

Optimizing the Machine Learning Lifecycle and MLOps

The Emergence of Cloud Data Platforms

BY KEVIN PETRIE
FEBRUARY 2022

This publication may not be reproduced or distributed without Eckerson Group's prior permission.

RESEARCH SPONSORED BY SNOWFLAKE



About the Author



Kevin Petrie is VP of Research at Eckerson Group. For 25 years Kevin 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, consulting, and advisory 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!](#)



About This Report

This report is based on research and conversations with numerous industry practitioners and vendors. It is sponsored by Snowflake, which has exclusive permission to syndicate its content.

Table of Contents

Executive Summary	4
Rise of the Thinking Machines	5
The Machine Learning Lifecycle	7
Optimizing the Machine Learning Lifecycle and MLOps on a Cloud Data Platform	10
Next Steps	14
About Eckerson Group	16
About the Sponsor	17

Executive Summary

To make machine learning succeed at scale, data science teams must standardize and streamline the Machine Learning (ML) lifecycle (also referred to as ML Operations or MLOps) that spans data and feature engineering, model development, and model production. While data science teams can start ML initiatives with a piecemeal approach, as they grow they need to take a holistic approach to the lifecycle. They can benefit by standardizing on a platform that provides the necessary scalability, reproducibility, and governance. Traditionally, data scientists use ML platforms to achieve this. But they still needed to manage data and data pipelines across multiple repositories.

Now a new option has emerged: cloud data platforms that merge data warehouse and data lake constructs. Like an ML platform, the cloud data platform offers lifecycle speed, scale of production, governance, and support for the ecosystem of ML tools. But it also goes further and offers the ability to consolidate enterprise data, collaborate across functions and organizations, and integrate ML into operational workflows.

Effective data science teams should adopt these guiding principles as they consider standardizing their ML lifecycle and MLOps on a cloud data platform.

- > Seek the right balance between standardization and customization.** *Standardizing steps in the ML lifecycle improves productivity. But your team will still need to customize certain projects, for example by procuring hard-to-find datasets, or creating specialized models. Be sure to select and configure a cloud data platform that offers unfettered access to the ML ecosystem of libraries, development interfaces such as notebooks, programming languages and tools. Also be sure your data science team has the creativity, domain knowledge, and familiarity with these tools to customize where needed.*
- > Evaluate cloud data platforms based on their support of the ML lifecycle.** *As you start and then scale up your ML projects, evaluate the ability of your existing cloud data platform—or potentially a new one—to support your ML lifecycle. Ask vendors about their current capabilities and roadmap to help you automate and standardize the many aspects of building and operating ML models. Make this a “must-have” criterion for evaluating cloud data platforms.*
- > Plan with future requirements in mind.** *Expect your supply of data to grow, along with business demand to use it. This means you need a cloud-based platform that can elastically scale its infrastructure to accommodate new types of data, new data sources, use cases, ML models, and users, much like a factory floor with multiple assembly lines. It also means you need to keep recruiting new talent, and building new skills with your existing data science team.*

Rise of the Thinking Machines

Machine learning evokes the image of self-guided robots, when in fact ML models need continual management and supervision.

Organizations can optimize these aspects of machine learning operations (MLOps) by building and deploying ML models on a cloud data platform—while still tapping into a comprehensive ecosystem of ML tools.

Machine learning models need care, feeding, and oversight to ensure they do good rather than harm.

This report explores machine learning, the ML lifecycle, and the role of the cloud data platform in optimizing that lifecycle.

Key technologies

To start, let's define the key technologies involved.

Artificial intelligence is software that mimics human cognition in order to solve various problems.

Machine learning is the most popular type of AI, in which an algorithm discovers patterns in data and “learns” the relationship between data inputs and outcomes. This learning (or “training”) process creates an ML model that studies the most telling data inputs—known as features—and generates a score that predicts, classifies, or recommends future outcomes. ML scores also might derive from anomalies that influence future outcomes.

The **cloud data platform** is a combination of data warehouse and data lake constructs that support various workloads, including business intelligence (BI), machine learning, and data applications. Like a data warehouse, it queries and transforms data, for example using structured query language (SQL) commands or other languages such as Python and Java; delivers query outputs to BI tools; and secures and governs data usage. Like a data lake, it stores myriad types of structured, semi-structured, and unstructured data objects; and applies AI/ML models to those objects. Cloud data platforms support high performance and scalability on elastic cloud storage.

