Vendor Profile

Cloudera: IoT Data Management and Analytics with an Open Source Platform

Stacy Crook Stewart Bond Carl W. Olofson

IDC OPINION

Based in Palo Alto, California, Cloudera is a provider of open source-based data management and analytics software. The company was originally formed in 2008 to deliver a commercial distribution of Apache Hadoop that targeted enterprises. Ten years later, Cloudera merged with another major provider of open source-based big data software, Hortonworks. While Cloudera was involved with IoT from an analytics standpoint, the merger with Hortonworks provided Cloudera additional capabilities from an edge and data movement perspective. Further:

- Today, Cloudera is focused on providing what it refers to as an "enterprise data cloud" called Cloudera Data Platform (CDP). Cloudera DataFlow (CDF) – formerly Hortonworks DataFlow (HDF) – is a scalable, real-time streaming analytics platform that sits on CDP. CDF ingests, curates, and analyzes data to help organizations find key insights that can then translate into actions. CDF provides the foundation for Cloudera's involvement in IoT; however, the company's capabilities for data at rest are also important to support the full life cycle of IoT data management.
- IoT projects often begin at the line of business (LOB) to solve a specific problem. As an IT-oriented provider of IoT solutions, Cloudera will benefit by partnering with other technology and services companies that have a close relationship with the line of business. IT stakeholders, however, will appreciate Cloudera's flexible, multicloud approach; the ability to define unified security and governance policies across both IoT and non-IoT data; and the broad set of IoT data management and analytics capabilities the company offers.

IN THIS VENDOR PROFILE

This IDC Vendor Profile provides perspective on Cloudera's strategy for the IoT market.

SITUATION OVERVIEW

Cloudera participates in the enterprise IoT software market. This is a broad space that includes both IoT-specific software products and non-IoT-specific software that can be leveraged for IoT use cases. The companies that participate in this market include a combination of IT vendors, operational technology (OT) vendors, and those that operate in the IoT connectivity space.

The reason there is such a varied group of vendor types that participate in this market is simple. There are many technology components required for an IoT project and many different types of IoT use cases. For instance, a company that wants to develop a smart sprinkler system for the consumer market is a different kind of IoT project than an organization that wants to connect its industrial

August 2020, IDC #US46724420

equipment to the internet. The core requirements of each project might look similar in terms of the need for endpoints, networks, and server-side components to gather, integrate, and analyze data, but the specific endpoints (and their associated protocols), the network topologies, and software architecture could look quite different.

For the past decade or so, organizations have been moving more and more of their workloads into the cloud. IoT has traditionally been viewed as a driver of cloud computing as it offers the compute power and scale needed to process the large volume of data IoT devices emit. More recently, however, IoT use cases have also played a key role in pushing forward another major enterprise computing trend called edge computing. Edge computing allows organizations to store, process, and analyze data close to where it was originally generated. By cutting out the round trip to the cloud for data processing, organizations can save costs and reduce latency for use cases that require real-time analysis. Because of this trend, many technology vendors that originally came out with cloud-centric IoT solutions have now added an edge component to those offerings. Cloudera, for example, gained edge-oriented capabilities with the Hortonworks merger in 2018.

Our belief is that most IoT projects will require a hybrid architecture, where some data is processed and analyzed at the edge in near real time, and other data is analyzed and processed in the cloud (or another type of large centralized datacenter). This requires a real-time data architecture to be in place that can rapidly ingest, transform, and analyze data in stream. However, the algorithms that process the streaming data must be built using historical data. Therefore, organizations often need to have both hot path (real time) and cold path (historical) querying capabilities available for their IoT data. Because of these requirements, vendors looking to participate in the IoT data management realm need to be able to provide both batch and real-time data processing capabilities.

Company Overview

Based in Palo Alto, California, Cloudera is a provider of open source–based data management and analytics software. The company has over 3,000 employees and does business in 85 countries with over 2,000 customers. Cloudera is run by CEO Robert Bearden, former cofounder and CEO of Hortonworks, a company Cloudera merged with in 2018.

