CLOUDERA

The Ultimate Guide to Enterprise ML Platforms

Table of Contents

Executive Summary

Artificial Intelligence (AI) and Machine Learning (ML) adoption is growing faster than ever. With ML, organizations can leverage large, growing treasure troves of data from numerous sources to streamline operations, drive innovation and build competitive advantages.

However, despite investing millions of dollars in ML initiatives and AI use cases, many are not successful and have a high failure rate. This is because ML requires a vast and complex surrounding infrastructure as ML isn't one single activity but rather an end-to-end iterative workflow with many phases that require owning the full ML lifecycle.

Because of these requirements, it is critical that customers deploy an end-to-end platform to ingest data, create, deploy, and maintain ML models, and make insights actionable by the business. Selecting the ultimate ML platform for your business requires accounting for accelerated performance, security, governance, and continuous monitoring needed to scale production ML use cases. Clients also need to have unprecedented flexibility and choice to deploy this ML platform on-premises, public clouds, or hybrid multi-clouds.

Worldwide, with Cloudera Data Platform, many enterprise organizations are now able to make ML ubiquitous, practical, repeatable, simple, functional, and cost effective.

Implementing AI/ML Can be Simple

Artificial Intelligence (AI) and Machine Learning (ML) are key technologies for organizations who want to benefit from the massive insights buried in their data marts, data warehouses, Apache Hadoop lakes, and spreadsheets. ML enables organizations to make internal processes faster and cheaper, build better products and services, create brand new products, or completely reinvent processes.

Consequently, ML adoption is accelerating fueled by the growing volume and variety of data and the rapid advances in cloud/edge computing and compute/storage performance. However, many ML initiatives and AI use cases are not successful and have a high failure rate.

According to VentureBeat AI[1], a technology publishing company, 87% of data science projects never make it into production. And a global survey by Dimensional Research[2] concluded that 78% of the AI/ML projects stall at some stage before deployment.

Another study[3] indicates that only 17% of all AI initiatives are in production, another 15% are in development, and 17% are in proof of concept. In contrast, over half, or 51%, have failed, indicating an exceptionally high failure rate.

87% of data science projects never make it into production.

Even with millions of dollars invested in analytics and AI/ML, most companies still struggle to establish an efficient and programmatic way to scale analytics and AI/ML. This is because deploying and scaling AI/ML in production has numerous challenges. The primary one is the technical debt[4] which creates a barrier to production. Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small orange box (Figure 1) in the middle. The required surrounding infrastructure to actually make ML work is vast and complex.

To overcome these hidden challenges in their journey to AI, clients need an end-to-end data and ML platform that streamlines and simplifies ML model implementation. Cloudera Data Platform provides this ultimate ML platform that addresses the many thorny AI/ML deployment challenges.

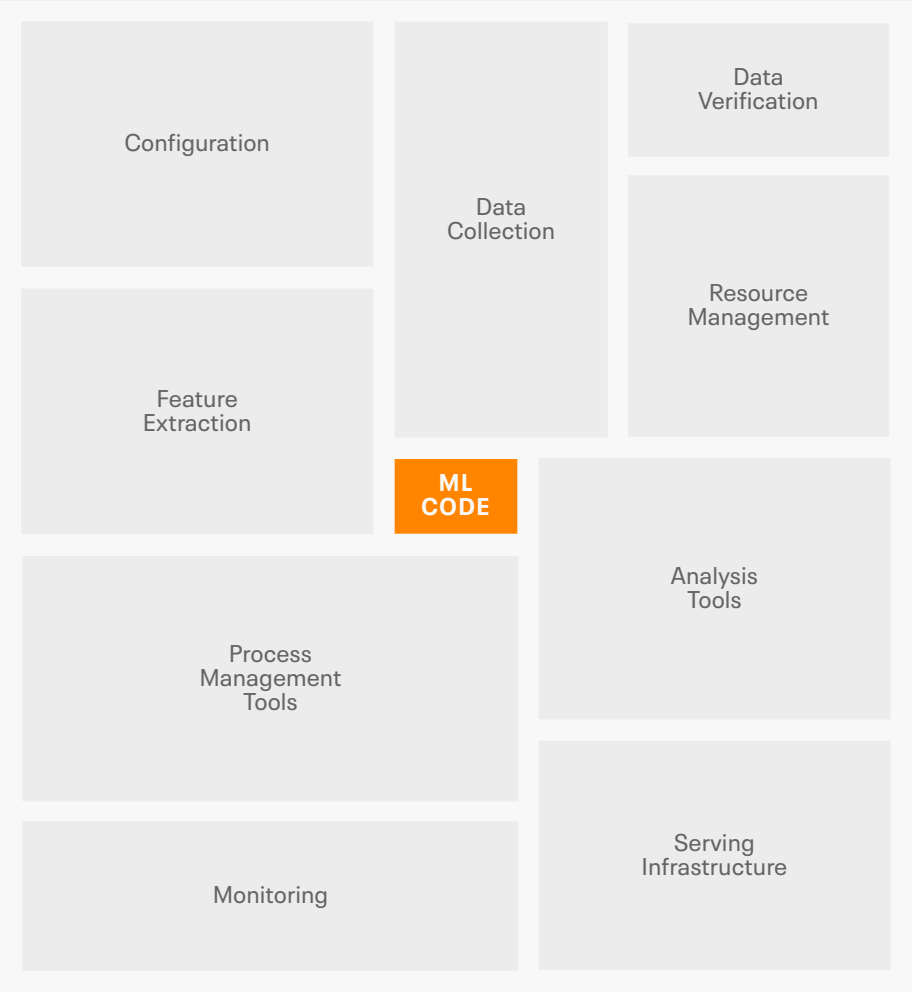


Figure 1: Hidden ML life cycle challenges

Challenges Deploying ML Models and Powering Predictive Use Cases

Developing and putting machine learning capabilities to work for the business requires more than machine learning tools and technologies. It requires end-to-end iterative data and analytics pipelines:

- Starting with the data source: via streaming or batch ingested into the pipeline
- Data engineering is needed to clean and prep for machine learning analytics
- Then data science tools are required to build and train models
- A complete production ML tool set is essential for deploying, monitoring, and retraining models
- Analytic services are needed to embed and operationalize that intelligence in the fabric of business through a combination of:
  - Predictive services/APIs
  - BI Dashboards
  - Data Products/Apps
  - Or intelligent edge devices
- Producing and validating insights needed for business enablement
- Lastly, ensuring data and pipelines are managed with security and governance throughout the lifecycle.