Cloud data platforms now not only support the **machine learning lifecycle**, they also can enable data science teams to streamline their ML Operations (MLOps). The ML lifecycle comprises the stages required for data science teams to build and operationalize ML models: data and feature engineering, model development, and model production. Traditionally enterprises use purpose-built machine

learning platforms—for example, **Amazon SageMaker**, **DataRobot**, **H2O**, or **Dataiku**—to support these activities. Their data science teams value the graphical interface of the ML platform, which reduces coding complexity, but might still rely on outside tools to connect with sources and ingest data. And they can benefit from integrations with a cloud data platform to handle high-scale, performance-sensitive workloads.

Why do enterprises need machine learning?

Enterprises continue to digitize their businesses to improve efficiency and create sticky customer experiences. This digital transformation accelerates interactions, transactions, and decision timeframes. It also generates reams of data that contain fresh insights about operations, customers, and competitors. Machine learning helps enterprises harness this data to gain competitive advantage. ML models can spot patterns within these reams of data in order to drive faster, more accurate decisions at a higher scale than humans could achieve on their own. This helps humans and applications take fast, smart action.

Machine learning helps enterprises harness reams of data to gain competitive advantage.

Machine learning use cases

Use cases for machine learning abound. Common examples include the following:

- > **Sales/inventory predictions:** A consumer packaged goods company predicts sales and resulting inventory requirements to assist supply chain planning ahead of the holiday shopping season.
- > **Fraud detection:** A credit card company classifies the risk of fraudulent transactions based on merchant location, product selection, and transaction size; and how those attributes compare with that buyer's history.
- > **Customer recommendations:** A video game company recommends avatars for players to use, based on players' skill levels, game choices, and prior avatar selections.
- > **Document processing:** A life sciences company summarizes and classifies medical journal articles for vaccine research, based on article authors, key concepts, and relevance to the research focus.
- > **Preventive maintenance:** A shipping company predicts when container ship engines will break down based on maintenance histories and IoT sensors that show wear and tear.

The Machine Learning Lifecycle

Let's drill into the three stages of the ML lifecycle: data and feature engineering, model development, and model production. While data science teams might start with a “do-it-yourself” approach, as they scale they standardize their lifecycle on an ML platform or now a cloud data platform. Along the way they can leverage libraries, notebooks, and other elements of the ML ecosystem.

Data and feature engineering

In this stage, the data scientist collaborates with the data engineer to integrate and transform their historical input data. They join, re-format, cleanse, and filter the data, then identify or derive the “features” that predict historical outcomes. For example, they might decide based on conversations with service technicians that tire pressure and brake pad erosion contribute to a lot of truck breakdowns. They decide those two types of IoT sensor data will become features. The data scientist also might need to label historical outcomes—for example, to distinguish roadside breakdowns from accidents.

The data scientist and data engineer might perform these tasks within a managed feature store such as **Tecton** and **Rasgo**. The feature store helps transform input data, define features, track them in a registry, and serve them to ML models. It layers onto the cloud data platform, integrating with its data and consuming its compute resources. Data science teams also now can use native feature store capabilities as part of an orchestrated process within the cloud data platform itself.

Model development

Once the data scientist has defined the features within their historical input data, they select an ML technique that they can use to define the relationship between features and outcomes. They might select a simple regression technique, which predicts a continuous set of numerical outcomes on an X-Y axis, or a classification technique that creates a decision tree with a series of “if/then” decisions. The data scientist trains their selected algorithm on historical data by running the algorithm and comparing its outputs to actual results. Then they adjust features and parameters, and run the algorithm again. Once the algorithm creates accurate outputs, they have a final model, ready for production.

Data scientists might download algorithms from AI/ML libraries such as **PyTorch** or **TensorFlow**, and use tools within those libraries to develop and train models. They transform data, visualize data, and develop models in AI/ML notebooks such as **Jupyter** and **Apache Zeppelin**. Cloud data platforms provide open integration with these libraries and tools to assist such efforts.

