

## 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 updates data integration platform with increased focus on nontechnical users

**MATT ASLETT**

**09 JUN 2017**

The latest version of the Cask Data Application Platform adds support for Apache Spark 2.x and Apache Hive 2.x, as well as data preparation capabilities designed to appeal to less technical users.

---

THIS REPORT, LICENSED EXCLUSIVELY TO CASK, DEVELOPED AND AS PROVIDED BY 451 RESEARCH, LLC, SHALL BE OWNED IN ITS ENTIRETY BY 451 RESEARCH, LLC. THIS REPORT IS SOLELY INTENDED FOR USE BY THE RECIPIENT AND MAY NOT BE REPRODUCED OR RE-POSTED, IN WHOLE OR IN PART, BY THE RECIPIENT, WITHOUT EXPRESS PERMISSION FROM 451 RESEARCH.



©2017 451 Research, LLC | [WWW.451RESEARCH.COM](http://WWW.451RESEARCH.COM)

Having completed its transition to being a provider of data and application integration with the launch of version 4.0 of the Cask Data Application Platform (CDAP) in late 2016, Cask Data has now launched version 4.2, adding support for the 2.x versions of both Apache Spark and Apache Hive, as well as a specific focus on data preparation functionality for nontechnical users. The company also introduced an improved experience for Windows-based users and a new cloud-based sandbox to ease adoption.

---

## THE 451 TAKE

Cask has evolved in recent years, and now offers a platform that is comparable to a combination of data lake management and data integration and preparation functionality, but remains differentiated thanks to its application development and deployment functionality. That combination of functionality, along with Cask Market, puts it in good position to attract enterprises taking a more strategic view of the data lake and exploring options for products that provide an abstraction layer from the underlying data processing platform.

---

## CONTEXT

We last caught up with Cask Data in February, outlining how the company, founded in late 2011 as Continuity, had completed a transition that began with its change of name in 2014. It evolved from offering an application development and deployment platform for Apache Hadoop with data integration capabilities, to providing a data integration and management platform with application development and deployment capabilities.

The delivery of version 4.0 of CDAP in December 2016, positioned as a unified data integration platform for big data, completed that evolution. The development of CDAP did not stop there, however, and Cask has now delivered version 4.2 with a number of improvements focused on support for emerging Hadoop-related projects, new data connectivity options, improved functionality for nontechnical users, and new deployment options to make it easier for users to get up and running.

CDAP is perhaps best described as a unified data and application management platform for managing Hadoop-based data lake environments and the applications that run on it. CDAP offers functionality for data exploration, data cleansing and data transformation, as well as data ingestion, data pipelines, workflow management, metadata management and data lineage, not forgetting the original application development and deployment environment.

CDAP also features Cask Market, which is essentially an app store for big-data applications and components, including data transformation and processing pipelines, plug-ins, sample data sets and drivers. Cask Market is being used by some customers as an internal marketplace for their own applications. CDAP is Hadoop-distribution-agnostic, and therefore serves as an abstraction layer for enterprises that enables them to develop and manage data lake deployments across multiple Hadoop environments, including Cloudera, Hortonworks, MapR, Amazon Web Services, Microsoft Azure and Google Cloud.

A simplified user experience is one of the primary focuses for version 4.2, with Cask making it easier for users to ingest and start managing existing data sources. There is also a new Cloud Sandbox available on AWS and Azure, lowering the barriers for initial adoption, and an improved Windows installation experience for the 50% or so of CDAP customers running on Microsoft's operating system.

There are improvements to the self-service data-preparation functionality aimed specifically at nontechnical users, including automated integration with existing data sources (such as the Hadoop Distributed File System, relational databases and Apache Kafka), as well as automated operationalization of data pipelines.

Version 4.2 also delivers support for the 2.x versions of both Apache Spark and Apache Hive, and an improved scheduler with support for event-based triggers, as well as new pre-built functionality available via Cask Market, including change data capture (CDC) integration for Microsoft SQL Server, