REPORT REPRINT

# Cloudera updates DataFlow for IoT data processing at the edge

**MARCH 28 2019**

**By Matt Aslett**

The company has updated the DataFlow software it acquired along with Hortonworks for managing data in motion. Specifically, it has added capabilities for edge data collection, routing and management.

## Summary

Cloudera has updated the DataFlow software it acquired along with Hortonworks for managing data in motion. Specifically, it has added capabilities for edge data collection, routing and management.

## 451 TAKE

Data from 451 Research's Voice of the Enterprise: Internet of Things, Workloads and Key Projects shows that analytics is critical to the success of Internet of Things (IoT) projects and that processing of IoT data is increasingly being carried out at the edge. With DataFlow, Cloudera already had a differentiated offering for processing and analyzing data in motion. Cloudera Edge Management adds the ability to develop, deploy and monitor data-processing applications at the edge, which is likely to be a fundamental enabler of successful IoT projects going forward. We noted when Cloudera announced its intention to purchase Hortonworks that DataFlow was cited as one of the advantages of the deal from a product-synergy perspective. Therefore, it's no surprise to see the company moving fast to deliver integration between the newly renamed Cloudera DataFlow and Cloudera's CDH distribution of Apache Hadoop.

## Context

As we noted in January, obtaining the DataFlow platform for managing data in motion was one of the primary benefits of Cloudera's acquisition of former rival Hortonworks. As such, it was no surprise to see the company quickly rebranding the product Cloudera DataFlow and taking the necessary steps to ensure that DataFlow is supported on Cloudera's CDH distribution of Apache Hadoop, as well as integrated with its various software offerings.

The latter has been enabled by the launch of Cloudera Flow Management, one of what are now four core components of the Cloudera DataFlow platform, alongside Stream Processing, Streaming Analytics and another new offering: Cloudera Edge Management. As a reminder, what is now Cloudera DataFlow has always been based on several open source components, including Apache NiFi for data flow management, Apache Kafka for pub/sub messaging, Apache Storm for streaming analytics, Apache MiNiFi for edge processing, and integration with Apache Ambari and Apache Ranger for management and security.

While Cloudera Flow Management is a new product SKU, it is not really a new addition to the platform but represents the delivery of formal Cloudera support for Apache NiFi as well as NiFi Registry. This includes the ability to install, manage and monitor NiFi running on CDH (versions 5 and 6) clusters using Cloudera Manager, as well as support for NiFi processors for Cloudera-developed data processing and storage engines, including Apache Impala and Apache Kudu.

Cloudera Edge Management is truly a new addition to the DataFlow platform and sees the company bolstering its capabilities for data collection, routing and monitoring at the edge. The offering builds on Apache MiNiFi, a pared-down, agent-based version of NiFi that was added to DataFlow in late 2016, but also includes NiFi Registry as well as a new capability called Edge Flow Manager. Edge Flow Manager adds the ability to monitor and manage data flows involving MiNiFi agents – including multiple classes of agents – supporting the collection of data from edge devices, as well as the ability to push changes and intelligence, including machine learning models, from central data-processing resources to the edge.

Additionally, Cloudera DataFlow includes Cloudera Stream Processing, which is based on a combination of Apache Kafka and Streams Messaging Manager (which was introduced in 2018 and provides support for managing and monitoring Apache Kafka clusters), as well as Schema Registry (which serves as a shared repository of schemas). The fourth component of Cloudera DataFlow is Cloudera Streaming Analytics, which is based on the combination

of Apache Storm, Apache Kafka Streams and Steaming Analytics Manager, which was introduced in 2017 and provides a code-free environment for developing streaming applications and leverages the Druid analytics engine to enable the creation of analytics dashboards and reports.

## Competition

The combination of functionality delivered via Cloudera DataFlow means not only that it has a variety of potential competitors, but also that it is a differentiated offering, which made it a key driver for the purchase of former rival Hortonworks. Among the potential competitors are StreamSets, which was previously the primary partner offering positioned as an alternative to DataFlow for Cloudera customers, while we also see some overlap with Striim's streaming data-integration and -intelligence offering.

As the primary commercial company behind Kafka, Confluent must also be considered a rival, given Cloudera DataFlow's Kafka management capabilities, while MapR offers Event Store for Apache Kafka as part of the MapR Data Platform. Additionally, IBM offers the Kafka-based Event Streams for IBM Cloud; AWS offers Amazon Managed Streaming for Kafka; Microsoft offers Kafka for Azure HDInsight; and Google offers Google Cloud Pub/Sub, and has partnered with Confluent Cloud on GCP. Other potential contenders include Databricks as the primary supporter of Apache Spark; Ververica (which was known as data Artisans until its recent sale to Alibaba) with its support for Apache Flink; and Streamlio, which has assembled a streaming platform based on the combination of Apache Pulsar for messaging, Apache Heron for stream processing and Apache BookKeeper for persistent message storage.

## SWOT Analysis

### STRENGTHS

DataFlow provides users with a data-source-agnostic approach to managing the flow of data through the enterprise, and gives the new Cloudera cross- and upselling opportunities.

### WEAKNESSES

Stream processing has previously largely been addressed by enterprises on a stand-alone basis for specific application requirements, and the broader need for data flow management is not as well recognized as it could be.

### OPPORTUNITIES

We are noting increased interest in performing data processing at the edge, which should drive increased interest in Cloudera Edge Management.

### THREATS

There is no shortage of alternatives, including an increasing number of cloud services based on Apache Kafka, although the combination of functionality in Cloudera DataFlow is a differentiator.