However, as clients traverse their AI/ML journey from data to business value (Figure 2), they encounter many obstacles that often cause failure.

Figure 2 summarizes the top practical challenges associated with deploying AI/ML in production and the key requirements to address these impediments. A recent Cloudera paper[5] explores these issues in greater detail.

Here the focus is on detailing the key requirements of an end-to-end ML platform. What's needed to support deep and enterprise-wide collaboration across the iterative AI/ML journey? And most importantly, what are the key user-centric and operational elements of an ultimate ML platform to ensure deployment doesn't have to be hard and complex?

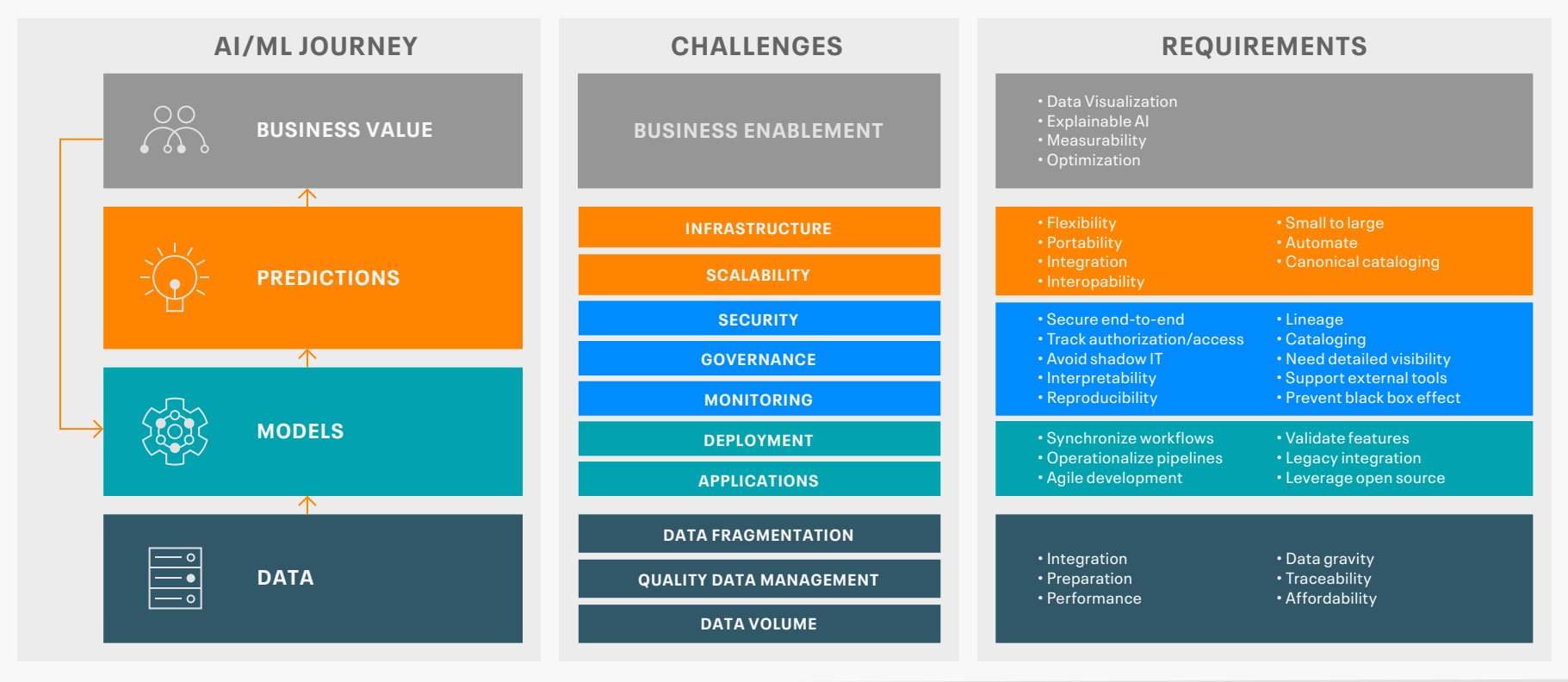


Figure 2: Challenges associated with deploying AI/ML

Key User-Centric Elements of an Ultimate ML Platform

ML initiatives are complex workflows and require that clients collect all the data they need, govern it to ensure it is trustworthy, analyze and build the necessary algorithms and be able to put the results into production. The individuals responsible for these activities are often disparate and disconnected and it requires a collaborative holistic approach (Your ML or AI journey is a team sport) to make this work efficiently.

An ideal ML platform must support a common collaborative environment capable of running all analytic processes in one place. This empowers customers to ingest data from many sources, build data pipelines, train ML models, get to production, and share insights to the business—all from the same secure, collaborative environment/ platform throughout the workflow.

A user-centric view of the various phases of a typical ML Journey (Data to Models to Predictions to Business Value) along with the key people/roles and their tasks are shown in Figure 3.

A typical ML environment is a collaborative effort of Data Engineers who acquire and process the data, Data Scientists who create the models, ML Engineers and ML DevOps personnel who deploy and run the applications and Business Analysts who interpret the results and make them actionable. The ultimate ML platform must support the activities of all these roles with common security and governance across the entire flow. An in-depth look at each role follows.

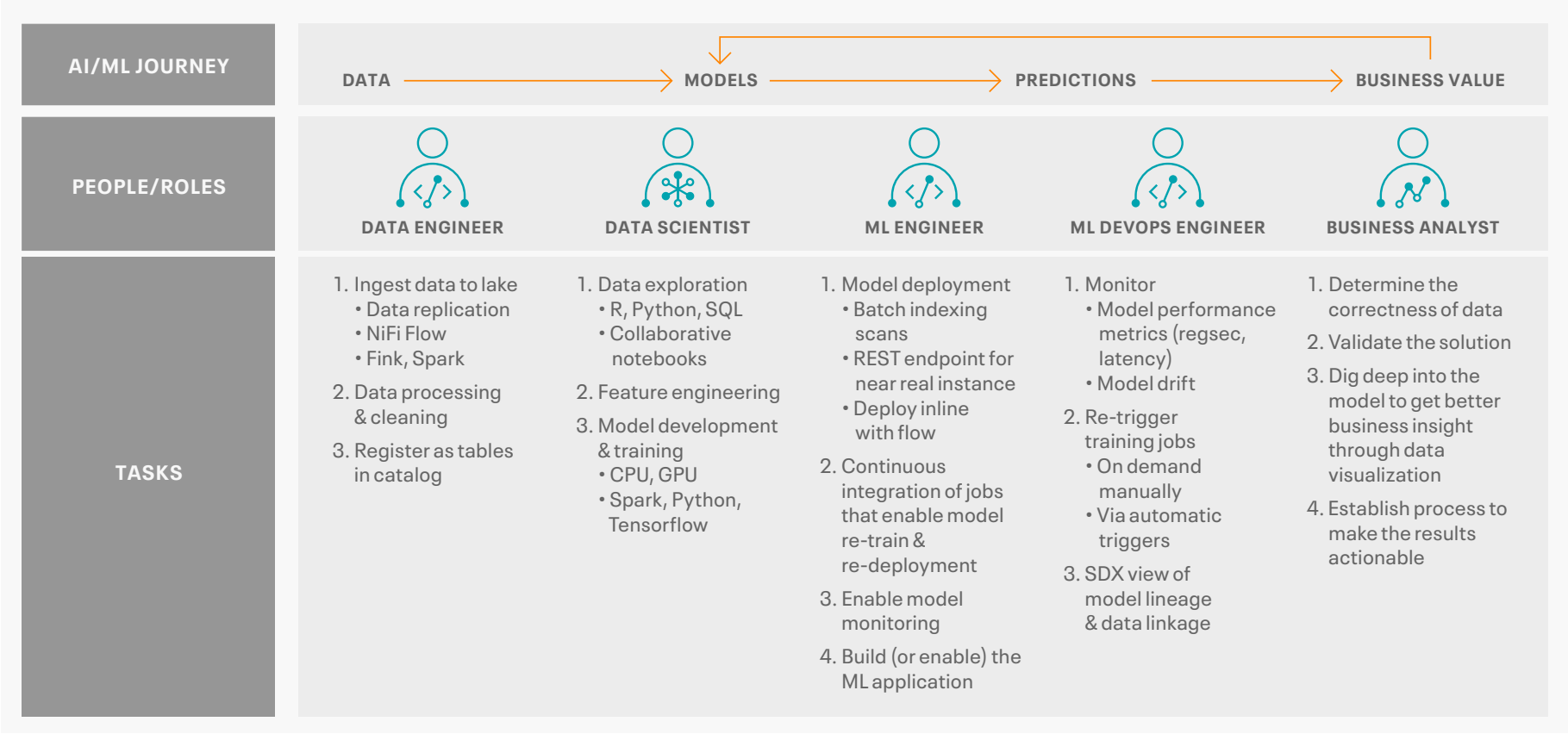


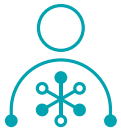
Figure 3: User-centric elements of an ultimate ML Platform

Data Engineers

Data Engineers must acquire the data to analyze, ingest into the data lake and prepare the data to make it ready for analysis. Raw data is typically not in a convenient format for a developer to run analysis since it was formatted by somebody else without that developer's requirements in mind. Also, raw data often contains semantic errors, missing entries, or inconsistent formatting, so it needs to be "cleaned" prior to analysis.

Data Engineers keep track of provenance, i.e., where each piece of data comes from and whether it is still up to date. It is important to accurately track provenance since data often needs to be re-acquired in the future to run updated experiments. A Data Engineer's tasks are:

- Ingest data to lake
  - Data replication
  - NiFi Flow
  - Fink, Spark
- Data processing and cleaning
- Register as tables in a catalog
- Automating ML data pipelines wherever they need to go



Data Scientists

Data Scientists wrangle (transform and map) data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream analytics. In addition, they are involved in data modeling and training tasks to produce a descriptive set of relationships between various types of information that are stored in a database.

A key focus is to create the most efficient method of storing information while still providing for complete access and reporting. There is also a constant need to test the model to determine if the objectives are being met. A Data Scientist's tasks are:

- Data exploration
  - R, Python, SQL
  - Data Visualizations
  - Collaborative notebooks
- Feature engineering
- Model development & training
  - CPU, GPU
  - Spark, Python, Tensorflow, etc.

ML Engineers

ML Engineers deploy the ML model in a production environment connected to business applications and make predictions using live data. The goal is to provide endpoints to run and control the developed pipeline, and easily integrate with other business systems using standard APIs. They evaluate batch vs. real-time prediction approaches in terms of cost, infrastructure, and complexity. A Data ML Engineer's tasks are:

- Model deployment
  - Batch indexing scans
  - REST endpoint for near real instance
  - Deploy in line with flow
- Continuous integration of jobs that enable model retrain & redeployment
- Enable model monitoring
- Build (or enable) the ML application

ML DevOps Engineers

ML DevOps Engineers monitor model performance, capture any performance degradation, and update models as necessary. Automation helps enterprise-level, end-to-end data science operationalization with minimum effort and maximum impact. This enables operationalizing complex AI/ML projects.
An ML DevOps Engineer's tasks are:

- Monitor
  - Model performance metrics (regsec, latency)
  - Model drift
- Re-trigger training jobs
  - On demand manually
  - oVia automatic triggers
- Tracking and management of data and model lineage
- Enabling software engineering workflows for data and AI-powered products.

Business Analysts

Business Analysts are essential for AI/ML initiatives. With their business and industry knowledge, they think critically, facilitate/influence collaborative decision-making, resolve conflicts, and solve tough business problems. Key responsibilities include:

- Determine the correctness of data
- Validate the solution
- Dig deep into the model to get better business insight through data visualization
- Establish processes to make the insights actionable.

One important thing to remember about these roles is that not every organization will have all of these functions — and that's OK. Different team members may take ownership of various aspects of the lifecycle and there's no single right answer as to how responsibilities are distributed — It depends on your needs, use case, and scale.

The ideal ML platform also requires several operational elements.

Key Operational Elements of an Ultimate ML Platform

Point solutions are not optimal for the AI/ML journey since they break IT security and governance, require data scientists to move data which could create shadow IT and other issues. What's needed is a holistic solution (Figure 4) to address these issues and other challenges associated with monitoring, deployment, governance, security, scalability, and infrastructure.

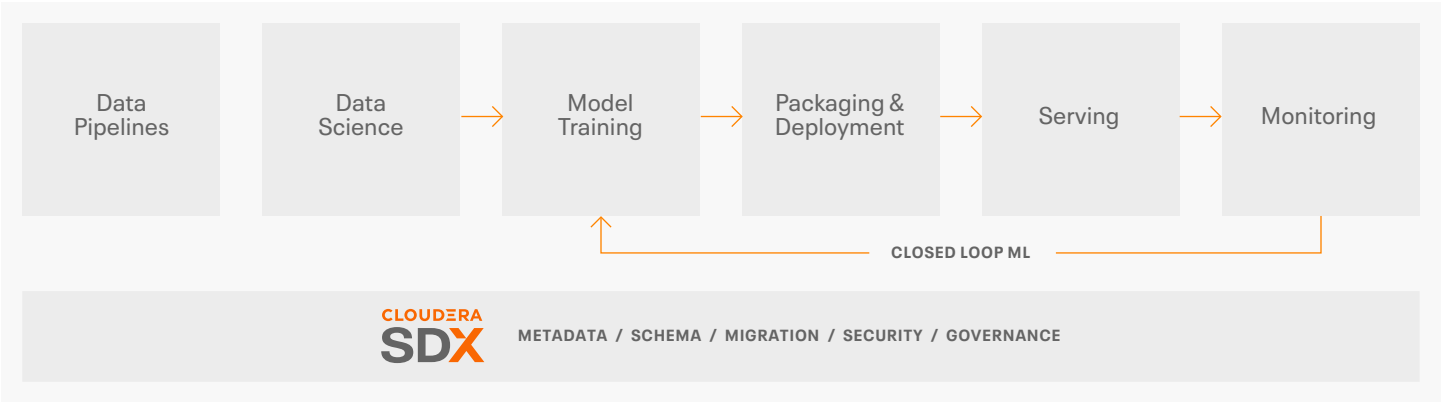


Figure 4: Operational elements of an ultimate ML platform

The operational elements of an ideal ML platform are:

- **Collaboration and transparency** to get/collect the right data, clean/prepare that data in the right format, create/automate ML pipelines, and enable data scientists with proper IDs, tools, and runtimes.
- **End-to-end, unified, and integrated** platform to avoid point solutions, synchronize and operationalize the data and pipelines, and help validate/test numerous features.
- **Unified model monitoring** platform for all the deployed models from a single pane of glass. Visual cues and alerts can be set to track both technical benchmarks and custom mathematical metrics.
- **Visibility of models and features** within teams and across organizations for model governance.
- **End-to-end governance and enterprise security** to deliver production models with inherited security, unified authorization, and access tracking.
- **Consistent microservices based architecture** to scale from small to large volumes of data and automate model creation.
- **Flexibility to run anywhere** on-premises or in multiple clouds and the ability to leverage common security, privacy, and governance standards across the AI/ML journey.
- **Port and integrate** legacy applications, existing silos of information, and support different infrastructures—multi-cloud, on-premises, and hybrid.
- **Catalog of prepackaged ML use cases, models, and algorithms** on wide ranging topics (e.g., fraud detection, Image analysis) to leverage past experiences, saving time and improving productivity.
- **Integrated business enablement/ data visualization** tools that enable business users to build dashboards, offer visual recommendations and provide predictive insights.

Lastly, production model deployment is one of the most difficult 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. Often, there is a discrepancy between the programming language in which a ML model is written and the languages the production system can understand, and recoding the model can delay model implementation. Deploying models to production means integrating the model with existing services in the production environment by creating endpoints which other services can call. Also, a complete production ML tool set is crucial for deploying, monitoring, and retraining models if businesses want to leverage ML at scale.

Only Cloudera has all the key solution components to deliver the ultimate ML platform.

Cloudera Solutions for the Ultimate AI/ML Platform

The Cloudera Data Platform (CDP), an open data platform, provides a rich portfolio of products and services (Figure 5) to help clients overcome ML challenges and adopt best practices to industrialize AI.

Each of the persona(s): Data Engineer, Data Scientist, ML Engineer, ML DevOps Engineer and Business Analyst play an important role in executing the various phases of the ML workflow. Figure 5 also shows the Cloudera solution and which persona is the likely (but not limited to) user of the solution. Brief descriptions of Cloudera solutions follow.

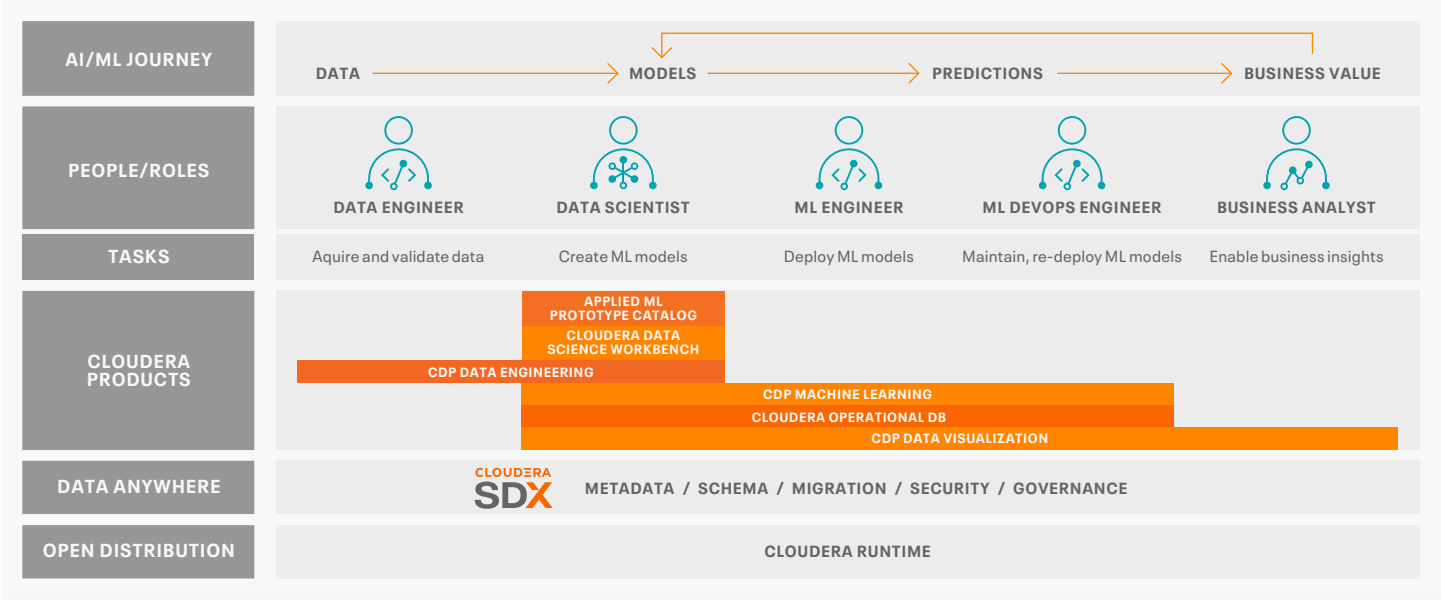


Figure 5: Cloudera solutions for the AI/ML journey

- **Cloudera Data Platform (CDP):** Is the industry's first Enterprise Data Cloud and offers a full complement of open-source data management and multi-function analytics, with the agility, elasticity, and ease of use of a public cloud-like experience. It provides a single control plane to manage infrastructure, data, and analytic workloads across hybrid and multi-cloud environments with shared services to safeguard data privacy, regulatory compliance, and cybersecurity threats across all cloud environments.
- **Cloudera Data Engineering (CDE):** Building on what has become the de facto computing framework for modern data engineering—Apache Spark—CDE is an all-inclusive data engineering toolset that enables orchestration automation with native Apache Airflow, advanced pipeline monitoring, visual troubleshooting, and comprehensive pipeline management tools to streamline ETL processes across enterprise analytics teams. For data science and machine learning teams, this means the data used in ML models is optimized, always-on, and perpetually accurate—enabling more AI use cases with lower risk for decision-makers.
- **Cloudera Machine Learning (CML):** Provides a secure, self-service enterprise data science platform that lets data scientists manage their own analytics pipelines, thus accelerating ML projects from exploration to production. It allows data scientists to bring their existing skills and tools, such as R, Python, and Scala. CML delivers an all-inclusive ML solution across private cloud (on-premises), multi-public cloud, and hybrid cloud deployments.

  Data scientists can build models directly from this data without moving or transferring any workloads, then deploy models into production with just a few clicks. Getting to production and creating trust with decision-makers is one of the biggest obstacles to successful AI use cases. This is why, in addition, to secure self-service access to ML data pipelines, libraries, runtimes, and IDEs; CDP Machine Learning enables best-in-class production ML capabilities and insight sharing features that make it simple to deploy, monitor, govern, and deliver results everywhere across the business.

  CML also enables a complete production ML toolkit, complete with model cataloging and the ability to monitor not just model performance, but also individual predictions down to the feature level. This is especially useful for deeply analyzing and ground-truthing models in production, then automating model retraining based on changing accuracy resulting from continuous learning.
- **Cloudera Shared Data Experience (SDX):** Cloudera's integrated set of security and governance technologies built on metadata and delivering consistent context across all analytics and public as well as private clouds. An intrinsic part of CDP, SDX reduces security risk and operational costs by delivering consistent data context across deployments. Multi-tenant data access and governance policies are set once, and automatically enforced across the data lifecycle in hybrid as well as multi-clouds. With these, organizations can deploy fully secured and governed data lakes faster and at scale.

