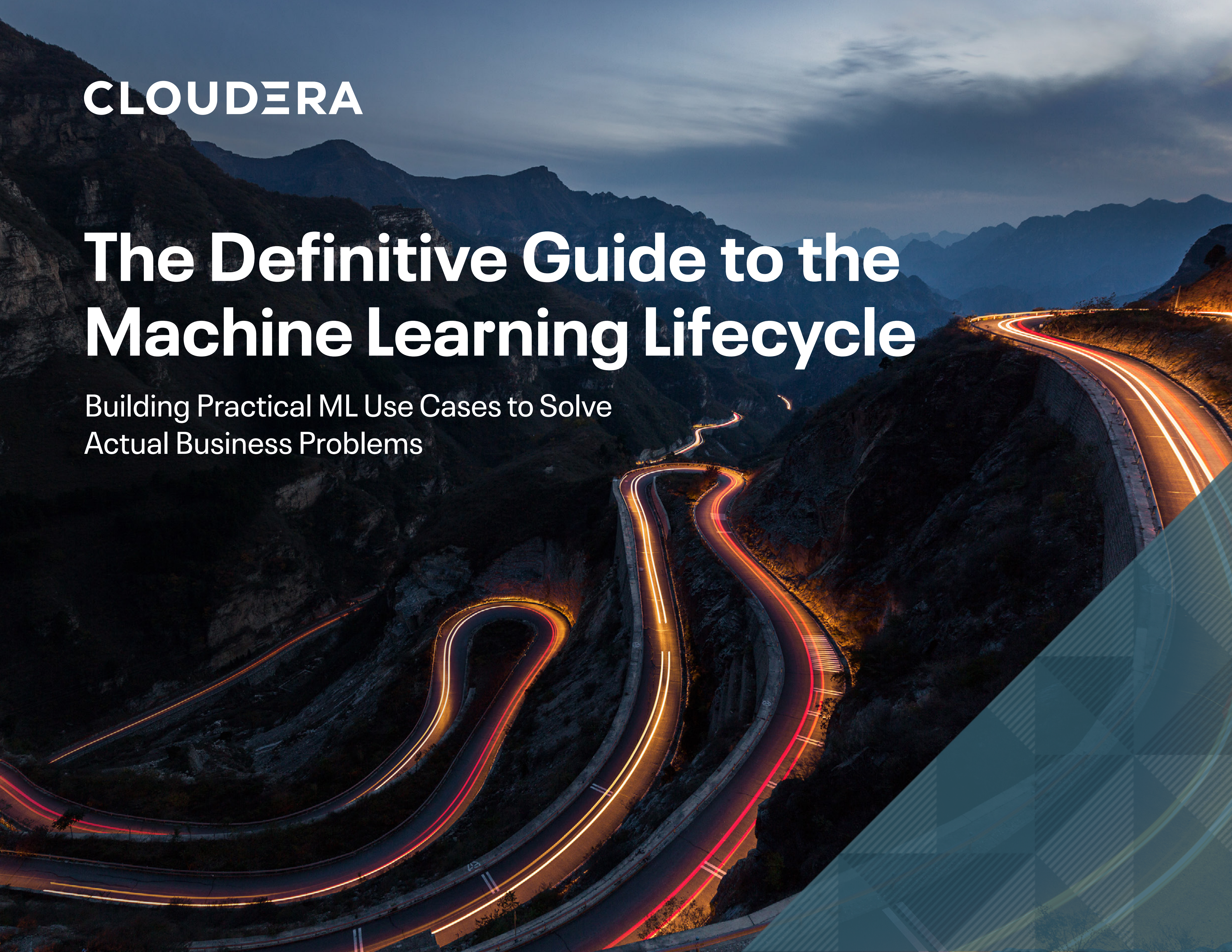
CLOUDERA



The Definitive Guide to the Machine Learning Lifecycle

Building Practical ML Use Cases to Solve Actual Business Problems

Table of Contents

Building ML Use Cases that Matter to the Business

92% of organizations are accelerating the pace of AI investment—but only **29%** are achieving transformational business outcomes.[1]

Organizations across every industry recognize the potential business value of AI, whether it's improving customer engagement or providing greater healthcare. Unfortunately, for the vast majority, the key to successfully building and implementing machine learning (ML) use cases that yield tangible results still remains elusive. Almost 80% of all AI and ML projects stall out due to problems with data quality, labeling, and building trusted models.[2] Additionally, 92% of organizations say cultural barriers prevent them from capitalizing on the business benefits ML has to offer.[1]

One clear reason many organizations come up short with AI is because of an overemphasis on building ML models. But to find success with AI, the models that power ML use cases can neither be created in a vacuum nor controlled by one. Those who insist on taking such a siloed approach to the ML lifecycle are likely to find themselves going nowhere fast. After all, 88% of enterprise ML projects never make it beyond the experimental stage.[3]

This is due to:

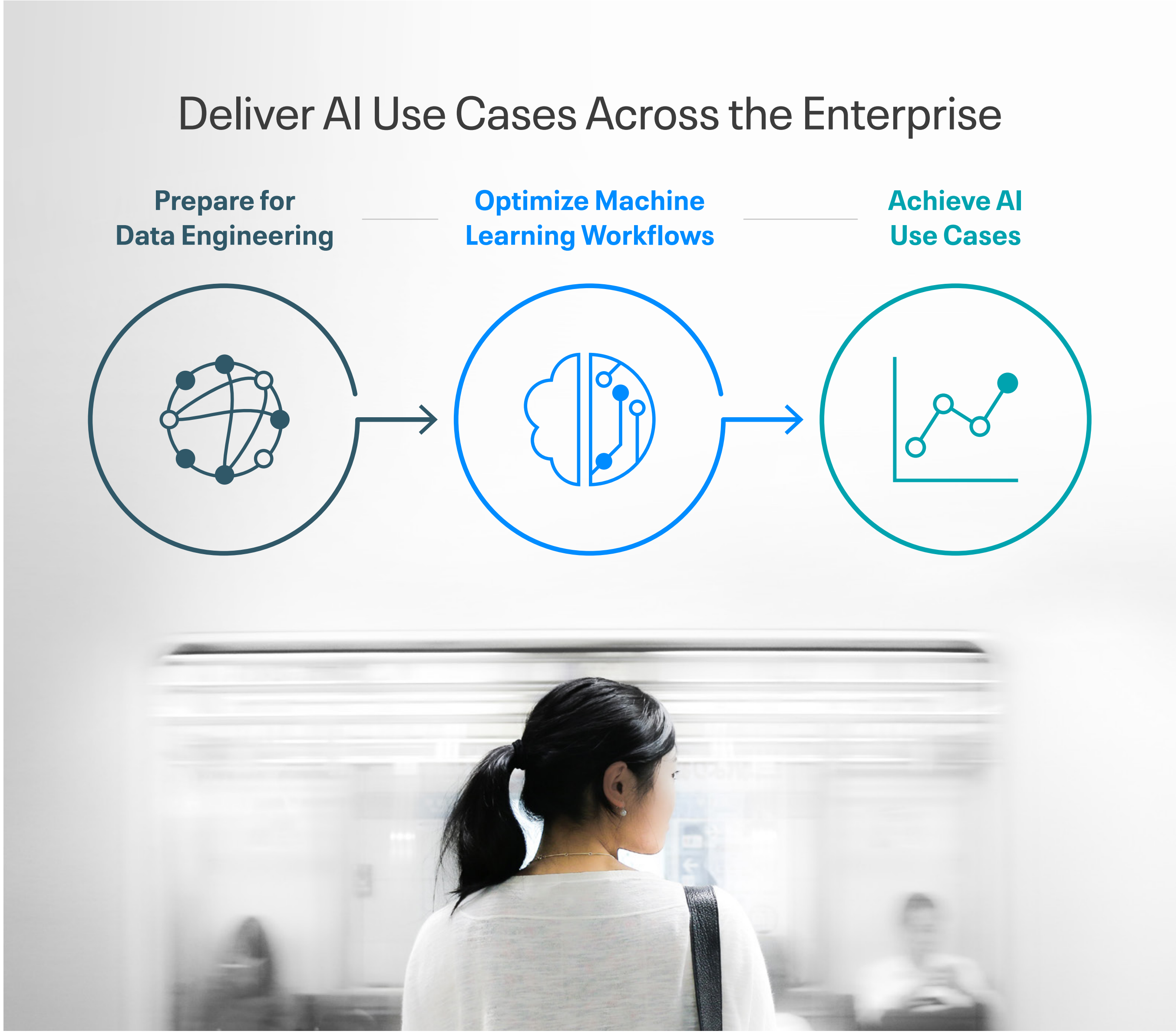
- A lack of proper data control and governance
- Technical limitations
- Accessibility and model interpretation problems
- Organizational issues
- An inability to apply ML at the necessary enterprise scale to power AI use cases in a meaningful way

88%

of enterprise ML projects never make it beyond the experimental stage.[3]

Instead, to build ML use cases for your organization that deliver results, you need a holistic approach that takes into account the full machine learning lifecycle — what comes before, what happens during, and what happens after ML modeling—from the ingestion of raw data all the way to how your ML models are monitored and maintained over time.

This guide reveals how to power more trusted and explainable ML use cases across your organization. Keep reading to learn what you can do to take control of the ML lifecycle—so that you can scale ML use cases that matter to your business.



Deliver AI Use Cases Across the Enterprise

Prepare for Data Engineering

Optimize Machine Learning Workflows

Achieve AI Use Cases

Chapter 1: It All Comes Down to Data

By 2022, 90% of organizations will explicitly mention information as a critical enterprise asset and analytics as an essential competency.[4]

Data is your most valuable asset. So, before you can even think about building and deploying ML models, your first course of action is to make sure you have the right end-to-end data management system in place. How well you continuously govern, monitor, and manage data across your entire organization will play a major role in the success and sustainability of your ML initiatives.

While automation, business predictions, and product innovations are only a few examples of what ML can achieve, those goals are only attainable by creating and maintaining ML algorithms—and an algorithm can only be as accurate as the data that shapes and feeds it.

Many organizations, however, now find themselves unable to effectively analyze a greater amount of data from a rising number of sources. Diverse data sets now live across a fragmented IT landscape consisting of edge, data center (private cloud), and public cloud environments. And if you're like most organizations today, you have approximately two public and private clouds in use and are increasingly tasked with ingesting and analyzing more data at the edge in real time.[5]

If you want to unlock the value of data in this landscape, you need to take full control of the data lifecycle throughout every environment and system along the way—from edge to AI.

"AI techniques are becoming a standard part of the typical business analysis and business intelligence that organizations do. There's consistent movement toward more and more advanced analytics and interesting insights that can be made from data today.

-Sushil Thomas
VP of Engineering, Machine Learning, Cloudera

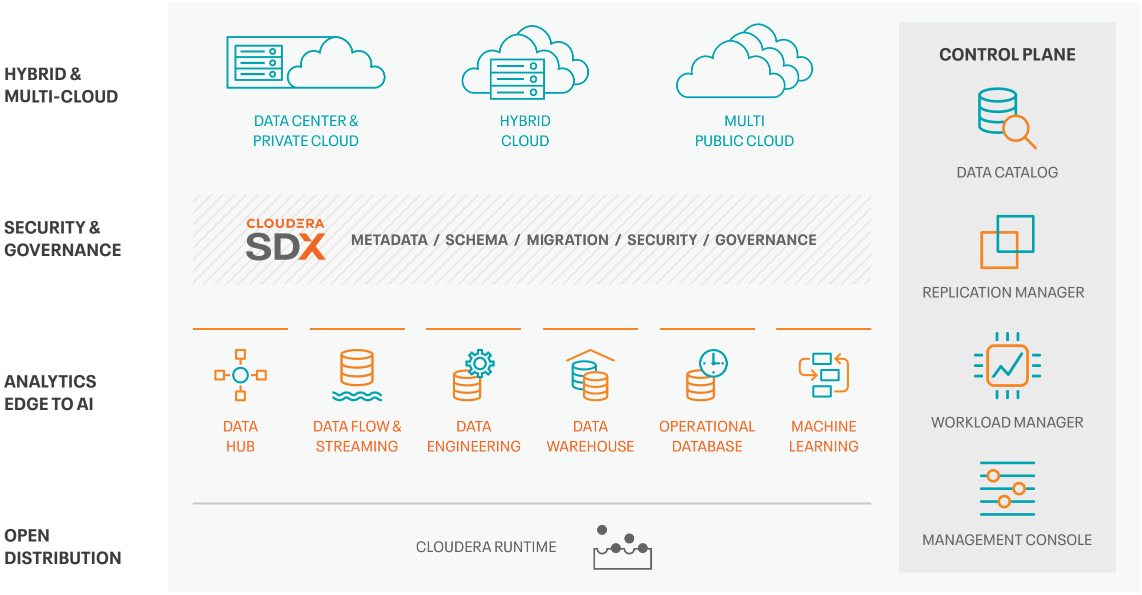
Demolishing your data silos

The people in your organization need self-service access to data for ML use cases and a variety of other reasons. Data engineers rely on raw data to create high-quality, production-ready pipelines, as well as to power analytics workstreams. Enterprise data science teams need access to business data and the tools and computing resources required for ML workflows. And business users need access to critical data for analytics and AI-driven insights for better decision making.

But there's a not-so-slight problem: To grant access to data, IT teams traditionally must create and manage multiple copies of the same data sets and stay ahead of the barrage of access requests. These are time-consuming tasks that are both manual and costly. And the more these IT teams struggle to maintain security, governance, and control across a complex IT landscape, the more likely they are to restrict access to data in an effort to ensure compliance. As a result, an organization's ability to implement ML use cases at all—let alone at enterprise scale—is greatly impacted.

If you can eliminate silos, then your data scientists, data engineers, and business users are one step closer to realizing the full analytical potential of your data across the complete data lifecycle for ML use cases. However, whatever method you choose to remove those silos must also ensure that your data can be managed holistically across the IT landscape, so that IT teams can efficiently enforce governance, security, and control.

Take Control with an Enterprise Data Cloud



HYBRID & MULTI-CLOUD

DATA CENTER & PRIVATE CLOUD

HYBRID CLOUD

MULTI PUBLIC CLOUD

SECURITY & GOVERNANCE

CLOUDERA SDX

METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE

ANALYTICS EDGE TO AI

DATA HUB

DATA FLOW & STREAMING

DATA ENGINEERING

DATA WAREHOUSE

OPERATIONAL DATABASE

MACHINE LEARNING

OPEN DISTRIBUTION

CLOUDERA RUNTIME

CONTROL PLANE

DATA CATALOG

REPLICATION MANAGER

WORKLOAD MANAGER

MANAGEMENT CONSOLE

An enterprise data cloud is a big data platform that helps you manage the deluge of data across edge, private cloud, and public cloud environments, as well as throughout the data lifecycle.

By implementing an enterprise data cloud, you can adopt the most valuable and transformative business and ML use cases by realizing the full analytical potential of all your siloed data.

Cloudera Data Platform (CDP) is a first-of-its-kind enterprise data cloud. It uses an open, hybrid data architecture to power data-driven decisions and predictive actions by seamlessly connecting your data across your fragmented IT landscape.

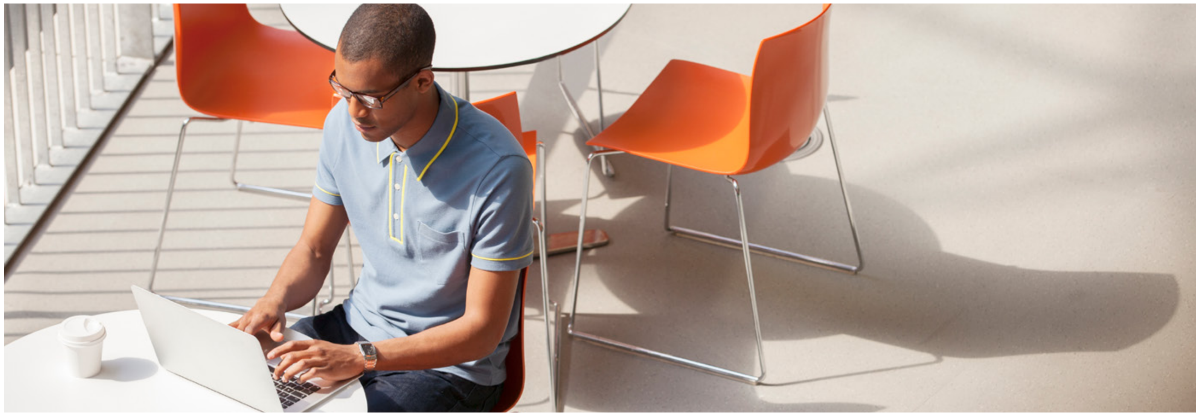
With CDP, data engineers, data scientists, and business users get quick, easy access to critical data, while IT teams more easily secure and govern the entire data lifecycle through a single pane of glass. With CDP, you can:

- Run multiple analytics and ML models against the same diverse data sets for AI use cases
- Eliminate complexities involved with managing petabytes of data and multiple workloads across hybrid environments
- Ensure that data workloads run at their highest possible performance and cost efficiency

Processing and Curating Large Data Volumes in Near Real Time—at Scale

Controlling and managing your data efficiently is only a single piece of the ML lifecycle puzzle. Another crucial element in building ML use cases is the processing and curating of large data volumes at incredible speeds—and doing so at scale.

Today, the vast majority of data engineers look to Apache Spark™ as their framework of choice for extract/transform/load (ETL) jobs to process large data volumes in near real time. Spark accelerates the ingestion, exploration, modeling, curating, and cataloging of data types from multiple sources, enabling users to build batch or streaming pipelines with relative ease. Data engineers also often rely on Apache Airflow for the curation and orchestration of complex data used by ML models and other workflows. Essentially, Spark and Airflow are responsible for the quality and quantity of the data that continuously feeds the ML models that in turn power business use cases.



But for all of its processing prowess, Spark requires significant manual work under the hood. ETL jobs inherently have resource-intensive and time-consuming requirements that can impact any ML workflows down the line. When a data pipeline is ready for deployment, data engineers have to provision it with adequate resources, account for the job in capacity planning, and schedule it. And even after deployment, data engineers have to ensure that the right dependencies are carried over into production, and then continuously monitor the job for problems.

On top of that, having to actually debug or performance-tune the job will only result in lost time, wasted resources, and additional headaches. For example, data engineers have to manually gather and scrutinize all of the necessary logs in hopes of finding a bottleneck or underlying issue. And when it comes time to upgrade to the latest version of Spark, the entire cluster—that could stretch across any number of teams—has to come down, bringing work to a screeching halt.

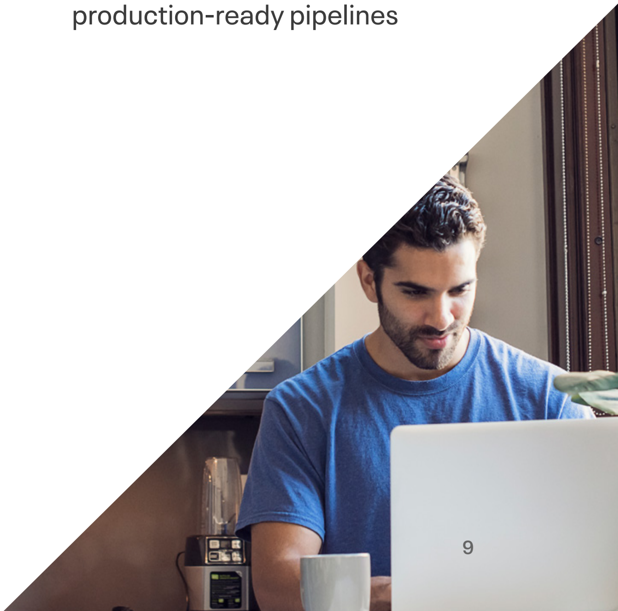
So, while Spark can process large data volumes at incredible speeds, it struggles with effective data engineering in production at scale—and this shortcoming ultimately has a negative impact on your AI initiatives.

How to Leverage Spark and Airflow for ML at Scale

Cloudera Data Engineering (CDE) for CDP lets data engineers use Spark to process large data volumes while streamlining the data pipelines driving your ML workflows and use cases. Data pipeline management is optimized—and data is accelerated from ingest to insight at scale.

With CDE, data engineers can:

- Automatically deliver curated, high-quality data pipelines while also ensuring security and governance compliance
- Control costs with its ability to deploy to any environment and on-demand automatic resource scaling
- Enable agile, self-service data engineering while provisioning and ensuring isolation across users
- Efficiently deploy and monitor the lifecycle of every job to quickly solve issues and continuously maintain production-ready pipelines
- Utilize ML pipeline curation and automation of Apache Airflow
- Scale on demand without disrupting other workloads
- Use APIs to automate the lifecycle management of clusters, applications, and more



Chapter 2: Setting Data Scientists up for Success

Give your data scientists a sandbox to play in, so they can develop models that power your ML use cases—but keep that sandbox safe.

Once you have control over the data lifecycle and your data engineers are able to process and manage large data volumes at scale, it's time to move a bit farther down the ML lifecycle and mobilize your data scientists.

To create predictive and analytics applications from ML models, data science teams need more than immediate and direct access to curated pipelines; they also need the freedom of choice, experimentation, and resource scalability.

One challenge you may encounter early on may be finding the right balance between giving your data scientists the freedom to experiment while simultaneously enforcing appropriate governance and security.

