Cloudera opens up about going all in on open source

AUGUST 15 2019

By James Curtis

Having completed its purchase of Hortonworks in January, the vendor has been working to not only reconcile products between the two firms but also determine which open source business model to adopt. That appears to be settled now, with Cloudera committing to go 100% open source, closely following the Red Hat model.

451 Research®

Introduction

Its acquisition of Hortonworks officially closed in January, and Cloudera has been on a mission to integrate the two firms ever since. With product reconciliations mostly addressed, however, the company still needed to resolve which open source model it would adopt. That has now been settled, with Cloudera announcing that it will adopt a full, 100% open source model. What this means is that former propriety Cloudera products will now be open-sourced as the vendor marches toward an open source model patterned after Red Hat's, albeit with some caveats.

451 TAKE

Since Cloudera bought Hortonworks, it has been fielding questions about how it planned to reconcile the product portfolios of the two companies and which open source model would be adopted. With the forthcoming release of the Cloudera Data Platform, many of the product decisions have been addressed. As for the open source model, Cloudera initially noted that it would support both models (propriety mixed with open source and 100% open source), but details were scarce. The fact that Cloudera has announced that it's going down the path of 100% open source, modeled after Red Hat, makes a lot of sense, particularly because existing customers likely chose Hortonworks for its approach to open source. Cloudera and Hortonworks both have deep legacies with open source, given that there are many individuals within the combined entity that were early creators or contributors to many of the open source projects Cloudera plans to use. So in that sense, the vendor is on solid footing supporting open source and can now focus on delivering product unification.

Context

In October 2018, Cloudera announced that it had reached an agreement to buy Hortonworks in a $2.1bn stock swap deal that officially closed in January. Prior to the purchase, Cloudera and Hortonworks were rivals of sorts, although both shared a collective interest in and commitment to Apache Hadoop and related projects that would become known as the broader Hadoop ecosystem.

Spun off of Yahoo in late 2011, Hortonworks' commitment to Hadoop resulted in the release of its own Hortonworks Data Platform, with a resolute commitment to a 100% open source software model. Cloudera, on the other hand, was founded in 2008 and by 2009 had released its first commercial distribution of Apache Hadoop. While Hortonworks took a 100% open source approach, Cloudera adopted a hybrid strategy, mixing open source with its proprietary tooling consisting of, for example, Cloudera Manager, Cloudera Navigator, Cloudera Director and, later, Cloudera Data Science Workbench.

Licensing strategy

With the two firms having different go-to-market strategies, it made sense for Cloudera to ultimately choose one of the approaches as its primary model. Management notes that in the leadup to a 100% open source decision, the company leaned heavily on the proven strategy implemented by Red Hat, although there will be some differences in how Cloudera's open source policy will play out.

The company plans to license all product sources based on Open Source Initiative-approved licenses, specifically as either Apache Software Foundation Licenses, Version 2 (ASL) or the Affero General Public Licenses, Version 3 (AGPL). Existing projects will retain their present license structure, licensed as either ASL or AGPL, as there are no plans to change existing license agreements. However, products yet to be open-sourced – Cloudera Manager, Cloudera Navigator and Cloudera Data Science Workbench, for instance – will be licensed as AGPL.

Cloudera notes its full commitment to open source and will continue to follow the practice of making contributions to the upstream source first, including to any new open source projects. Access to binaries, however, will only be available from Cloudera and will require a subscription agreement with the company to access, which is a departure from how it previously distributed its binaries. The reason for putting binaries behind a paywall is that it provides some level of protection for the vendor. The binaries contain Cloudera-specific IP that turns the many disparate open source projects into an enterprise-grade functioning system.

Questions remain, however, about when these changes will occur and, of course, what the impact will be on existing customers. Starting in September, a subscription agreement with Cloudera will be required to access the binaries, impacting primarily new clients. For existing customers under subscription, there will be little to no impact, other than a small 'click-through' exercise to transition over, according to management. For those organizations not on subscription, the current releases will continue to be accessible, although functionality will diverge over time given the new binaries. For forthcoming products expected to be open-sourced, the company has targeted February 2020 for getting them under open source license.

Competition

Because Hortonworks is now owned by Cloudera, the company's most direct competition has come from the distributed-processing framework vendors. These include MapR, which was recently acquired by HPE, so it may still be a rival in the short term but likely not in the long term as the buyer works to integrate MapR's technology within its portfolio; Qubole, which offers a data platform that drives a collective-engines approach based on open source processing tools and is available only on the cloud; Databricks with its cloud-based analytics platform powered by Apache Spark; and Cazena with its managed data lake as a service. Cloudera, however, seems to be setting its sights on vying with established data management providers – Oracle, Microsoft, IBM, SAP and Teradata, as well as Amazon Web Services and Google Cloud Platform (GCP) – mostly because it has multiple products built on its platform covering data science to streaming to data warehousing.

Oracle has been actively positioning its core database functionality via 'autonomous' capabilities and has its Oracle Autonomous Data Warehouse and Oracle Big Data Cloud. Microsoft Azure offers HDInsight (based on the Hortonworks Data Platform), Azure SQL Data Warehouse and Azure Databricks. Furthermore, Microsoft is embedding Apache Hadoop and Apache Spark as part of SQL Server 2019. GCP features some competing services such as Cloud Dataproc and BigQuery, while AWS's offerings include Amazon EMR, Amazon Kinesis and Amazon Redshift.

Another contender is SAP's Cloud Platform Big Data Services (formerly Altiscale), which provides Apache Hadoop, Apache Spark, Apache Hive and Apache Pig services; SAP HANA can also be included here. Teradata provides its Vantage offering. IBM recently launched an updated version of its Db2 database, which includes built-in support for data science tools and frameworks for the development of artificial intelligence applications, and functions as a single SQL engine code base for the multiple products and services that share the Db2 brand. Still others include Actian's VectorH (Hadoop) and recently released Actian Avalanche, a managed cloud service based on VectorH, along with other data-warehouse vendors such as Snowflake, Pivotal's Greenplum, Yellowbrick Data, Exasol and Micro Focus' Vertica.

SWOT Analysis

STRENGTHS

Cloudera's reputation, as well as its commitment to and support of open source software, are well-known, given its significantly deep technical resources that include individuals that were early creators of many of the open source projects the company uses.

WEAKNESSES

A bit of mixed messaging initially may have confused some as Cloudera noted that it would support both companies' open source, go-to-market strategies before eventually settling on the approach most closely associated with Hortonworks and modeled after Red Hat.

OPPORTUNITIES

The adoption of a 100% open model gives Cloudera a strong, clear message going forward and likely puts at rest some possible concerns from existing Hortonworks customers who may have had concerns about certain projects or platform capabilities going forward.

THREATS

The competitive field has significantly grown to now include a strong number of well-established data management providers, many with varied and complex portfolios.