- **Cloudera Data Visualization:** Enables everyone across the ML lifecycle to share insights quickly and easily and build complete predictive reporting applications in a drag and drop interface—directly inside of Cloudera Machine Learning without moving or copying data to third party tools. ML models can be exposed and queried to make new predictions in an end-user application—effectively completing the ML lifecycle and delivering true end-to-end ML that makes it easy to adopt and scale AI use cases across the business.
- **Cloudera Operational Database:** Empowers developers to automate and simplify database management with capabilities like auto-scale, auto-heal, and auto-tune. It is fully integrated with CDP, enabling end-to-end visibility and security with SDX as well as seamless integrations with CDP services such as Data Engineering, Machine Learning, and Data Warehouse.
- **Applied ML Prototypes (AMPs):** Enable data scientists to go from an idea to a fully working ML use case in a fraction of the time, with an end-to-end framework for building, deploying, and monitoring business-ready ML applications instantly. AMPs (Figure 6) move the starting line for any ML project by enabling data scientists to start with a full end-to-end project developed for a similar use case, including all the best practices, frameworks, trained and deployed ML models, as well as prebuilt predictive business applications, out of the box. This means that ML development teams can tackle their own ML business use cases more quickly, for everything from churn modeling, to sentiment analysis, to anomaly detection and beyond.

Clients have unprecedented flexibility and choice to deliver fast ROI with the Cloudera Data Platform.

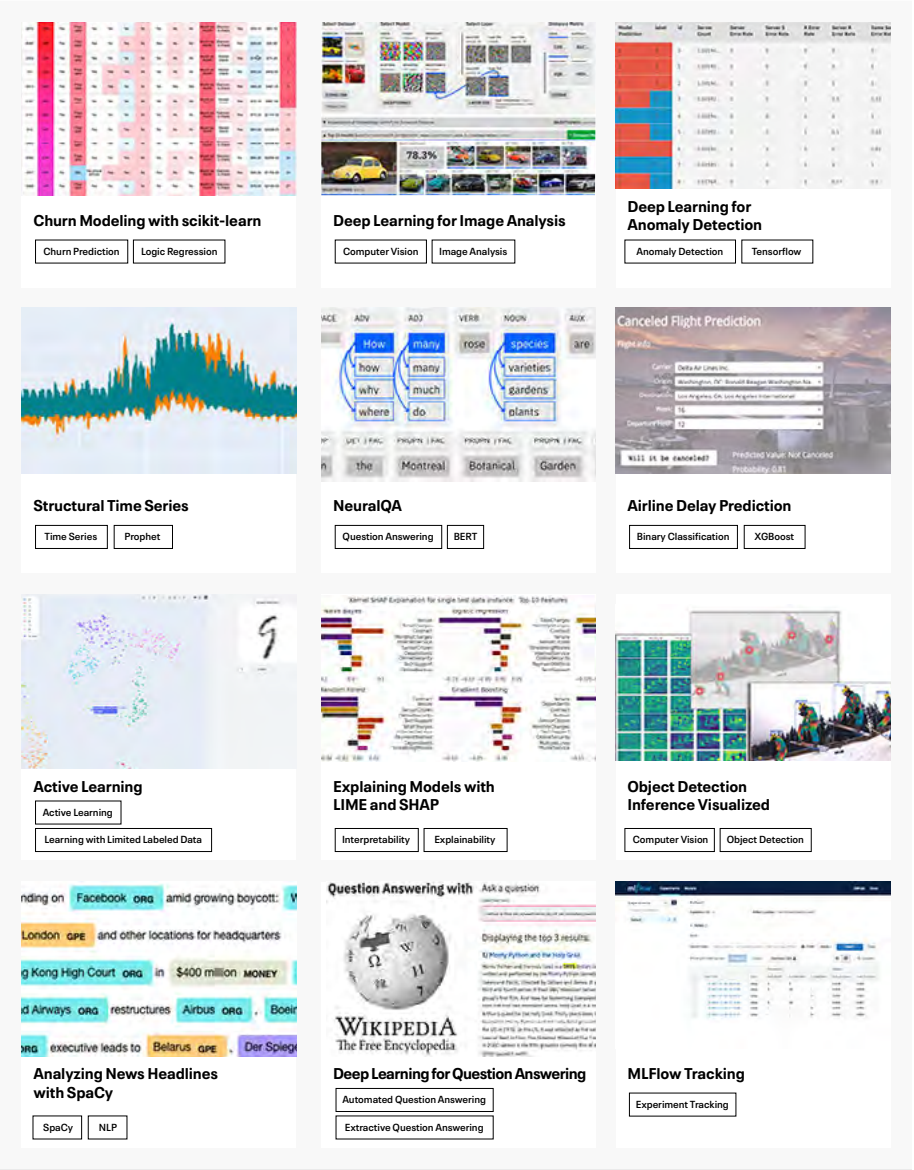


Figure 6: Applied Machine Learning Prototypes (AMPs) in CML

Deploying Machine Learning Projects on CDP

CDP is the core of the deployment platform. This next generation platform enables analytics from the edge to AI, including CML with a single shared data experience for enterprise security, governance and automation in any on-premises, private cloud, hybrid cloud, or multi-public cloud environments (Figure 7).

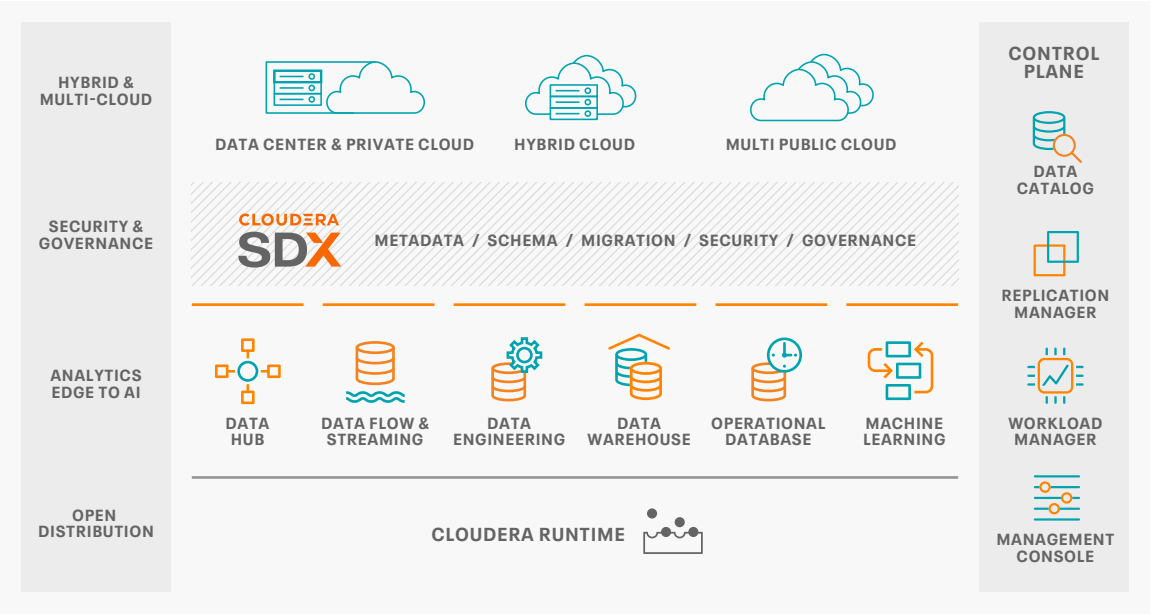


Figure 7: Deployment options for the Cloudera ML platform

Key capabilities of this deployment platform include :

Support for Data Center and Private Cloud, Hybrid, Multi-Cloud

- Move data and applications without rewriting and retraining
- Separate data management strategy from infrastructure strategy
- Manage all environments from a single pane of glass

Multi-Function and Open

- Deploy one platform to address current and future workload needs
- Connect disparate workload types to develop Edge2AI applications on one platform
- Open source and open API

Secure and Governed

- Manage data security and governance centrally
- Automate application security at all layers
- Reduce time to value with enterprise-grade productivity tools

Cloud Native Experience Everywhere

- Easy to use with self-serve capabilities
- Elasticity and agility to meet changing demands of workloads and business
- Simple to manage and maintain environments and applications.

Cloudera also provides comprehensive services to enable clients to deploy and manage this ultimate ML platform.

Cloudera Services for AI/ML

Cloudera AI/ML services (Figure 8) help customers fill their capability gaps in ML from a business, process, people, technical and organizational perspective. These services help organizations through the planning, feasibility, prototyping, and deployment of real projects.

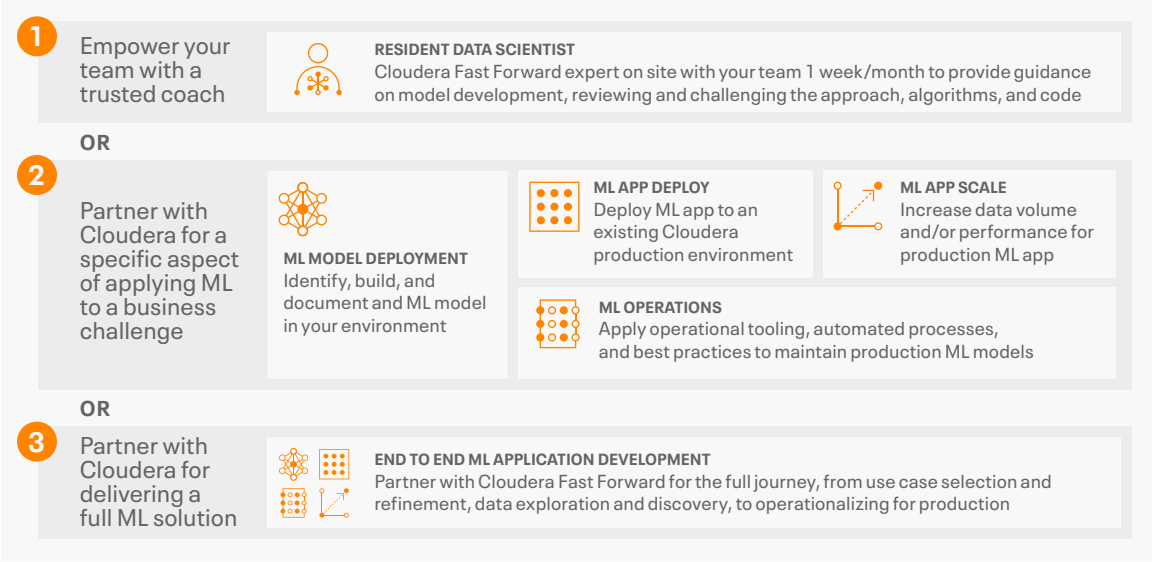


Figure 8: Cloudera AI/ML services

The key services provided include:

- **Resident Data Scientist:** CFFL experts on site with client teams for 1 week/month
- **ML Application Development:** Identify, build, and document an ML app in customer environment
- **ML Application Deploy:** Deploy existing application to existing Cloudera production environment
- **ML Application Scale**: Increase data and/or performance for existing production ML app
- **ML Operations:** Apply operational tooling, automated processes, and best practices to maintain production ML models.
- **End-To-End ML Application Development:** Partner with Cloudera to deliver a complete project from use case selection to production.