While standardization and control are important, your data scientists have to remain agile and nimble as they experiment. Give them the room to experiment rapidly, fail early and often, learn from their choices, and try new things.

"

From an organizational perspective, the important thing is that your data scientists spend less time being frustrated and spend more time being productive.

–Sushil Thomas
VP of Engineering, Machine Learning, Cloudera

Enabling Self-Service Data Science

When it comes to developing models for ML use cases, data science teams need access to data pipelines, scalable compute resources, and their preferred runtimes, IDEs, libraries, and data science tools of choice.

Something important to note is that whatever platform and tools data scientists use today will lay the groundwork for tomorrow. So, look for a platform that bridges two primary and continuous phases of the ML lifecycle:

PHASE ONE
covers holistic ML development and the building of the ML models by the data scientists.

A formula for success in phase one

You don't want to box in data scientists with a platform that restricts access to data, preferred tools, or hinders collaboration. At the same time, you still need to enforce governance and security. For phase one, make sure there's a platform in place that gives your data science teams practical access to the necessary data pipelines, along with the compute resources and libraries they need. Also, it's important to establish efficient collaboration across disciplines, so that data scientists can communicate effortlessly with data engineers and anyone else involved in the building and maintaining of ML models.

Accelerating ML use cases

Data scientists need strong exploration and visualization skills. They also need a deep understanding of dozens of machine learning techniques; if they're lacking in expertise on a given technique deemed appropriate for a specific use case, then they have to invest valuable time into learning it. Other challenges—such as figuring out the surrounding architecture for productionizing, operationalization, model monitoring, etc.—are also matters to figure out before the development of the ML use case can get under way. How agile and fast you are in addressing these challenges will set the pace for the entire process.

If you want to build an ML use case faster, you need a platform that enables data scientists to jumpstart the building of the use case itself.

With the right platform in place that offers pre-built ML projects complete with pre-canned models and apps, your data scientists have a running start at building ML-powered business use cases. Look for a platform with ML prototypes that show the approach taken for a particular case, as well as demonstrate the use of a specific technique, tool, or library. Your data scientists will see where they're going, know how to get there, and they won't have to travel nearly as far.

PHASE TWO focuses on model production, scaling, and ongoing operations.

Making the jump to phase two

Regarding phase two, the ability to put ML models into large-scale production cannot be overstated. The key to success is a platform with continuous integration tools that let data scientists deploy ML models anywhere those models are required to operate. Also, it should be noted that the bulk of phase two centers around keeping models in production accurate and up to date, which in itself poses a huge challenge.

For many organizations, most ML models never make it into large-scale production because cross-disciplinary teams lack a way to collaborate, share knowledge, and scale models for use. And for ML models that do make into production, the arduous task of maintaining ML model hygiene begins, lest the model drifts and impacts the use case it serves.

Don't just give your data scientists everything they need to experiment. Give them everything they need to build and deploy models into full-scale production—and then monitor and maintain those models once they're deployed. It's the only way your ML use cases will ever get off the ground and yield a positive business impact.

Build, Deploy, and Operate ML Models Faster

Cloudera Machine Learning (CML) for CDP unifies self-service data science and data engineering in a single, portable service for multi-function analytics on data anywhere.

CML also features Applied Machine Learning Prototypes (AMPs), a catalog of end-to-end reference ML projects that you can deploy at the click of a button. There are dozens of high-quality AMPs already available within CML, each one ready to deploy out of the box and into your ML workspace with:

- A templated business use case (including deep learning for image analysis, fraud detection, anomaly detection, and more)
- All the required data
- Instructional, step-by-step guidance and explanations
- An end-to-end workflow, from data to auto-deployed model and application

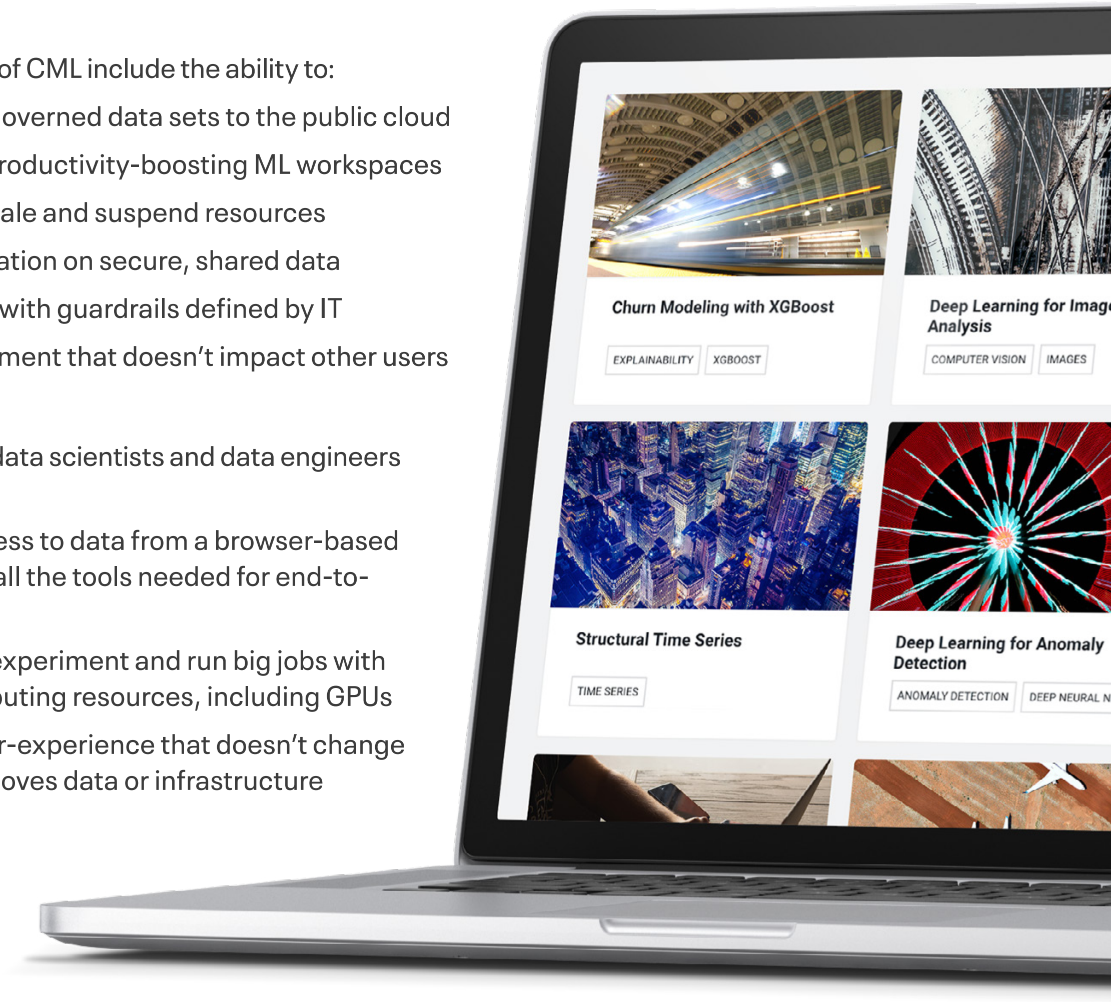
Each AMP also features completely open project code and the fully built solution, so data scientists start building an ML use case with 90% of the project already under their belt.