Cloudera was formed in 2008 to deliver a commercial distribution of Apache Hadoop that targeted enterprises. In addition to Hadoop, Cloudera engineers contribute to a constellation of other Apache open source projects, including Spark, Hive, Avro, and HBase, as well as others. The distribution was launched in March 2009. Over the years, Cloudera has formed major partnerships with Oracle Corp., Dell, Teradata, and Microsoft. Cloudera and Hortonworks announced their merger on October 3, 2018. Following that merger, the combined company redoubled its commitment to open source and to the concept of a unified data platform. With the acquisition of Hortonworks, IBM joined the list of strategic partners. Although originally focused solely on Hadoop and related open source projects, Cloudera has broadened its scope to address a wide range of data management and analytic needs, all based on open source technology.

Today, Cloudera is focused on providing what it refers to as an "enterprise data cloud" called the Cloudera Data Platform. The Cloudera Data Platform is a multifunctional open source–based platform that can run on the customer's clouds or on-premises datacenter of choice. Various services run on the platform to enable analytics, including a data hub, a data warehouse, an operational database, streaming data capabilities, and machine learning (ML) tools. Each of these services sit upon on a common foundation called Cloudera Shared Data Experience (SDX). Cloudera SDX provides unified data and metadata security and governance policies, no matter where the analytics workloads are running.

Company Strategy

IoT

Product

Cloudera DataFlow is a scalable, real-time streaming analytics platform that sits on CDP. CDF ingests, curates, and analyzes data to help organizations find key insights that can then translate into actions. CDF provides the foundation for Cloudera's involvement in IoT. The CDP data-at-rest capabilities also come into play when IoT data is stored longer term, often so that the data can be used to train ML models.

Three main components to the CDF offering are streams messaging, edge and flow management, and stream processing and analytics. Table 1 provides the purpose and technology underlying each of these components.

TABLE 1

Cloudera DataFlow Description

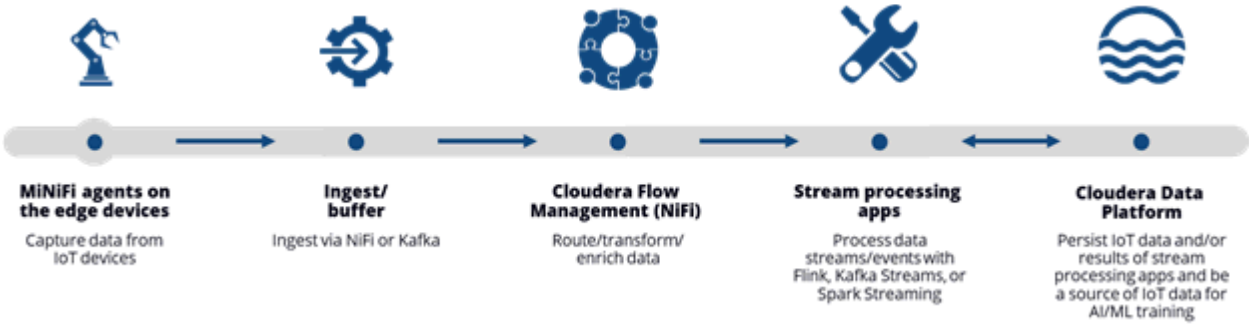
Product Components	Purpose	Open Source and Cloudera Tech
Streams messaging	Messaging, monitoring, and replication	▪ Apache Kafka ▪ Streams Messaging Manager ▪ Streams Replication Manager
Edge and flow management	Data ingestion, intelligence, and monitoring from edge to cloud	▪ Apache NiFi ▪ Apache MiNiFi ▪ Edge Flow Manager
Stream processing and analytics	Event processing for real-time insights	▪ Apache Flink ▪ Spark Streaming ▪ Kafka Streams

Source: IDC, 2020

The way data typically flows through CDF is in the following manner: Apache MiNiFi agents (available in C++ or Java) sit on edge locations and pick up IoT data. MiNiFi can process the data at the edge and route the refined data directly to Kafka, NiFi, or even a cloud data store. At the enterprise or within your cloud, Apache NiFi then picks up the data, routes/transforms/enriches the data as required, and then publishes it to a Kafka data syndication service. The subscribing stream processing apps can then process the real-time data using Apache Flink, Kafka Streams, or Spark Streaming. Figure 1 shows a schematic of typical Cloudera DataFlow.

FIGURE 1

Typical Cloudera DataFlow Schematic



Source: IDC, 2020

The edge management components of the system play a vital role. The Edge Flow Manager allows organizations to define and deploy their flows to the thousands of edge nodes using a no-code user interface and integrate the flows with NiFi registry as well. It allows them to monitor thousands of edge agents at once and update those agents as required. Machine learning model files are also deployed and updated at the edge using this capability. These models can be built with the machine learning capabilities of CDP.

Another important capability of CDF is the default metadata and data lineage capabilities offered by Apache NiFi. A common issue associated with organizations that have implemented their own IoT data capturing solution is a lack of intelligence about the context of IoT data: Where did the data originate, when did it originate, and how did it get to the data storage or processing utility? Apache NiFi tracks data as it moves throughout the system, records and makes events available, handles the merging and splitting of data, and allows users to view attributes and content at specific points in time. This metadata is a part of NiFi by default and populates the SDX component described previously, which sits underneath CDP and CDF to capture intelligence about data.

Go to Market

As an enterprise technology company, Cloudera is mainly focused on business-to-business use cases for IoT. The top industries the company is focused on include public sector, transportation, utilities, healthcare, manufacturing, and retail. The top use cases where Cloudera has seen traction so far are within asset tracking and monitoring, utility monitoring, predictive maintenance, patient monitoring, and usage-based insurance.

IoT projects often start at the line of business within enterprise and public sector organizations. As a provider of horizontal data management technology, Cloudera's traditional buyers tend to sit in the IT and engineering organization. Therefore, while the company has its own vertical specialists that can help drive these conversations with existing customers, one of the best ways for a company like Cloudera to approach the IoT market is through partners that have the relationship with LOB. Partnerships with a company such as Cloudera behoove the LOB specialist as well, as IT is getting brought into IoT projects earlier and earlier in the cycle these days, and often hold much of the project budget.

Cloudera is actively engaging with partners on several fronts. Prior to the merger with Cloudera, Hortonworks had formed an alliance with PTC and HPE that was announced at HPE Discover in Madrid. Since then, Cloudera has aligned at a field team level with these companies to jointly offer IoT solutions to their customers. Cloudera also has a relationship with Red Hat, another key player in open source IoT, specifically around OpenShift, Red Hat's container platform. More recently, Cloudera has become a member of a consortium that will focus on Industry 4.0 adoption within the automotive industry. Other companies engaged in this consortium include Dell and Intel for hardware, PTC and Rockwell Automation for digital solutions, Accenture for consulting and implementation services, and Microsoft Azure for cloud infrastructure.

FUTURE OUTLOOK

At this point in time, it is fair to say that IoT has officially crossed the chasm from the early market to the mainstream market. According to IDC's *Worldwide IoT Decision Maker Survey,* 85% of respondents said that in 2019 they had budget earmarked for IoT projects. However, the main challenge we have witnessed within IoT initiatives over the past several years is not the ability to get a pilot going but in the scaling and growth of that pilot into a mature project.

There are many challenges customers deal with, from both a technology perspective and a business perspective. Some of those issues arise in the early days of the project, such as getting the right devices and infrastructure setup. Other pitfalls emerge later when organizations begin to gather their data and run analytics on that data. From a data and analytics perspective specifically, we see customers struggling in the following areas:

- Cultural/organizational challenges such as a lack of executive support, no data-driven culture, or too many disparate initiatives across enterprise
- Data concerns such as poor data quality, siloed data, too many regulatory concerns, or no inventory of available data
- Metadata/data intelligence concerns such as a lack of metadata about IoT data origin, lineage, context, meaning, timeliness, and relevance
- Challenges related to data refresh cycle times, such as when sensor data is delivered too slowly to take immediate actions on insights
- Process challenges – for instance, analytic and business processes are disconnected, or analytics are not operationalized
- Issues with realizing ROI when organizations have difficulty attributing business outcomes to solution capabilities, often because the measurement of benefits is inconsistent or doesn't happen at all
- Skills-related challenges such as the lack of enough skilled data scientists and analytic talent and/or the lack of developers who are able to incorporate analytics into other enterprise apps

Cloudera's solution looks to address many of the challenges listed previously, demonstrating alignment with market needs. However, it is also important to recognize that some of these challenges are process based and cannot be solved with a product. It underlies the need to work with customers on a more consultative level, so they are set up for IoT success from a people, process, technology, and data perspective. Cloudera can play a role in this, but it also behooves the company to have partnerships with services companies, such as Accenture, that specialize in some of the more strategy-oriented aspects of IoT projects.

One other aspect of Cloudera's solution that we view in a positive light is the open source nature of the software. IDC research has shown that there are several reasons open source software can be an attractive option within the IoT market. These reasons include characteristics such as:

- **Ecosystem innovation and support:** One of the key reasons open source software underpins some of the most important technology trends today, such as analytics, is that in these large communities, innovation can happen much faster than within a single company. Open source projects that deal with real-time data analytics such as Apache Kafka, Apache Spark, and Apache Cassandra are being used by many organizations within their IoT deployments. There is also a strong support ecosystem to leverage within highly active open source projects.
- **Transferable skills:** Transferability of skill sets is another reason an organization may choose to use open source software. Oftentimes, open source projects leverage development languages and technologies well understood by a broad swath of developers, data scientists, and so forth, because if they didn't, it would be difficult for the project to pick up contributors. This is an important factor within IoT, where lack of skills is a commonly cited project inhibitor.
- **Development and deployment flexibility:** Leveraging open source software gives organizations full control of the software development process, the ability to deploy the code where they want, and the ability to make changes to the code as needed. This flexibility is useful when dealing with the distributed architecture that IoT projects often require.
- **Support for diverse environments:** IoT environments are inherently complex and can include many different types of hardware, software, networks, and communication protocols. And these variables can change at any time. Organizations may prefer to use open source tools that give them more flexibility to make changes to their hardware versus a combined hardware/software solution that locks them into a specific vendor's software.

ESSENTIAL GUIDANCE

Advice for Cloudera

- Emphasize the benefits of open source technology in IoT, such as rapid community-based innovation and support, application deployment flexibility, and support for diverse endpoint environments.
- Demonstrate the breadth of capabilities Cloudera brings to bear for the full life cycle of IoT data management, which encompasses functionality for data in motion as well as data at rest.
- Differentiate with edge management and data lineage capabilities, demonstrating the built-in capabilities of NiFi in capturing IoT data intelligence to enable stream processing in context and deliver appropriate outcomes for decisioning.
- As noted previously, Cloudera still has an emphasis on partnerships; however, there were more partnerships relative to the IoT solution space prior to the merger with Hortonworks. Some of the previous partnerships are not as relevant given native capabilities in CDF, but as an IoT data management platform provider, Cloudera will need to continue and build relationships with IoT technology and solution providers.

LEARN MORE

Related Research

- *Worldwide IoT Platform and Analytics Forecast, 2020-2024* (IDC #US45860220, June 2020)
- *IDC's Worldwide IoT Platforms and Analytics Taxonomy, 2020* (IDC #US45860020, June 2020)
- *IDC Market Glance: IoT Platforms and Analytics, 1Q20* (IDC #US45859120, March 2020)
- *Five Key Trends in the IoT Platforms and Analytics Market* (IDC #US45977320, February 2020)
- *The Evolution of the IoT Platform Market* (IDC #US46043519, February 2020)
- *Future of Intelligence: Insights at Scale* (IDC #US45720519, January 2020)

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com