Model production

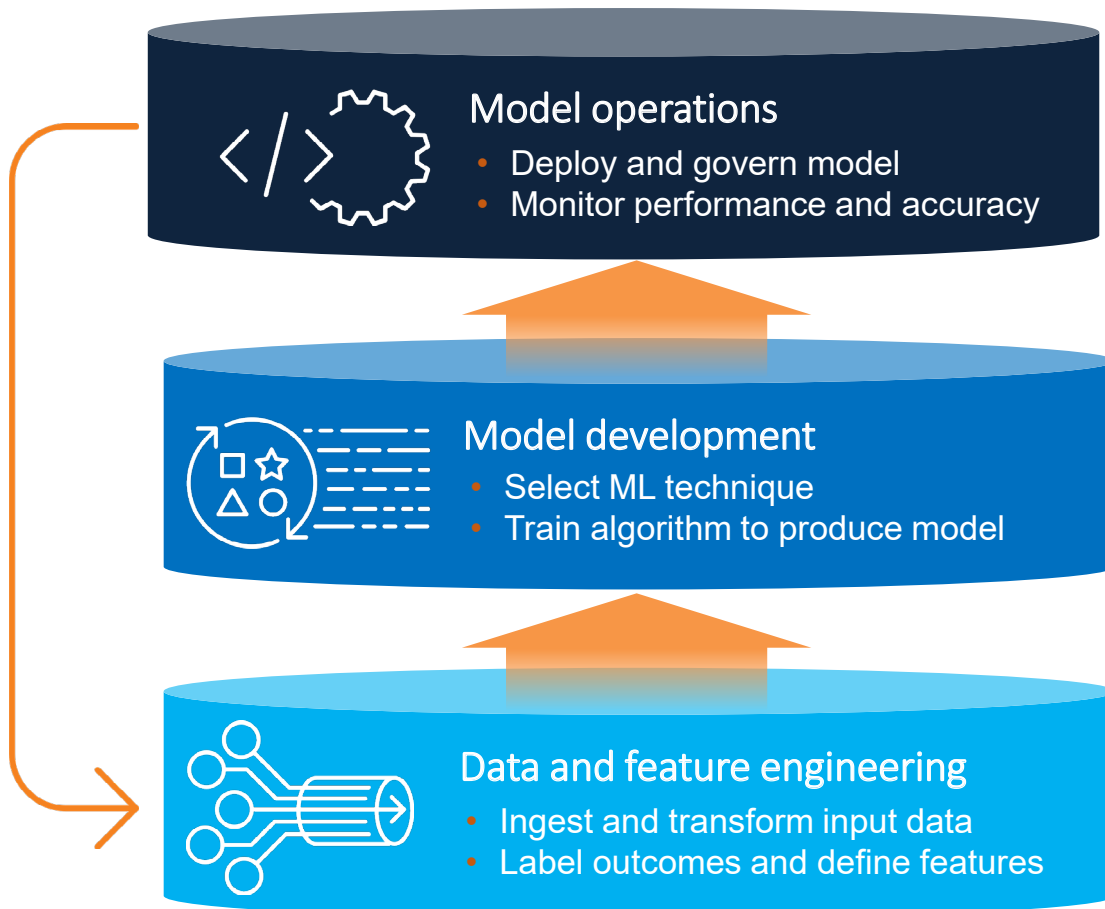
Now the ML engineer works with the data scientist to implement the ML model in production. They collaborate with the DevOps engineer to insert the model where it needs to go—perhaps in a supply-

chain management application, payment-processing workflow, or ecommerce website. They also collaborate with the data engineer to integrate the model with the right production data pipeline. Once they activate the model, they monitor its operational performance, cost, and potential bias. They also monitor model accuracy, for example to identify “model drift” when live business conditions change vs. historical training data. In addition, the ML engineer collaborates with governance officers to track those metrics in a model catalog.

Cloud data platforms integrate with the ML ecosystem to assist the production stage. ML engineers and developers might build and test software code that operationalizes ML models using development environments such as **GitHub**. Developers might use workflow tools such as **Airflow** to orchestrate workflows that operationalize ML model outputs.

Figure 1 illustrates the three stages of the ML lifecycle.

Figure 1. Stages of the Machine Learning Lifecycle



The ML lifecycle involves lots of iterations. When data science teams spot model drift, bias, or other performance issues, they must pull that model out of production for re-training on more recent historical data. They might need to select new features, select a new ML technique, or re-label historical outcomes, perhaps leveraging automated retraining capabilities within ML platforms or cloud data platforms. Success depends on vigilance and repetition.

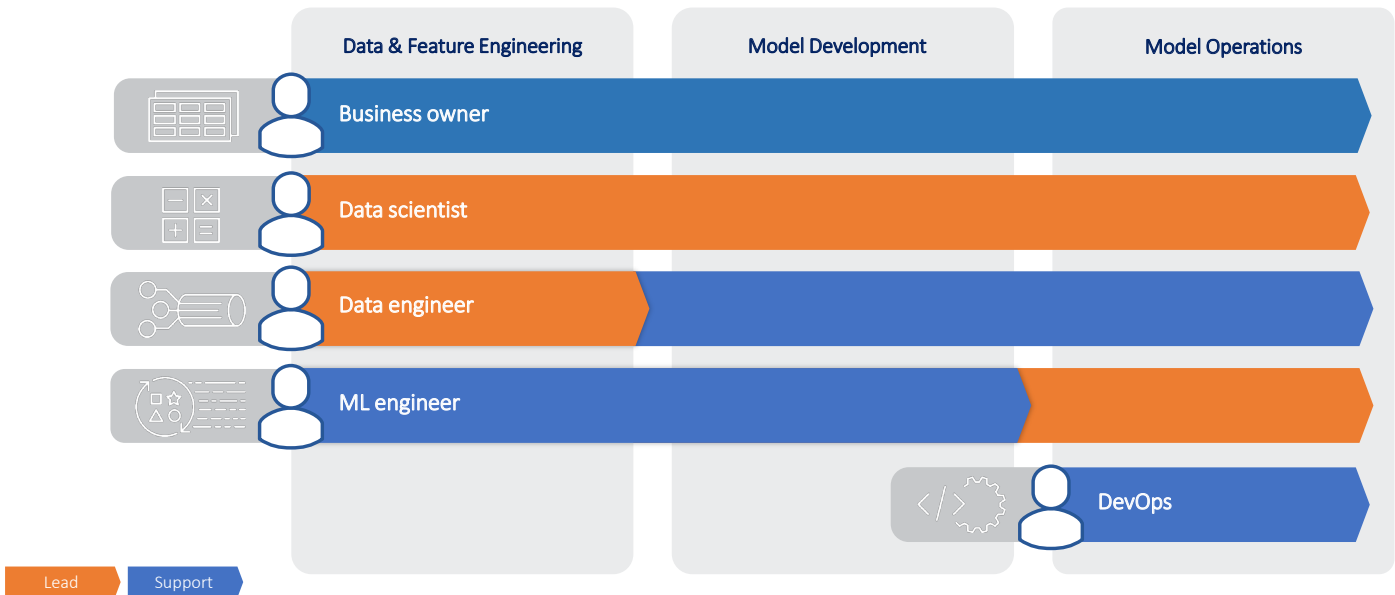
The machine learning lifecycle entails lots of iterations for re-training and re-development.

Team Roles

Data science teams depend on the tight collaboration of all the stakeholders described above as they execute the ML lifecycle. Data scientists lead the overall lifecycle, with oversight from the business owner. Along the way they collaborate with data engineers, ML engineers, and DevOps engineers. Business owners should coach the data scientist and overall team during each stage.

Figure 2 illustrates these team roles in the ML lifecycle.

Figure 2: Team Roles in the Machine Learning Lifecycle



Challenges

Modern enterprises depend on effective machine learning to stay sharp in a dynamic world. They cannot afford to drain their sales inventories, miss fraudulent transactions, or let container ships break down. Their data science teams need to build ML models that generate correct outputs, then operationalize those outputs to help applications and decision makers act on them in a timely and reliable fashion. However, three primary challenges stand in the way: scalability, reproducibility, and governance.

- > **Scalability.** Data science teams struggle to remove silos and integrate large, growing datasets, as well as models and code. They need scalable processing infrastructure to train, deploy, and change out many models, leveraging scalable processing, in order to maintain accuracy.
- > **Reproducibility.** Data science teams struggle to avoid reinventing the wheel with each project. Repeatable processes and model results can only consistently be achieved when teams share and reuse one another's features, models, and code.
- > **Governance.** Increasing public scrutiny and regulatory pressure force enterprises to reduce the risk of inaccurate, biased or opaque models, as well as the mishandling of personally identifiable information (PII).