Other advantages of CML include the ability to:

- Easily replicate governed data sets to the public cloud
- Quickly deploy productivity-boosting ML workspaces
- Automatically scale and suspend resources
- Fast experimentation on secure, shared data
- Elastic compute with guardrails defined by IT
- IT policy enforcement that doesn't impact other users or workloads

Additionally, both data scientists and data engineers benefit from:

- Self-service access to data from a browser-based workspace with all the tools needed for end-to-end ML
- The freedom to experiment and run big jobs with right-sized computing resources, including GPUs
- A consistent user-experience that doesn't change if the business moves data or infrastructure



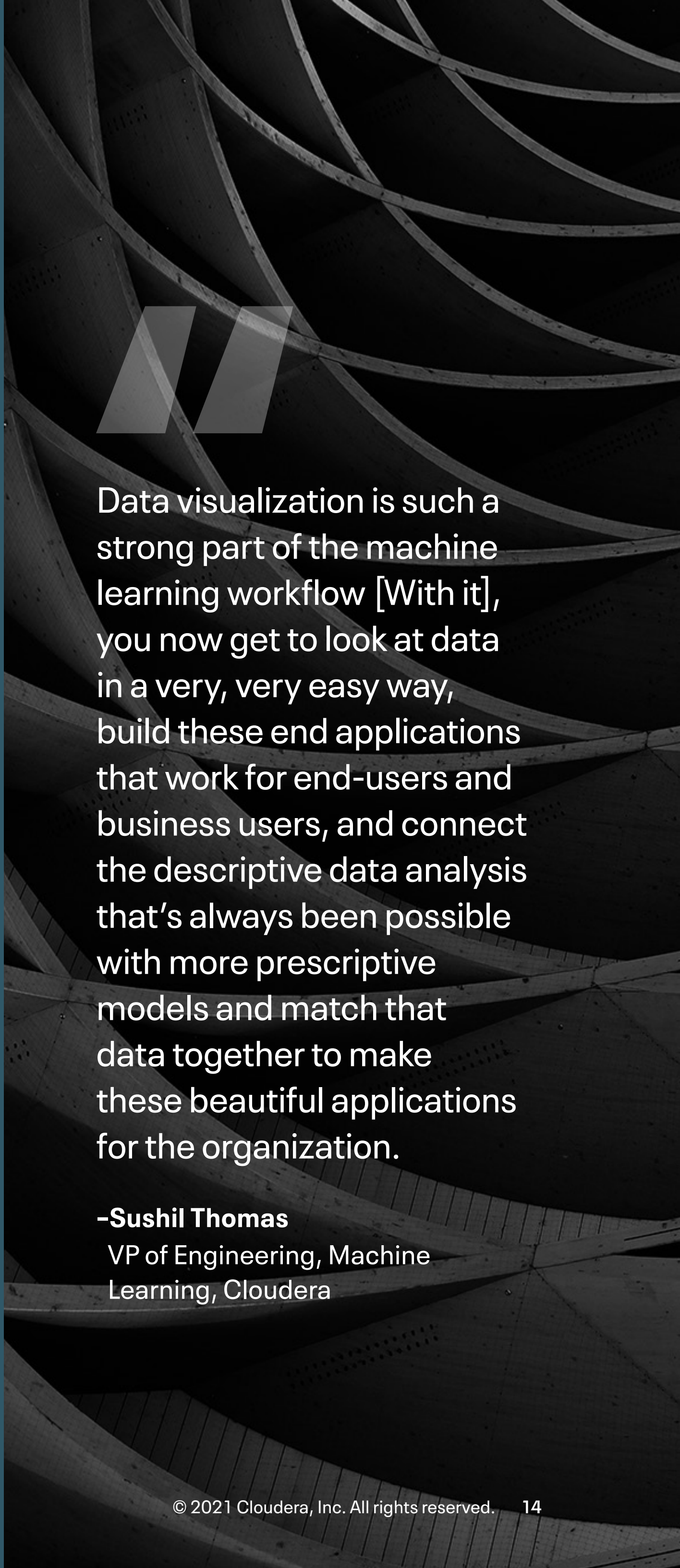
Chapter 3: Showcasing the Business Value of ML

