Eckerson Group | Deep Dive

# Deep Dive on Machine Learning Platforms

## Three Products to Consider

KEVIN PETRIE

JUNE 2022

RESEARCH SPONSORED BY

CLOUDERA

# About the Author

**Kevin Petrie** has deciphered what technology means to practitioners, as an industry analyst, writer, instructor, marketer and services leader. Kevin launched, built and led a profitable data services team for EMC Pivotal in the Americas and EMEA, and ran field training at the data integration software provider Attunity (now part of Qlik). A frequent public speaker and author of two books on data streaming, Kevin also is a data management instructor at eLearningCurve.

---

# About Eckerson Group

**Eckerson Group** is a global research and consulting firm that helps organizations get more value from data. Our experts think critically, write clearly, and present persuasively about data analytics. They specialize in data strategy, data architecture, self-service analytics, master data management, data governance, and data science. Organizations rely on us to demystify data and analytics and develop business-driven strategies that harness the power of data. **Learn what Eckerson Group can do for you!**

Eckerson Group
RESEARCH CONSULTING ADVISORY
GET • MORE • VALUE • FROM • YOUR • DATA

---

# About This Report

Research for this report comes primarily from numerous briefings with software vendors. This report is sponsored by Cloudera and Tibco, who have exclusive permission to syndicate its content.

*This is an excerpt of a larger report profiling three machine learning platforms. To read the full report, click **here**.*

# Table of Contents



*This is an excerpt of a larger report profiling three machine learning platforms. To read the full report, click **here**.*

# Introduction

Thinking machines based on artificial intelligence need lots of care, feeding, and supervision.

Artificial intelligence (AI) software makes decisions, creates content, and handles other tasks that normally require human cognition. Machine learning (ML), the most common category of AI, learns patterns in data to predict, recommend, or classify outcomes. It depends on an ML model, which is an equation that defines the relationship between the most telling data inputs, known as features, and outcomes.

ML models can run amok in many ways. They might find the wrong patterns, make wrong predictions, or fail to spot business changes. To reduce risk and deliver results, enterprises need to build and manage multiple ML models in a controlled way. They need to maintain business oversight as they deploy, optimize, and swap out models to maintain accuracy. They also need to foster open collaboration between business owners, data scientists, ML engineers, and data engineers. Most enterprises cannot meet these requirements with homegrown tools.

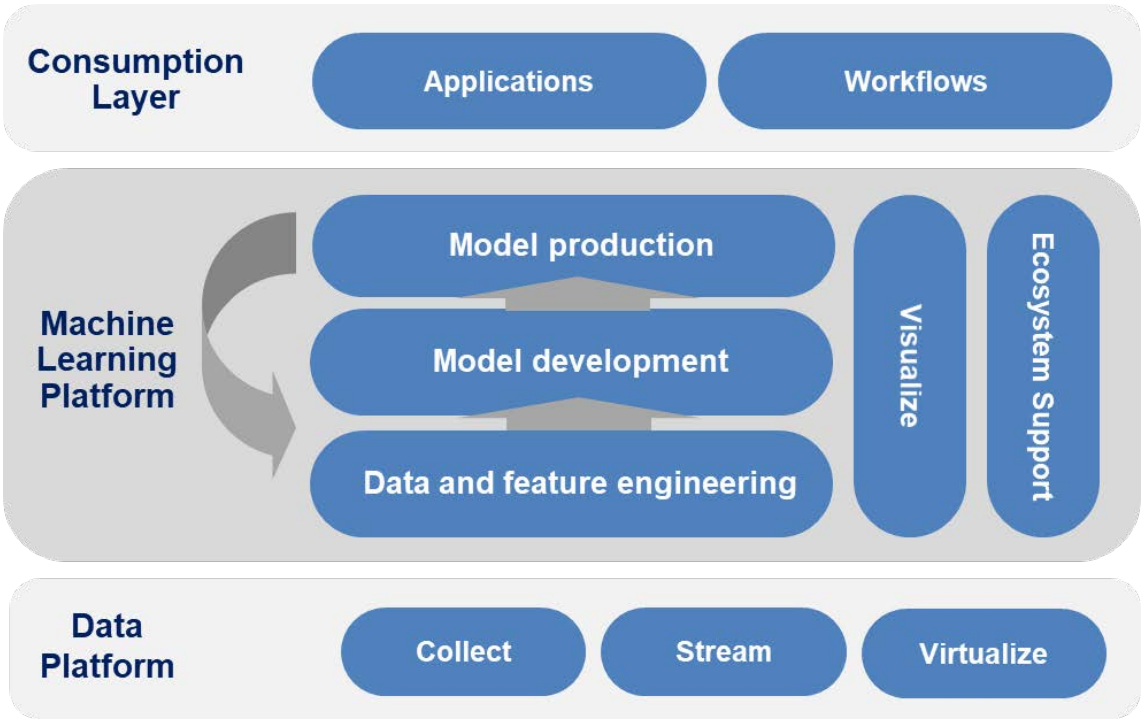
*Enterprises need to foster open collaboration between business owners, data scientists, ML engineers, and data engineers.*

