REPORT REPRINT

This Impact Report was published by 451 Research as part of our syndicated Market Insight subscription service and subsequently licensed for commercial use by Cask.

Cask Data completes evolution to data integration platform with CDAP 4

MATT ASLETT

2 FEB 2017

The company recently announced the launch of CDAP 4, marking the completion of its evolution from Hadoop-based application development to also address data preparation, data integration and data management.

451 Research®

Initially launched to provide an application development and deployment platform for Apache Hadoop, Cask Data recently announced the launch of CDAP 4 (Cask Data Application Platform), now positioned as a unified integration platform for big data with application development and deployment functionality. CDAP provides data preparation, data integration and data management, along with the original application development capabilities. With the latest release, the company has added Cask Market, an app store for Hadoop-based applications and reusable templates.

THE 451 TAKE

While we initiated coverage on what is now Cask (and was then Continuuity) from within the data platforms and analytics team thanks to our focus on all things Hadoop-related, the company's initial focus on application development and deployment was not really in our wheelhouse. That explains why the company dropped off our radar for some time, as well as why it now very much has our attention again, thanks to its positioning as a unified data integration platform for big data. The addition of data ingestion, metadata management and data preparation capabilities to CDAP provide access to the data preparation, data integration and data management functionalities that complement the original application development and deployment capabilities, and should be attractive to enterprises looking to develop data-driven applications designed to exploit Hadoop as a data lake.

CONTEXT

Way back in the mists of time (2012), 451 Research initiated coverage on a new startup called Continuuity, formed to build an application development and deployment platform for Apache Hadoop. Much has changed since then. In September 2014, the company changed its name to become Cask Data (although it prefers simply Cask) and set out on a mission to address one of the limitations its early application development and deployment customers had encountered: If you can't get the data to the application, it will be of limited value.

While data integration capabilities were first included in CDAP 3, launched in May 2015, it was with the general availability of CDAP 4 in December 2016 that Cask first positioned CDAP as a 'unified data integration platform for big data,' signifying the evolution from offering an application development and deployment platform with data integration capabilities to a data integration and management platform with application development and deployment capabilities.

There has also been a change in the CEO role at Cask since our first interaction with the company. Founding CEO Todd Papaioannou left the company mid-2013, and is now equity partner at Data Collective. Fellow founder Jonathan Gray took on the role of CEO at that time, with another founder, Nitin Motgi, assuming the CTO role Gray vacated.

Gray was previously part of Facebook's HBase engineering team, while Motgi was part of Yahoo's cloud computing team, which pioneered the development and adoption of Hadoop. Other long-term senior executive include chief architect Andreas Neumann and COO Vikram Bhan, and the company counts more than 50 employees in all.

Details on the number of paying customers are harder to come by, although Cask will say it is in the double digits and that it is expecting to triple the total number this year. Cask adds that it saw initial success with a relatively small number of very large customers – particularly those in the telecommunications and financial services sectors that tend to build rather than buy data and application platforms. The company is now looking to ramp up by targeting larger numbers of smaller customers in industries that tend to buy rather than build, such as retail, healthcare and insurance.

Cask has raised over $37m in funding to date, including a $10m series A round provided in late 2012 by Battery Ventures, Andreessen Horowitz, Ignition Partners, Data Collective and Amplify Partners, as well as a $20m series B round in late 2015 led by Safeguard Scientifics and involving all existing investors. Cloudera also chipped in with an undisclosed investment in early 2015, while Ericsson made a strategic minority investment in mid-2016. AT&T is also a strategic investor.

PRODUCTS

As noted above, CDAP is now positioned as a unified data integration platform for big data, and has evolved from an application development and deployment environment that was initially known as AppFabric and then Reactor. That core application development and deployment functionality remains at the heart of CDAP, which is 100% open source (Apache License 2.0) and is an abstraction layer on Apache Hadoop and Apache Spark that enables the rapid development and management of real-time and batch applications.

CDAP is designed to take advantage of the Apache HBase NoSQL database, Apache Hive for analytics, Apache Tephra for globally consistent transactions on HBase, and the Tigon open source stream-processing framework, which was developed by Cask in conjunction with AT&T Labs.

The addition of data integration functionality into CDAP began with v2 (abstractions and integrations) and accelerated with v3, which added support for data ingestion and data pipelines, as well as workflow, metadata management and data lineage. Specifically, with version 3.2 Cask introduced an interactive application for building, running and managing ETL (extract, transform and load) data pipelines in a data lake environment, while version 3.4 saw the introduction of data discovery, metadata tracking, data lineage and usage analytics functionality.

CDAP 4 sees the introduction of a self-service environment for data exploration, data cleansing and data transformation, as well as visualization, which is integrated with the data pipeline management functionality for operationalization. These data ingestion, management and preparation capabilities are not available as stand-alone products. Indeed, they are best thought of as user interfaces to select functionality in the underlying platform, and are offered as self-service extensions to CDAP. With CDAP 4, Cask also added Cask Market, which is best thought of as an app store for big-data applications and components, including data transformation and processing pipelines, plug-ins, sample data sets, and drivers.

All of the above functionality is open source and freely available by downloading CDAP, and Cask also offers the CDAP Enterprise Subscription, which provides stress, performance and compatibility testing; Web-, phone- and email-based support; and hot fixes (among other things). Cask also offers training and professional services.

CDAP is certified with Hadoop distributions from Cloudera, Hortonworks and MapR, and is also available on AWS, Google Cloud and Microsoft Azure. The ability to run on multiple distributions on-premises and in the cloud, and to provide interoperability between them, is said by Cask to be a major argument in favor of using CDAP rather than attempting to build something similar.

COMPETITION

There are a variety of vendors competing in the data integration and data management market with products and services that overlap with CDAP, although the evolution of CDAP from application development and deployment platform to unified integration platform for big data with application development and deployment functionality means that it is well differentiated. In addition to the application development and deployment capabilities, CDAP's open source licensing and native support for Hadoop are differentiators, along with the fact that it is primarily adopted by developers and engineers rather than the data management professionals that are the traditional adopters of data integration and data management products and services.

As such Cask Data reports that its greatest competition comes from large enterprises trying to build CDAP-like platforms themselves from open source components – and spending valuable time and resources on writing their own custom integration – or even using the open source CDAP without a commercial support subscription. Beyond that, the company does now sees its competition coming from data integration vendors such as Informatica and Talend, rather than application development platform providers, at least to the extent that they are the incumbent data integration and data management providers - which puts them on a shortlist of new data integration and data management projects even if they lack some of the differentiating features mentioned above. We might also expect Cask to be compared with other data integration providers such as IBM, Oracle, SAS Institute, SnapLogic and Hitachi's Pentaho, in the cases where they are the incumbent provider.

The introduction of metadata management and data preparation functionality means that we might also expect comparisons to be made between CDAP and the data lake management/governance and self-service data-preparation vendors, including the likes of Trifacta, Paxata and Datawatch for self-service data preparation; Alation, Waterline Data and Tamr for data management/governance; and Unifi Software, Podium Data and Zaloni, which offer both. Again Cask would argue that these products only offer a portion of CDAP's overall functionality – with Cask having additional operations/management and application development functionality. We would agree that there is an element of false competition here: some of the self-service data-preparation products are comparable to CDAP's data management and data preparation functionality, but this functionality is not available separately, and the combined CDAP can best thought of as a platform for IT that can be used to enable self-service data preparation.

SWOT ANALYSIS

STRENGTHS

Cask now offers a comprehensive platform for the development of Hadoop- and Spark-based applications enabled by data integration and management.

WEAKNESSES

Since CDAP has evolved over time, the company is still not well known as a data integration vendor, and will need to raise its profile given the breadth of competition in this space.

OPPORTUNITIES

As enterprises look to expand on early Hadoop projects with data lake initiatives, we believe the combination of self-service preparation and data management/governance will be key.

THREATS

There are some big vendors in this space with considerably more experience in relation to data integration than Cask.