Working with 1000s of organizations, Cloudera is at the forefront of developing best practices and solutions to empower customers to build, deploy and manage production ML workflows across the enterprise on an open, enterprise platform with their data, skills and intellectual property (IP). Only Cloudera offers an ultimate ML platform, tools and expert guidance and services to help clients worldwide unlock business value from AI/ML.

Client Examples

The ultimate ML platform from Cloudera is allowing customers to meet the challenges of managing big data and is delivering substantial benefits across various industries. Four representative clients are highlighted here.

IQVIA, INC.

Challenge

IQVIA CORE™ platform wanted to:

1. Integrate all data together
2. Bring analytics to the data
3. Be deployed as IQVIA's Human Data Science Cloud

Solution

Global data lake and compute infrastructure serving 100% of tenants, 1000s of users, and 100s of different use-cases.

Wide array of ML techniques including deep learning for 100s of models embedded in ML solutions for cohorting and longitudinal analytics.

Outcomes

Cloudera chosen for Big Data Factory architecture.

Delivers 10+ PB centrally managed and privacy-governed healthcare data and high-performance analytics infrastructure for thousands of global users at IQVIA customers for greater innovation and improving healthcare outcomes.

Number 1 global provider of advanced analytics, technology solutions, and contract research services to the life sciences industry.

Office for National Statistics

Situation

UK ONS' Data Access Platform (DAP) serves 600+ CDSW users, meeting new data science talent with modern, open (Python, R, Spark) tools for DS & DE and enabling migration from costly SAS and Oracle. CDSW is used for 100s of daily data pipelining jobs and daily monitoring ~50 deployed models.

Solution

Cloudera Fast Forward Labs (CFFL) consulting to guide organizational skills transformation and ML industrialization strategy. A CFFL Resident Data Scientist helps with DAP utilization, upskilling and new capabilities dev including ML applied to migration statistics and the Census.

Outcomes

UK ONS is meeting their organizational 100% analytics accuracy, timeliness, and security goals and providing improved digital statistics including 1.3% measured increase in GDP.

#1

UK's largest independent producer of official statistics and the recognized national statistical institute of the UK. To modernize the data and analytics platform, expanding available data souces for fine-grainde analysis and providing faster, more detailed statistical estimates including GDP.

EXPERIAN PLC

Challenge

Data across 250 different data warehouses. Took days to copy data from silos for analysis. Limited by performance and scalability restraints.

Solution

New data product enabling users to develop and share a range of benchmarking scorecard and customer analysis in real-time; maintaining security, speed, and scale.

Outcomes

Analytics used by top 15 US banks and lenders. Experian customers see up to 75% speed improvements for data processes and archiving.

Best Overall Analytics Platform winner of the Fintech Breakthrough Awards for their Cloudera-powered Ascend Analytics Platform.

UNITED OVERSEAS BANK

Challenge

Lacked a solution for deeper analytics capabilities across 2PB of growing data to enhance the bank's performance.

Solution

Utilizes CML in support of their AI and data science roadmap to drive adoption across the bank.

Outcomes

25 AI projects in progress. Anti-money laundering project reduced false positives by up to 60%.

40%

40% increase in ML operational efficiency across entire business.

Delivering the ultimate ML platform

Many clients are implementing high-value ML use cases in several industries. For this they need a reliable partner with deep expertise to overcome the many challenges with deploying and scaling ML in production and avoid the high failure rates typical with ML projects. In addition, a complete production ML tool set is essential for deploying, monitoring, and retraining models.

Cloudera provides the ultimate ML platform. It is an open, unified, collaborative, secure and governed enterprise-grade production platform to run and manage all ML models with transparency, consistency, trust and high-performance.

This platform drastically reduces time to value for production ML models by enabling Data Engineers, Data Scientists, ML Engineers, ML DevOps Engineers, and Business Analysts to collaborate in a single unified platform; purpose-built for agile, iterative production ML workflows with enterprise-grade governance, security, and monitoring capabilities built in.

This ultimate ML platform accelerates a client's ML and AI journey and provides them unprecedented choice and flexibility to deploy and scale anywhere. This makes ML ubiquitous, practical, repeatable, simple, functional, and cost effective. Clients can now not only make internal processes faster and cheaper but also build better products and services, create brand new products, or completely reinvent processes.

Learn More

Machine Learning on Cloudera Data Platform (CDP) enables your data science teams to own and accelerate the full ML lifecycle—from data ingest, to ML development workflows, to business impact in a secure and scalable purpose-built platform.

Learn more about ML on CDP.

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

Sources

1 "Why do 87% of data science projects never make it into production?," July 19, 2019. https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

2 "Survey: 96% of Enterprises Encounter Training Data Quality and Labeling Challenges in Machine Learning Projects," May 23, 2019. https://www.businesswire.com/news/home/20190523005183/en/Survey-96-Enterprises-Encounter-Training-Data-Quality

3 Ritu Jyoti, "Accelerate Your AI Journey with a Hyperconverged Data and Analytics Platform," February, 2020. https://www.ibm.com/downloads/cas/ENDG17K3

4 D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems," n.d. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

5 "4 Essential Platform Factors for Enterprise ML," 2020. https://www.cloudera.com/content/dam/www/marketing/resources/whitepapers/4-essential-factors-for-enterprise-ml.pdf.landing.html

CLOUDERA