Machine learning challenges include scalability, reproducibility, and governance.

Optimizing the Machine Learning Lifecycle and MLOps on a Cloud Data Platform

Cloud data platforms have recently emerged as a new option to manage the ML lifecycle and address the challenges of scalability, reproducibility, and governance not only for models but also for data.

Market evolution

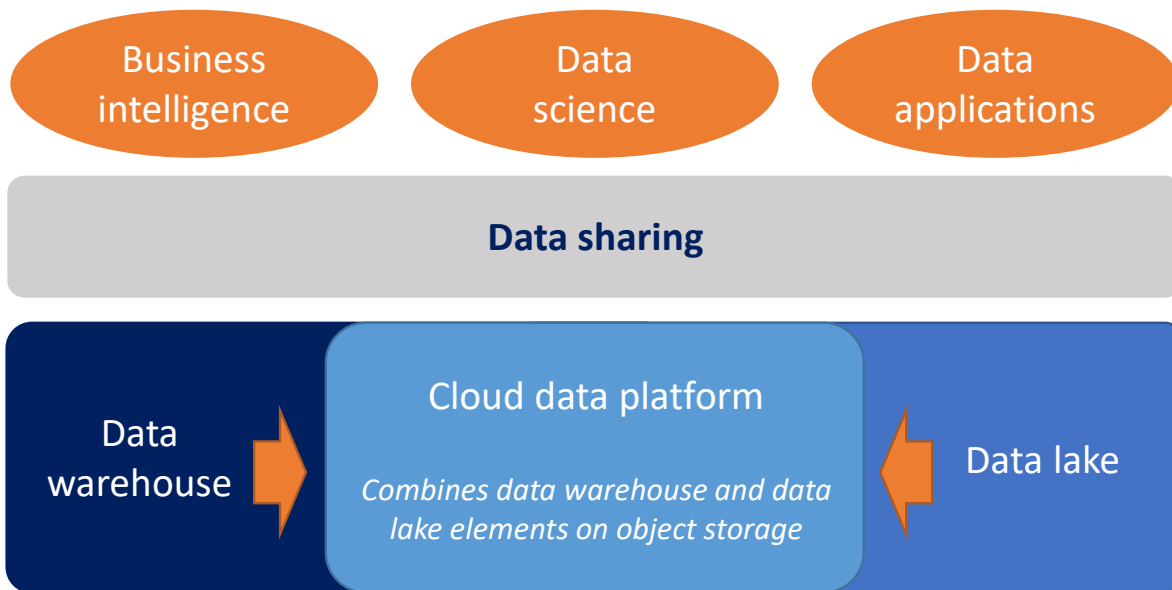
Cloud data platforms are the convergence of data lakes and data warehouses, assisted by cloud elasticity. As data types and sources proliferated, enterprises embraced the flexibility and scalability of cloud object stores. Legacy data warehouse vendors adopted object stores and added functionality to support all analytics needs, from basic reporting with SQL to data science. Meanwhile new cloud data platforms emerged with this combined functionality at its core.

New platforms such as Snowflake provide SQL data warehouse structures, modeling capabilities, and native support for other languages to support data science tools. They make it easier to support a

variety of workloads, data types, and volumes, even across clouds, and help share data both within and between enterprises.

The cloud data platform provides a common repository to support the overlapping worlds of BI, data science, and data applications, all supported by data sharing. Figure 3 illustrates the convergence of the data lake and data warehouse into the cloud data platform.

Figure 3. The Data Lake and Data Warehouse Constructs Converge into the Cloud Data Platform



The cloud data platform evolved further to support and help automate the ML lifecycle. Cloud providers now bundle ML capabilities with their cloud data platforms. For example, Snowflake’s partnership with **Anaconda** enables data science teams to package Python-based ML algorithms and manage the ML lifecycle in the Snowflake Data Cloud. This helps them tap the value of open-source libraries while addressing security and reliability requirements.

Cloud data platforms evolved to support and help automate the machine learning lifecycle.

Advantages of the Cloud Data Platform

Implemented and managed well, the cloud data platform can serve as a standardized ML model factory that resides on a foundation of governed data. Let’s consider its advantages.