AI use cases should deliver value to the business—but only 20% of analytic insights actually contribute toward meaningful business outcomes.[6]

With data engineers and data scientists empowered to deliver powerful ML models, it’s time to put ML use cases that benefit the business into action. This requires presenting knowledge derived from ML models as explainable, visualized insights. And too often, analytical insights produce no business impact. There are many reasons for this, one of which is that typical data visualization methods involve moving data from a centralized database or data lake to a downstream business intelligence platform or a third-party visualization tool.

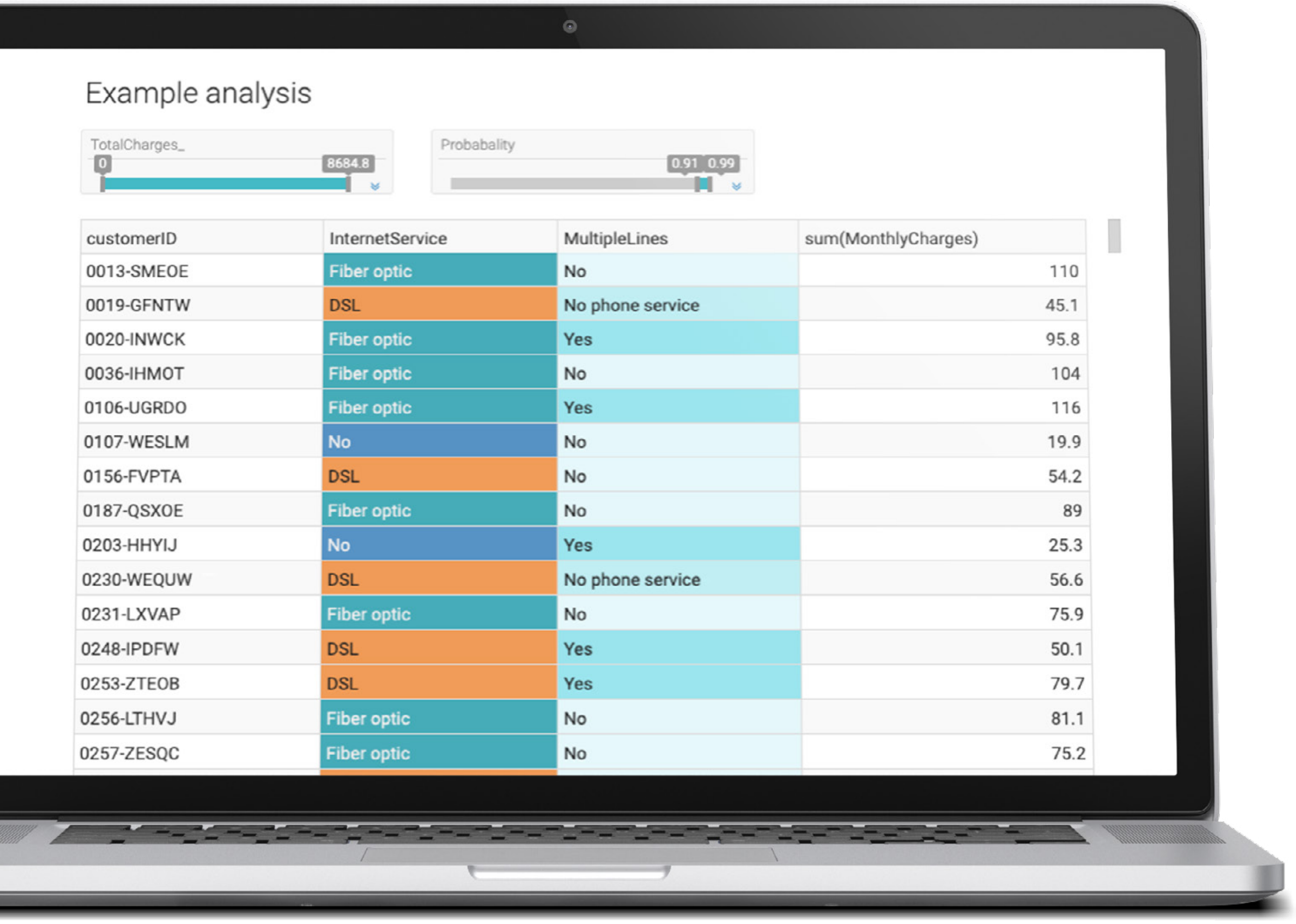
This takes time, requires work and coordination from multiple teams, and creates the additional overhead of securing multiple data copies. All the while, users are unable to see data in real time and share any insights gleaned automatically.

For success with ML use cases, all users across your organization—from the data engineers to the business stakeholders—need the ability to explore data and collaborate with each other to communicate insights. Most importantly, business teams in particular have to trust ML models before they take any action on the predictions those models make. This potential barrier to success is at the homestretch of the ML lifecycle—and it can stop your AI initiatives just shy of crossing the finish line.