## Enter the Machine Learning Platform

The ML platform enables these cross-functional teams to build and manage machine learning projects as part of an efficient, governed, and scalable process. It helps data science teams standardize features, ML models, and software code; then reuse and share those artifacts. When they operationalize the results, they have fewer surprises. (See figure 1.)

Figure 1 illustrates the ML platform architecture.

**Figure 1. Machine Learning Platform Architecture**




*The machine learning platform offers an environment in which data science teams can manage machine learning in an efficient, governed, and scalable way.*

The ML platform supports the three stages of the ML lifecycle: data and feature engineering, model development, and model production.

- **Data and feature engineering.** During this stage, the data scientist works with the data engineer to transform input data and label the historical outcomes. They derive the features from input data that best predict historical outcomes.
- **Model development.** The data scientist selects an ML technique and "trains" that algorithm on historical features. They check results and train it again until that algorithm delivers the desired accuracy.
- **Model production.** The ML engineer implements the ML model in production workflows, with the help of the data scientist. They monitor models' performance, accuracy, and cost, and organize their metadata in a governed catalog.

The ML platform also includes visualization and support for the ML ecosystem.

- **Visualization.** Data science teams visualize data during each stage of the lifecycle—for example, to

explore data, define features, or inspect the outputs of training or production models.

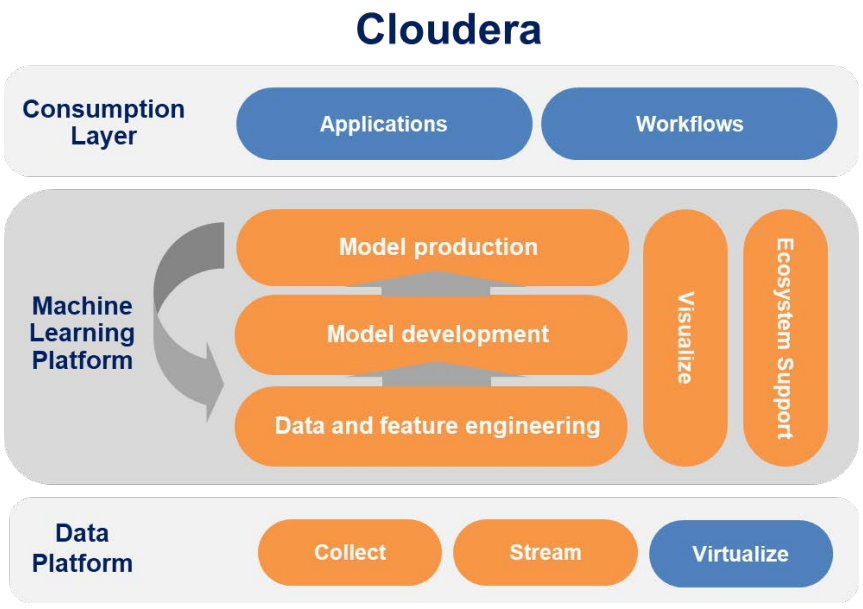
- **Support for the ML ecosystem.** Data science teams also use an ecosystem of feature stores that help build and manage features, libraries that share ML algorithms, notebooks that help build models, and catalogs that organize metadata for models and data. Some ML platforms contain similar native capabilities.