Advantages of both ML platforms and cloud data platforms

Cloud data platforms now provide data science teams with many elements of an ML platform—and integrate them with their data environment. Similar to ML platforms, cloud data platforms offer lifecycle speed, production scale, and governance. They tend to favor users with more programming and statistical skills because they offer less automation than ML platforms.

Lifecycle speed. Some cloud data platforms offer a graphical interface and automated workflows that accelerate the ML lifecycle for basic use cases. The data scientist might follow prompts to discover, ingest, and cleanse relevant historical datasets for their use case. They might visualize the characteristics of that data, then—with minimal scripting—define their features, specify their target variable, and label historical outcomes. They can select one or more recommended ML techniques, then follow prompts to train their algorithms against historical data, measure their accuracy, and adjust their features. Once they complete the training, the cloud data platform helps automate the packaging, testing, approval, and deployment of ML models in a production environment. It also helps data science teams reuse models and datasets. Cloud data platforms also should offer a code-oriented approach, often leveraging ecosystem tools, to help more skilled data scientists and developers build specialized models.

Scale of production. This guidance and automation enable data scientists to build, train, select, deploy, and reuse many models in parallel. They can visualize outputs for multiple algorithms in a single dashboard, then drill into different algorithms as needed to compare their accuracy and fitness for production environments. Once they take a model live, they can continue to run “challenger” models in the background, and swap them into production to remediate model drift, performance issues, or signs of bias. Much like an ML platform, the cloud data platform becomes like a high-scale factory with many assembly lines, which together produce a higher quantity and quality of ML models. It supports all these steps with elastic cloud compute resources that scale up and down as needed to support processing workloads.

Model governance. A cloud data platform also can help assess the quality of training data, for example by flagging duplicates and outliers, cleansing that data, and identifying PII to mask. It provides the data scientist, ML engineer, and governance officer with centralized views of model lineage, accuracy, and potential bias, as well as their handling of PII. It could also catalog and display models, along with their lineage, approval history, and metadata, to assist governance and reuse. Data science teams can configure dashboards and alerts that identify signs of production model drift, bias, or PII mishandling.

Like the ML platform, the cloud data platform offers lifecycle speed, production scale, and governance.

Additional advantages of the cloud data platform

The cloud data platform offers the additional advantages of data consolidation, cross-functional collaboration, and integrated workflows.

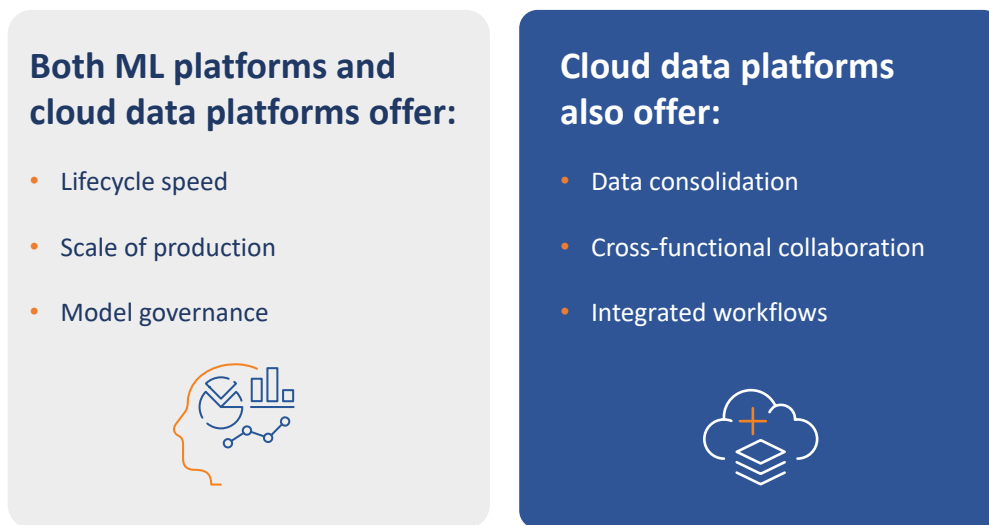
Data consolidation. Data science teams can discover, explore, and transform historical data; join with external data sets through data sharing; then define features; and train models—all within the cloud data platform. They also might operationalize models and monitor their production performance and accuracy within the platform. By applying the ML lifecycle to the same source of truth as the rest of the business, they reduce the risk of contradictory data copies. This improves data quality, protects sensitive data, and makes governance easier because data stewards and other governance officers have just one platform to inspect and monitor. Some cloud data platforms also help share data between enterprises, across clouds or regions, to further enrich ML training and production datasets. In the future, look for cloud data platforms to take the next step of combining data cataloging with model cataloging, so that governance officers can centralize their control of both datasets and ML models.

