REPORT REPRINT

Cloudera illuminates its machine learning strategy

MARCH 19 2020

By Krishna Roy

The vendor is looking to provide organizations that have invested (or plan to invest) in the Cloudera Data Platform with enterprise-grade machine learning. We take a look at what Cloudera has delivered so far and the deliverables set to come this year.

451 Research®

Introduction

Cloudera entered the vibrant but increasingly crowded machine learning sector in 2017 with Cloudera Data Science Workbench, which is integrated with the Cloudera Data Platform to provide enterprise-ready capabilities such as data security and governance. While CDSW continues to be available for on-premises deployment, it has been joined by Cloudera Machine Learning – the next evolution of CDSW. Cloudera Machine Learning it is a native hybrid and multicloud service in-line with the company's broader corporate mission to offer enterprises a native hybrid cloud experience. Additionally, Cloudera offers a portfolio of services through its Professional Services and Fast Forward Labs teams to assist organizations in various aspects of their machine learning initiatives. Cloudera's ML services, which include strategic advice and custom application deployment, are offered as a complement to its software offerings, which have a busy roadmap ahead.

451 TAKE

Cloud-based AI platforms are key to machine learning adoption – 56% of companies are currently using these tools, and 48% say it is their primary strategy, according to 451 Research's latest Voice of the Enterprise: AI & Machine Learning survey. Cloudera Machine Learning has therefore hit the market at the right time. The company's ability to provide data science teams with software to accelerate, manage and scale machine learning workflows while enabling IT and MLOps teams to secure and govern these workflows – in addition to offering consulting (in recognition that successful machine learning is as much about people and processes as it is about software) – singles it out from the crowd. However, baking machine learning into CDP is a double-edged sword – it means organizations not currently using Cloudera's platform are likely to need to invest in another data platform, which they may be reluctant to do.

Context

Data science is a big focus for Cloudera as it seeks to enable enterprises to deploy machine learning models into production in a rapid, repeatable and easy-to-scale way. The vendor's initial offering to enable this was CDSW, an on-premises offering based on its March 2016 acquisition of data science tool provider Sense.io. Cloudera Machine Learning for the Cloudera Data Platform is the next evolution of CDSW – it's a fully cloud-native machine learning service for hybrid and multicloud deployment. The company unleashed it in September 2019.

Packaged machine learning software natively integrated into CDP in order to bring essential enterprise-ready capabilities, such as shared metadata, schema, security and governance (also shared by the rest of the components on CDP), is one aspect of the vendor's machine learning strategy. However, it recognizes that successful machine learning in the enterprise (and at scale) isn't just a matter of having the right software tooling. It requires transforming skills and processes, which are the preserve of the firm's Fast Forward Labs and Professional Services teams. The former came from Cloudera's acquisition of applied ML research firm Fast Forward Labs in September 2017.

Cloudera notes that it has hundreds of its largest enterprise customers using its wares for data science and machine learning, with 30% deployed in the public cloud and the largest deployments rolled out to 500 users.

Strategy

Cloudera Machine Learning is available as an 'experience' (or fully containerized workspace) in CDP alongside six others: Data Warehouse, Data Catalog, Data Hub Clusters, Workload Manager, Replication Manager and Management Console. These are part of the company's positioning of CDP as an 'enterprise data cloud for all enterprise needs.' Users are provided with a role-based interface that surfaces different capabilities, including different data access policies, depending on who is using it.

Cloudera Machine Learning is primarily sold through a direct sales model, using a consumption-based pricing model. CDSW is also sold directly to customers, although the company is using partnerships to build an indirect sales model for both offerings.

Cloudera Machine Learning's ability to support hybrid deployments and multiple clouds rests on utilizing Kubernetes and containers. It is currently available on Amazon Web Services and in the Microsoft Azure Marketplace – with availability on Google Cloud Platform coming later in 2020.

In keeping with many data science offerings, Cloudera Machine Learning and CDSW don't house proprietary machine learning algorithms, but hook into libraries and integrated development environments (IDEs) in order to provide one environment to build, train, deploy and operationalize ML models in a scalable manner. All Python, R and Scala ecosystems are supported, in addition to RStudio, PyCharm and Jupyter Notebook IDEs.

Furthermore, Cloudera recognizes that machine-learning-driven data science is a team sport, requiring functionality for user personas other than data scientists. The company is largely addressing data scientists, data engineers, ML engineers and MLOps personas, but has capabilities under development to cater to nonexperts.