The lifecycle involves frequent iterations to respond to a dynamic business environment. When issues arise in production, data science teams need to go back and refresh training data, change features, or switch models.

The ML platform sits between two additional architectural elements: the data platform and consumption layer.

- **The data platform** presents data for model training and production. It might **collect** data for storage in a repository such as a cloud object store, **stream** data for processing in memory, or **virtualize** data for processing in a logical layer that spans different locations.
- **The consumption layer** includes **applications** that embed ML model outputs as well as **workflows** that stitch together multiple applications and models. For example, a workflow might include an ecommerce application, ML model for fraud detection, and payment processing application.

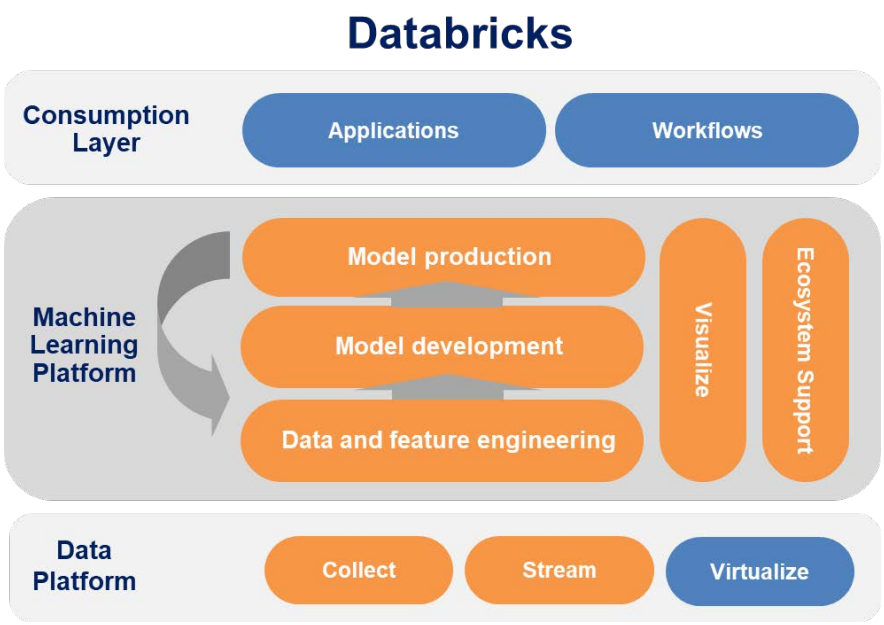
Together these elements give data science teams the flexibility they need to scale, adapt, and innovate based on changing business requirements. This report profiles three product approaches, each of which couples an ML platform with a data platform.