Cross-functional collaboration. A cloud data platform offers a shared workspace for data scientists, data analysts, data engineers, ML engineers, and DevOps engineers to collaborate. They can share the same datasets, models, and analytical outputs, while still using their development tool of choice. For example, a data scientist can review datasets jointly with a data analyst and learn from their experience with BI projects before finalizing their ML model features. Once they produce a final ML model, the data scientist can share it with the data analyst for reuse in their BI dashboards and reports. They also can help in measuring model accuracy by comparing model results to actual outcomes within the same platform.

Integrated workflows. This consolidation, sharing, and collaboration make it easier for teams to integrate ML model outputs with production workflows. For example, a data scientist might train an ML-driven chatbot on customer service records, then hand it to a DevOps engineer to insert into a customer service portal. An ML engineer and DevOps engineer might operationalize an ML model that automatically recommends customer upsell opportunities to sales reps through notifications in **Salesforce** or **Slack**. Integrations like these are easier because the ML and operational applications run on the same data store.

A cloud data platform also offers data consolidation, cross-functional collaboration, and integrated workflows.

Figure 4 summarizes these advantages of managing the ML lifecycle on the cloud data platform.

Figure 4. Advantages of Managing Machine Learning on the Cloud Data Platform

The risk of lock-in

As with other analytics projects, consolidating the ML lifecycle on a cloud data platform can create the risk of lock-in. Enterprises should select a cloud data platform that provides open access to various data formats using standard protocols, application programming interfaces (APIs), and tools. The platform also should provide open standards for the use of their preferred programming languages. By investing in a platform like this, and avoiding proprietary future enhancements, enterprises can keep their data, tools, and code both portable and interoperable to address changing requirements.

Next Steps

Machine learning, the most popular type of AI, provides enterprises with a much-needed weapon to compete in today's complex, fast-changing business environment. While data science teams can start ML initiatives with a piecemeal approach, they need to standardize on a platform in order to achieve the necessary scalability, reproducibility, and governance. A cloud data platform—like a dedicated ML platform—enables them to accelerate the ML lifecycle and MLOps, produce more models, and govern them. A cloud data platform goes further to help data science teams consolidate data, foster collaboration across teams, and insert ML models into enterprise workflows.

Effective data science teams should adopt these guiding principles as they consider standardizing their ML lifecycle and MLOps on a cloud data platform.

- > **Seek the right balance between standardization and customization.** Standardizing steps in the ML lifecycle improves productivity. But your team will still need to customize certain projects, for example by procuring hard-to-find datasets, or creating specialized models. Be sure to select and configure a cloud data platform that offers unfettered access to the ML ecosystem of libraries, notebooks, and tools. Also be sure your data science team has the creativity, domain knowledge, and familiarity with these tools to customize where needed.
- > **Evaluate cloud data platforms based on their support of the ML lifecycle.** As you start and then scale up your ML projects, evaluate the ability of your existing cloud data platform—or potentially a new one—to support your ML lifecycle. Ask vendors about their current capabilities and roadmap to help you automate and standardize the many aspects of building and operating ML models. Make this a “must-have” criterion for evaluating cloud data platforms.
- > **Plan with future requirements in mind.** Expect your supply of data to grow, along with business demand to use it. This means you need a cloud-based platform that can scale elastic infrastructure to accommodate new data sources, use cases, ML models, use cases, and users, much like a factory floor with multiple assembly lines. It also means you need to keep recruiting new talent, and building new skills with your existing data science team.

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 competitive intelligence and market positioning guidance to software vendors to improve their go-to-market strategies.

Eckerson Group is a global research, consulting, and advisory 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!**



About the Sponsor

Snowflake delivers the Data Cloud—a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake’s platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. The Snowflake Data Marketplace is a key data sharing vehicle inside the Data Cloud. Snowflake customers can access an ever-growing ecosystem of data from SaaS & commercial data providers. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. Learn more at [snowflake.com](https://www.snowflake.com).