> Data visualization is such a strong part of the machine learning workflow [With it], you now get to look at data in a very, very easy way, build these end applications that work for end-users and business users, and connect the descriptive data analysis that’s always been possible with more prescriptive models and match that data together to make these beautiful applications for the organization.
>
> **-Sushil Thomas**
> VP of Engineering, Machine Learning, Cloudera

Delivering Actionable ML Use Cases



Example analysis

TotalCharges_ 0 8684.8

Probabality 0.91 0.99

customerID	InternetService	MultipleLines	sum(MonthlyCharges)
0013-SMEOE	Fiber optic	No	110
0019-GFNTW	DSL	No phone service	45.1
0020-INWCK	Fiber optic	Yes	95.8
0036-IHMOT	Fiber optic	No	104
0106-UGRDO	Fiber optic	Yes	116
0107-WESLM	No	No	19.9
0156-FVPTA	DSL	No	54.2
0187-QSXOE	Fiber optic	No	89
0203-HHYIJ	No	Yes	25.3
0230-WEQUW	DSL	No phone service	56.6
0231-LXVAP	Fiber optic	No	75.9
0248-IPDFW	DSL	Yes	50.1
0253-ZTEOB	DSL	Yes	79.7
0256-LTHVJ	Fiber optic	No	81.1
0257-ZESQC	Fiber optic	No	75.2

Business teams will not act on predictive insights unless they trust the ML model where those insights are derived. Furthermore, business teams want to interact with the data themselves, to see it visualized in a way that's easy to understand and use for decisive action.

The end-goal for the vast majority of ML use cases in the enterprise is to deliver business-critical, explainable insights and AI-powered predictions that business users can count on and readily share across the organization.

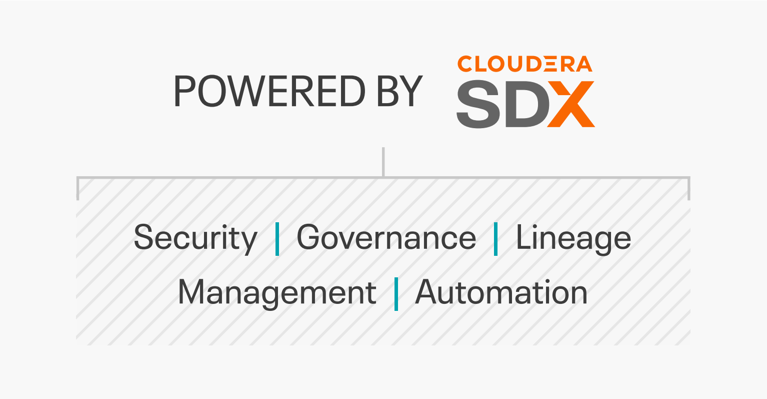
But success in this regard hinges on whether or not you have successfully taken control of the data lifecycle. If high-quality data isn't presented effectively and in a way that's engaging, it can get overlooked. And if models aren't properly monitored and maintained, then accuracy slips and confidence in the insights presented will waver.

Find a way to empower data engineers and data scientists to not only deliver visualized insights that demand attention—but to also protect the integrity of the pipelines and models that bring those visualizations to life.

Protect the Integrity of Your ML Lifecycle

CDP features an always-on Shared Data Experience (SDX) layer that enables holistic security, governance, and compliance across the full data lifecycle. SDX can also be used in CML to deploy models to production with access, governance, and security rules inherited directly from CDP.

Additionally, SDX for models enables full ML lifecycle governance with lineage tracking and auditing capabilities that make finding the origins of data used in the training of any given model easy. This means the data behind your models are explainable and interpretable, fostering greater confidence from business users.


POWERED BY CLOUDERA SDX

Security | Governance | Lineage

Management | Automation

Intuitively See and Share Visualized Data

CDP Data Visualization (DV) lets data engineers, data scientists, and business users quickly and easily explore data, collaborate, and communicate explainable insights everywhere. With CDP DV, there's no data movement between third-party tools. And it lets you connect and visualize data from any data lake, data warehouse, or CDP service, meaning you and other stakeholders can tap into data and visualize it anywhere in the data lifecycle. Likewise, you can use the drag-and-drop interface of CDP DV to build predictive applications based on any living ML models served in CML.

Advantages of CDP DV include:

Fast, intelligent reporting with built-in, AI-powered Natural Language Search and Visual Recommendations

Ease of use in an intuitive, visual UI for instant insight sharing without moving data

Accelerated collaboration with a consistent, integrated data visualization experience across all data and business teams

Make an Impact on Your Business with ML That Scales

Finding success with your ML initiatives first requires taking control over the entire ML lifecycle. And while this guide walked through what it takes to power more trusted and explainable business use cases, it's up to you to take the next step toward making AI use cases that matter to the business.

Learn more about how Cloudera can help you achieve success with ML faster by reading the Capabilities and Approach for AI, at Scale, in the Enterprise Whitepaper.

Sources

[1] Big Data and AI Executive Survey 2021, NewVantage Partners.

[2] Dimensional Research, "Poor data quality causing majority of artificial intelligence projects to stall," Bob Violino. July 2019.

[3] McKinsey, "Artificial Intelligence the Next Digital Frontier?," July 2017.

[4] Gartner, "Why Data and Analytics Are Key to Digital Transformation," Christy Pettey. March 8, 2019.

[5] 2020 State of the Cloud Report, Flexera.

[6] Gartner, "Our Top Data and Analytics Predictions for 2019," Andrew White. Q3 January 2019.

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 cloudera.com | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

Privacy Policy | Terms of Service

CLOUDERA