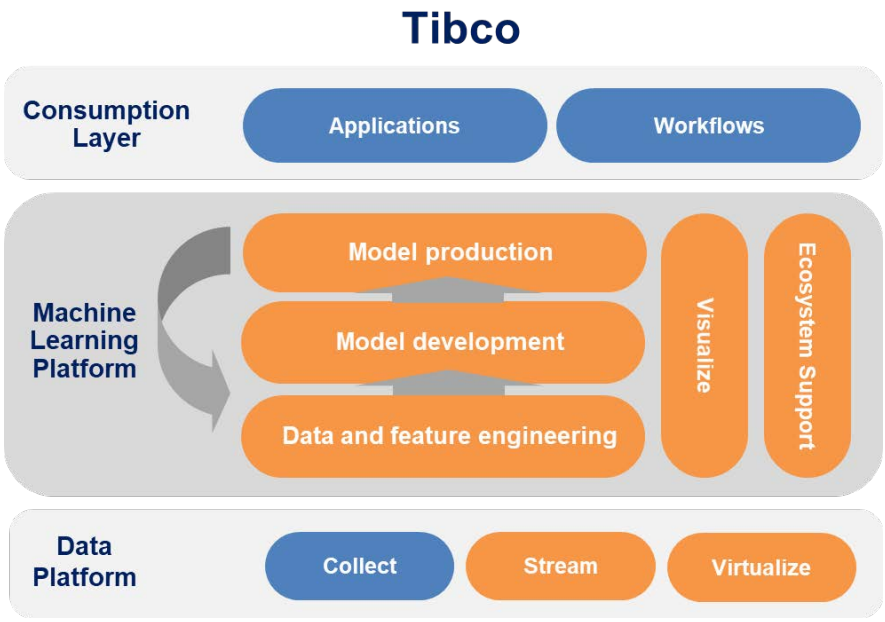

**Cloudera** offers an ML platform on top of its hybrid data platform. Cloudera collects and streams data in hybrid environments that include cloud and on-premises infrastructure. It helps manage data and feature engineering, model development, and model production. It also assists the lifecycle with visualization and ML ecosystem support. Building on its heritage as a data lake provider on-premesis, Cloudera helps data

science teams extend their ML initiatives to cloud and hybrid environments.

**Databricks** offers an ML platform on top of its cloud data platform. Databricks collects and streams data on cloud infrastructure; and helps manage data and feature engineering, model development, and model production. It also assists the lifecycle with visualization and ML ecosystem support. Databricks does not support hybrid environments. Building on its heritage in Apache Spark and open-source innovation, Databricks helps data scientists employ advanced AI/ML tools.




**Tibco** offers an ML platform on top of its hybrid data platform. Tibco streams and virtualizes data on hybrid infrastructure; and helps manage data and feature engineering, model development, and model production. It also assists the lifecycle with ML ecosystem support. Building on its own heritage in BI visualization, Tibco helps business owners and "citizen" data scientists oversee and participate in the ML lifecycle.




Each of these products has its own differentiators and solves distinct customer problems. This report seeks to help business, technical, and analytics leaders learn the key differentiators of the three products above and gain a better understanding of which product is best suited to meet the unique requirements of their organizations.

# Cloudera Machine Learning

**Founded:** 2008

**Product:** Cloudera Machine Learning

**Launched:** 2019

**CEO:** Robert Beardon

## Executive Summary

*Cloudera helps data science teams at midsized and large enterprises modernize, standardize, and simplify how they manage ML projects in distributed, heterogeneous environments. Cloudera unifies data management with consistent governance policies, data structures, and metadata. It unifies ML lifecycle management with containerized workspaces and data visualization capabilities. It also offers reusable project elements and streamlines infrastructure requirements. Cloudera works best for enterprises that need to accelerate their ML rollouts on hybrid environments, potentially including legacy Hadoop data lakes.*

## Background

### Company

In 2009, former Oracle VP Mike Olson recruited Doug Cutting, co-creator of Apache Hadoop software, to join the Cloudera venture he'd started with veterans of Facebook, Google, and Yahoo!. Their mission: empower enterprises to analyze high volumes of multi-structured data on commodity hardware. Olson and team offered a Hadoop distribution to help integrate and query data, write analytical programs, and manage what came to be known as data lakes. Cloudera raised nearly $1 billion, reached $261 million in annual revenue, and went public in 2017.

Over time, enterprises started to balk at the complexity of managing various Hadoop components on their data centers' inflexible hardware. Many of them directed new projects to pre-packaged data lakes and data warehouses that run on elastic object stores such as Amazon S3. Cloudera pivoted to object stores, acquired its close rival Hortonworks, and went private with a sale to private equity investors for $5.3 billion last year.

Today Cloudera, led by Hortonworks co-founder Robert Bearden, centers its strategy on the "hybrid data cloud." Cloudera enables enterprises to run any type of analytics on distributed infrastructure that might span multiple data centers and/or clouds, including Hadoop data lakes. It helps enterprises apply new tools to heritage Hadoop projects, spin up new projects on the cloud, and manage everything with a

unified governance framework. ML projects are a key growth driver for the Cloudera hybrid data cloud.

### Customers

Cloudera Machine Learning (CML) targets midsized and large enterprises that seek to simplify their ML lifecycle across private and public clouds. It enables data scientists and data engineers to perform exploratory analysis and define features. It also enables data scientists and ML engineers to standardize the ML lifecycle with reusable elements and streamline model training and deployment processes.

Cloudera sells to all verticals, and has an installed base that includes eight or more of the top ten global companies in financial services, telecommunications, auto manufacturing, pharmaceuticals, and technology. Data science teams implement CML to rapidly deploy new ML models in production applications and scale out their existing ML operations.

The ideal CML customer already uses the Cloudera hybrid data cloud for basic BI projects. Such customers base these projects on operational data—including semi- or unstructured data such as social media posts, clickstream logs, or IoT sensor feeds—that reside in the Hadoop file system or cloud object stores. Their data analysts and data scientists are ready to scale up their production of ML models for ever-rising volumes, varieties, and velocities of this data.

### Product

CML helps manage the ML lifecycle on data that the Cloudera hybrid data cloud collects and streams. It also provides visualization capabilities to assist data and feature engineering, as well as the consumption of model outputs. Like other ML platforms, CML supports the ecosystem of open programming languages, catalogs, libraries, notebooks, and feature stores. Cloudera differentiates its ML offerings with unified management of distributed data, unified management of ML projects, reusability, and runtime efficiency. Together these capabilities help modernize, simplify, and standardize the ML lifecycle.

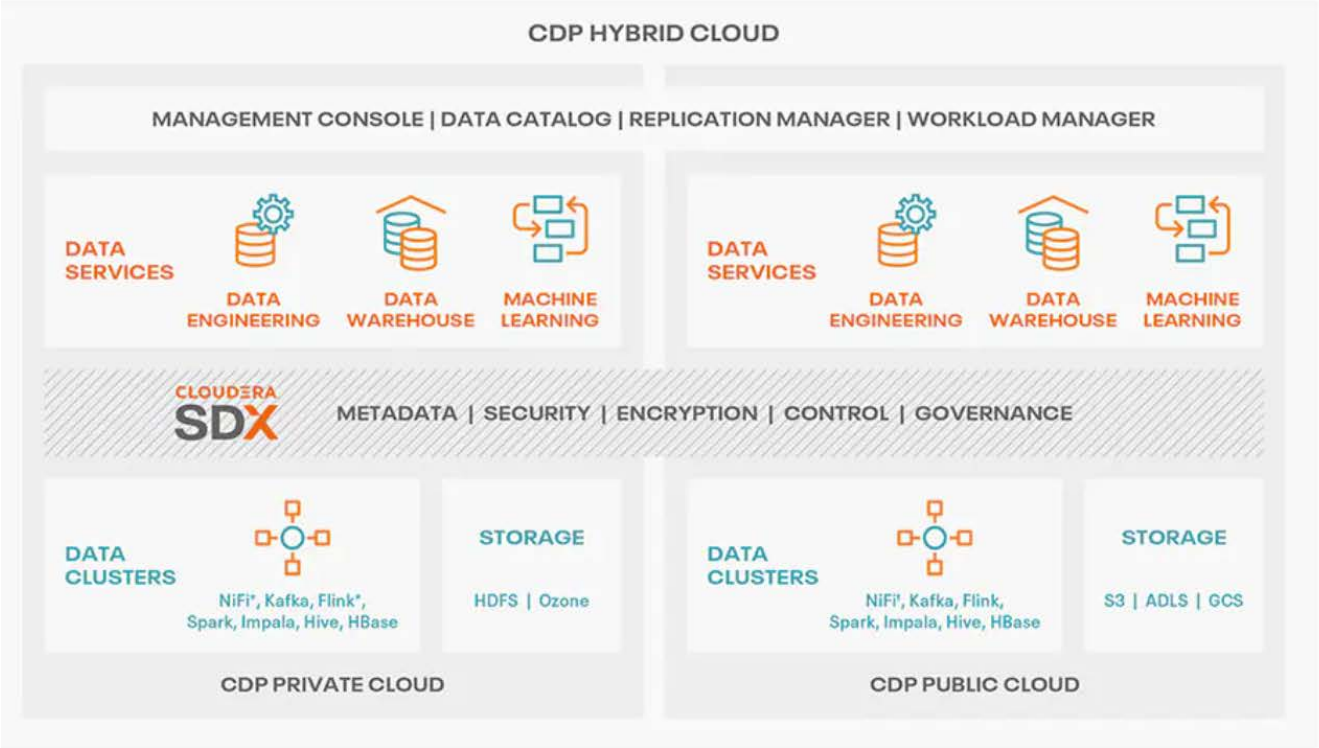
## Unified Management of Distributed Data

Let's start with the biggest differentiator, one that cloud-only platforms cannot match. The Cloudera hybrid data cloud helps customers unify how they manage the data that accumulates in different places. Many enterprises have data in Cloudera or Hortonworks data lakes that they implemented on-premises years ago, as well as new Cloudera data lakes on AWS, Azure, or Google infrastructure. Even as they start new Cloudera data lakes on the cloud, these enterprises keep a lot of data on premises or in edge systems thanks to migration complexity, data gravity, and local sovereignty requirements. And they hate to move between clouds thanks to pricey egress charges from cloud providers. This results in distributed, heterogeneous environments.

The Cloudera hybrid data cloud unifies the management of these environments with a "Shared Data Experience" (SDX) that includes consistent governance policies, data structures, and metadata. SDX automatically classifies data upon ingestion, then organizes and describes its structures, usage patterns, and searchable attributes in a central catalog. It builds on Cloudera's longstanding support of Apache projects for governance, using Apache Ranger to implement role-based access controls and Apache Atlas to track data and model lineage. It also identifies and protects personally identifiable information (PII) to assist compliance with privacy requirements.

Figure 2 illustrates the Cloudera hybrid data cloud.

**Figure 2. Cloudera Hybrid Data Cloud**




## Unified Management of ML Projects

Cloudera Machine Learning, meanwhile, helps users unify how they manage ML projects in these environments. This is the second most important differentiator. Data science teams use CML to create workspaces, each tied to a Kubernetes cluster that runs on elastic storage and compute. They create containerized projects in those workspaces, then transform and explore input data, define features, and then train and deploy ML models. They perform these tasks through the same governed framework, with consistent role-based access controls, tools, and workflows.

Data science teams can use both Cloudera and third-party ecosystem tools to manage each stage of the ML lifecycle. For example, data scientists and data engineers can use Cloudera Data Visualization (included in CML) to explore datasets, identify patterns and define features. Data scientists then can visualize ML model outputs while training or operating models and create intuitive dashboards to track model KPIs on an ongoing basis. ML engineers, meanwhile, can use developer tools such as Flask to build programs for ML models and automation tools such as Apache Airflow to deploy them in production workflows.

### Reusability

CML enables data science teams to reuse standard or custom artifacts across workspaces. For example, it provides reusable ML projects called Applied Machine Learning Prototypes (AMPs). These projects accelerate the ML lifecycle by packaging configuration files for each step—defining features, training a model, serving that model, deploying it, then monitoring and re-training it. Data science teams use AMP documentation of best practices as they adapt the config files, code, and ML model parameters to their business requirements, datasets, and environments. They also can create their own AMPs and reuse them across the enterprise.

Cloudera's Fast Forward Labs Research unit also publishes reports on ML techniques and best practices to help data science teams learn from their communities and increase the odds of success.

### Efficiency

The Cloudera hybrid data cloud streamlines infrastructure requirements by separating the consumption of storage and compute resources wherever they reside. This brings cloud-like efficiency to on-premises environments and ensures each workload only utilizes what it needs. Enterprises can spin up compute cycles without impacting storage, or they can add storage capacity without impacting compute.

CML also streamlines how data scientists and ML engineers implement the programs that train and operate ML models. They can create modular runtime environments whose developer tools, kernels, and libraries are only what that program and ML model need. If the program doesn't need to run Spark, its runtime will exclude Spark programming instructions. This modular, lightweight approach offers an alternative to the rigid, sometimes bloated runtimes of other ML platforms. It also makes it easier to reuse a program in a new environment. Rather than rewriting everything, the data scientist or ML engineer can just reassemble runtime modules to fit the new environment.

## Pricing

Cloudera offers its software for both public cloud and private cloud environments. It prices CML at 17 cents hourly per Cloudera Compute Unit, which includes both memory and CPU processing on public cloud infrastructure that AWS or Azure provide separately. CML prices increase based on the size of the

cloud instance. CML also is available as a Private Cloud Data Service for $650 per Cloudera Compute Unit per year, including dedicated storage and compute in the cloud. The other private cloud option is to run CML on-premises for $10,000 per node. CML includes Cloudera Data Visualization.

## Recommendation

Given its history on-premises, Cloudera is well positioned to help data science teams operate in the majority of enterprise data environments because they will remain hybrid for the foreseeable future. Cloudera helps modernize, standardize, and simplify ML projects on Cloudera's hybrid data cloud, which can span multiple data centers and/or clouds. The Cloudera Machine Learning product includes data visualization capabilities that assist each stage of the ML lifecycle. It offers unified management of distributed data, unified management of ML projects, reusability, and runtime efficiency.

The ideal Cloudera customer is large or midsized enterprise that seeks to:

- Modernize their ML lifecycle while still capturing value from heritage Hadoop data lakes.
- Reduce the complexity of ML projects by standardizing data management and creating reusable elements.
- Improve the efficiency of data processing, model training, and model production.
- Gain the flexibility to deploy more ML models, scale their operations, and iterate quickly, across both private and public clouds.

# Conclusion

In one sense, ML platforms serve as the foundation for a smart factory. They enable enterprises to standardize and control processes for designing, manufacturing, and delivering ML models, then carefully monitor model performance to ensure quality. Enterprise teams can tune their manufacturing processes, recall defective models, and swap them out with fresh versions. They can manage this ML lifecycle with unified interfaces and consistent, reusable elements.

But in another sense, ML platforms serve as a living environment. They enable enterprises to cultivate innovation and customize models as they respond to changing conditions. Enterprise teams, starting with business owners, can empower creative experts to find problems, devise solutions, and bring them back to the team. They can foster both individual initiative and collaboration, incorporating tools from the ML ecosystem.

This report profiled three leading examples of ML platforms, each of which balances these competing roles in a different way. Together, they demonstrate the range of options available to help your business and data science teams succeed with machine learning. However, there are many additional options within that range. The following questions will help you evaluate any ML platform to find the right answer to your business and technical needs:

- **How does this platform streamline the ML lifecycle?** Many moving parts contribute to the creation, production, and operation of ML models. Look for a platform that reduces friction by helping standardize elements and automate processes.
- **Does it foster the right level of innovation and customization?** Dynamic markets can change your business objectives or technical requirements on a dime. Evaluate platforms based on their ability to support rapid change, for example by helping retrain models, import new algorithms, or spin up a new project.
- **Does it support cross-functional collaboration?** Business owners, data scientists, data engineers, and ML engineers each play a critical role in the ML lifecycle—and each depend on one another. Look for a platform that enables these team members to communicate and pass the baton without needing extensive training.
- **How does it help manage risk?** ML project risks include angry customers, disrupted operations, and compliance issues. Be sure to find a platform that helps your business owners and data science teams keep these risks at an acceptable level by governing factors such as data usage, model accuracy, and model bias.

> **How does this platform fit into your existing data environment?** ML platforms must align with your existing hybrid, cloud, or multi-cloud environment. Evaluate ML platforms based on the level of effort it requires to integrate with the tools, data stores, and pipelines you already use.

# About Eckerson Group

Wayne Eckerson, a globally-known author, speaker, and consultant, formed Eckerson Group to help organizations get more value from data and analytics. His goal is to provide organizations with expert guidance during every step of their data and analytics journey.

Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps data analytics leaders abreast of the latest trends, techniques, and tools in the field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate business requirements into compelling strategies and solutions.
- **Our advisors** provide one-on-one coaching and mentoring to data leaders and help software vendors develop go-to-market strategies.

Eckerson Group is a global research and consulting firm that focuses solely on data and analytics. Our experts specialize in data governance, self-service analytics, data architecture, data science, data management, and business intelligence.

Our clients say we are hard-working, insightful, and humble. It all stems from our love of data and our desire to help organizations turn insights into action. We are a family of continuous learners, interpreting the world of data and analytics for you.

Get more value from your data. Put an expert on your side. Learn what Eckerson Group can do for you!

Eckerson Group

RESEARCH CONSULTING ADVISORY

GET • MORE • VALUE • FROM • YOUR • DATA

# About the Sponsor


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.