Cloudera Machine Learning's purview is analyzing and exploring data; training models by running, tracking and comparing them; deploying and monitoring models as APIs; and managing shared resources. It houses features for data collection and verification and feature extraction. Facilities to configure the environment required to train and deploy machine learning models, including the necessary CPUs and GPUs, are also included. Additionally, isolated yet elastic compute and storage, process management tools, and machine resource and infrastructure monitoring are part of the offering.

CDSW contains the same functionality, and is how the company's machine learning tools are deployed for Cloudera Enterprise Data Hub and the Hortonworks Data Platform. For on-premises deployments on the new Cloudera Data Platform, CDSW sits alongside CDP Data Center, which is essentially the company's more traditional on-premises Hadoop offering. At present, CDSW implements a sidecar architecture, which means it's deployed on Kubernetes, on edge nodes sitting next to a Hadoop cluster.

Cloudera Machine Learning, in contrast, doesn't require Hadoop, which is a key difference between the two. CDSW doesn't support autoscaling either, which is a key feature of Cloudera Machine Learning. Furthermore, Cloudera Machine Learning is expected to be available on-premises in its fully cloud-native form factor, including support for autoscaling, later in 2020. The offering will be called Cloudera Machine Learning for CDP Private Cloud.

Continually improving the user experience for data science teams is one focus for future investment. Bringing business users more easily into the data science mix is another. It will involve the introduction of a visual application builder to craft visualizations and templates that business users can understand, tied in with the machine-learning-driven predictive analytics. This forthcoming offering will utilize assets from Arcadia Data, which Cloudera acquired in September 2019.

Supporting machine learning in production is another significant investment area. It will be addressed by the delivery of a model catalog, which will use the existing data catalog in CDP as its foundation. Cloudera's existing data catalog is built on the Apache Atlas open data governance and metadata framework. The company is extending it so that it will be able to support the governance of large volumes of ML models, in order to create a model catalog.

Finally, Cloudera plans to deliver a software developer kit to read and track all metrics associated with each model, in order to better support machine learning initiatives once in production. The company's SDK will include capabilities to track all inputs and outputs to a model for model accuracy, as well as other capabilities for model lineage and 'explainability,' so that organizations can see where a model came from and what data they used, for instance. The metrics will be housed in a back-end metrics store.

Competition

Cloudera's strategy to provide native integration with a data platform to support machine-learning-driven predictions is similar to that of Databricks, in our opinion. However, Cloudera cites AWS SageMaker and Domino Data Labs as high-level competitors, noting that its ability to provide a shared data and multifunction analytics platform integrated with machine learning makes it more comprehensive in nature. AWS has the ability to offer something similar, but it would require buying other products, including Glue for data integration.

Cloudera views Domino Data as a point product because it isn't natively integrated with a data platform, instead running across multiple ones. This strategy has been adopted by the majority of data science platform vendors, in recognition that many enterprises have already made investments in data platforms and don't want to purchase another.

It is also worth noting that IBM is a partner as well as a potential competitor – Big Blue has a partnership with Cloudera to sell CDSW, but also competes in the broader machine learning sector. RStudio is another potential rival – while Cloudera supports RStudio's IDE, RStudio is also seeking to provide enterprises with a corporate-ready machine-learning-driven data science platform.

As previously noted, machine-learning-driven data science is a crowded sector. Google and Microsoft are other dominant vendors, and we suspect Cloudera will encounter them more frequently with the availability of Cloudera Machine Learning on their respective clouds.

SAS Institute, TIBCO, RapidMiner and Dataiku are other vendors targeting a similar audience. Furthermore, we wonder if Cloudera's foray deeper into MLOps will take the company into competition with MLOps specialists such as Seldon, HydroSphere.io, Algorithmia, Datakitchen, Datatron and Metis Machine.

SWOT Analysis

STRENGTHS Cloudera's credentials as a leading data platform provider enable it to offer enterprises a data-platform-centric approach to data science that is underpinned by consulting and professional services for differentiation.	**WEAKNESSES** Do we want yet another data platform specifically for data science? The answer will be 'no' for some organizations.
OPPORTUNITIES Existing customers represent the low-hanging fruit for machine learning, which, owing to pre-existing integration with CDP, makes the upsell process easier.	**THREATS** The machine-learning-driven data science platform sector is full of vendors (451 Research is tracking more than 35 vendors in this category), so Cloudera is operating in a crowded space.