Market Insight Report Reprint

Databricks adds to Lakehouse Platform with machine learning and data catalog enhancements

June 24 2021

by Matt Aslett

Data science and analytics platform provider Databricks continues to expand its Lakehouse offering, and recently added a new Databricks Machine Learning platform, as well as a new data catalog and data integration service. The company also unveiled a new open source initiative designed to securely enable data sharing across and between enterprises.

451 Research

S&P Global
Market Intelligence

Introduction

Databricks has come a long way since it was founded in 2013 to build a business around the Apache Spark data processing engine. Although Spark is still at the heart of Databricks' portfolio, the company's cloud service has been expanded over the years through both acquisition and internal development. Now known as the Databricks Lakehouse Platform – based on the concept of the data lakehouse, which is designed to combine the best of data-warehousing and data-lake architectures – the company's cloud service has seen a number of recent enhancements related to machine learning, data management and data sharing.

Machine learning has always been a key capability thanks to the underlying machine learning capabilities of Apache Spark, but with the launch of Databricks Machine Learning, the company has pulled together its existing capabilities into a dedicated offering targeted at data engineers, with the addition of automated machine learning and feature store functionality. In relation to data management, Databricks has also introduced Delta Live Tables for data integration pipeline management, as well as a new data catalog. The latter is underpinned by Delta Sharing, which is the company's latest open source initiative, designed to securely enable data sharing across and between enterprises.

THE 451 TAKE

Databricks continues to build out its Databricks Lakehouse Platform en route to a probable initial public offering, and now offers a broad and deep portfolio of cloud services that address data integration and transformation, SQL-based analysis, data science, and machine learning. Based on a variety of interdependent open source projects and related cloud services, the portfolio has the potential to be somewhat confusing to the uninitiated. In that context, pulling the machine learning capabilities together in the form of a single interface for machine learning engineers is a good move that complements the similar Databricks SQL interface for analysts. The addition of Unity Catalog will also enable Databricks customers to better manage their various data assets across the Databricks Lakehouse Platform estate, while Delta Sharing is an interesting addition, albeit in the early stages, that could play an important role in enabling vendor-neutral data sharing.

Details

Having expanded and expounded on the concept of the data lakehouse in 2020, Databricks has now adopted the terminology to describe its cloud-based data services. The Databricks Lakehouse Platform is designed to combine the low-cost storage and agility advantages of the data lake with the data structure and data management features of the data warehouse to support multiple workloads (including data integration and transformation, SQL-based analysis, data science, and machine learning).

The Databricks Lakehouse Platform is based on a number of open source projects (including Apache Spark for data processing, Delta Lake for structured data management, MLflow for machine learning lifecycle management and Redash for SQL-based analytics), with additional security and administration capabilities, delivered as a cloud service on Microsoft Azure, Amazon Web Services and Google Cloud.

At its recent Data + AI Summit virtual event, the company announced a number of new additions to the Databricks Lakehouse Platform focused on machine learning, data management and data sharing.

Machine learning has always been an integral part of the Databricks value proposition thanks to the capabilities of Spark and MLflow, but the new Databricks Machine Learning platform is designed to provide a dedicated interface targeted at machine learning engineers and to enable them to manage the machine learning lifecycle – from experimentation and development to deployment and management in production.

Databricks Machine Learning includes the company's Collaborative Notebooks for data scientists and the machine learning runtime for performing machine learning jobs created in multiple frameworks and tools, as well as the Managed MLflow service for model deployment and management and model serving functionality. Two new capabilities are Databricks AutoML, to accelerate model development with automated model generation and training (with in-built explainability), and Databricks Feature Store, to facilitate reuse of model features with lineage and usage tracking.

In addition to the Databricks Machine Learning interface, Databricks also launched new data management capabilities – specifically the Unity Catalog and Delta Live Tables. The latter is a cloud service for creating and managing data integration and transformation pipelines on the Delta Lake structured transactional layer. The service is designed to simplify pipeline creation by enabling users to specify desired outcomes, with Delta Live Tables automatically creating data transformation and validation instructions.

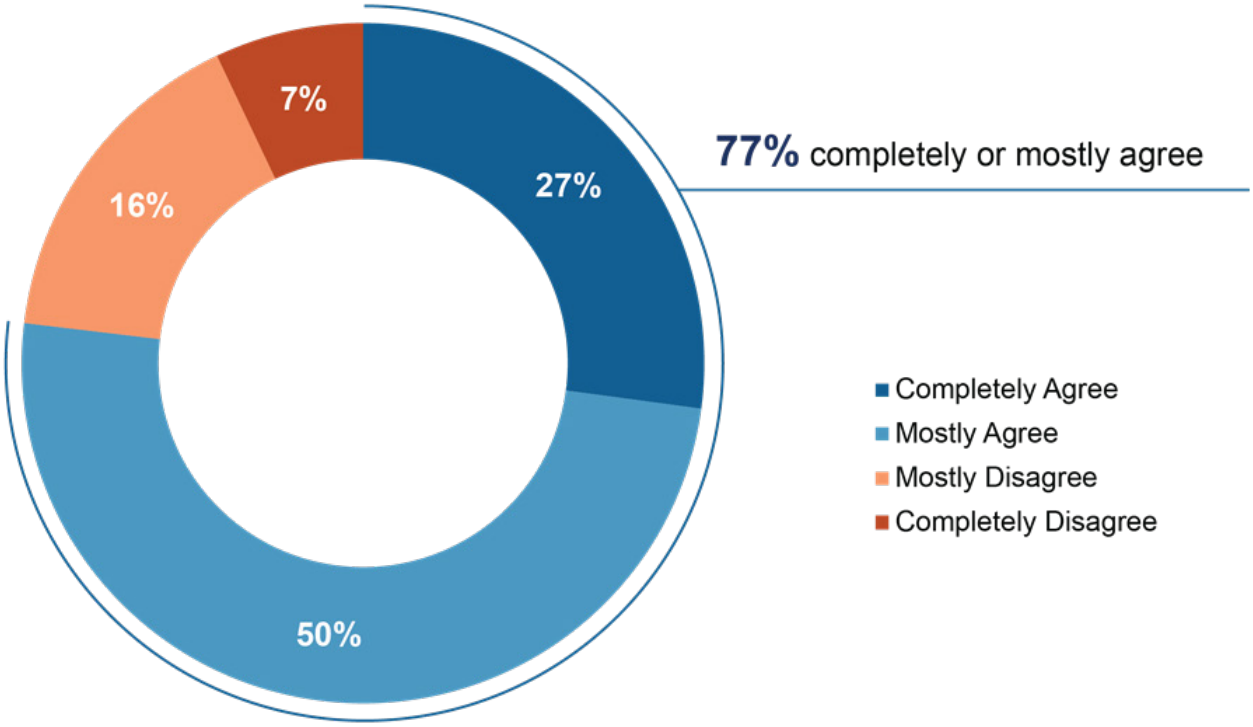
The Unity Catalog, meanwhile, is designed to operate across multiple clouds and integrate with external data catalogs (from Databricks partners such as Alation, Collibra, Immuta and Privacera). In addition to complementing these existing catalogs with a view across a user's Databricks Lakehouse Platform estate, the company notes that Unity Catalog is a platform for governance of all assets (including tables, views, files, dashboards and models) – not all of which will be covered by other catalogs.

Unity Catalog is based on Databricks' new Delta Sharing open source initiative. Part of the Linux Foundation's open source Delta Lake project, Delta Sharing is an open protocol for sharing data across and between enterprises. It enables organizations to share data in the Apache Parquet and Delta Lake formats without copying it, via an open protocol that can be implemented in SQL, visual analytics tools and programming languages – avoiding reliance on a single vendor, product or service.

A number of companies have signed up as supporters of the Delta Sharing initiative, including technology vendors such as Amazon Web Services, Google Cloud and Salesforce's Tableau, as well as data providers such as Nasdaq, NYSE, Foursquare and S&P Global (451 Research's parent company).

The ability to share data with partners, suppliers and even customers is becoming increasingly important to enterprises. More than three-quarters (77%) of respondents to 451 Research's Voice of the Enterprise: Data & Analytics, Data Management & Analytics 2020 agreed that the ability to share and collaborate more closely on data with partners, suppliers or customers is becoming increasingly critical to their organization.

Importance of Ability to Share Data With External Collaborators



Q: To what extent do you agree or disagree? – 'The ability to share and collaborate more closely on data with our partners, suppliers or customers is becoming increasingly critical to my organization.'

Base: All respondents, abbreviated fielding (n=398)

Source: 451 Research's Voice of the Enterprise: Data & Analytics, Data Management & Analytics 2020

CONTACTS

The Americas
+1 877 863 1306
market.intelligence@spglobal.com

Europe, Middle East & Africa
+44 20 7176 1234
market.intelligence@spglobal.com

Asia-Pacific
+852 2533 3565
market.intelligence@spglobal.com

www.spglobal.com/marketintelligence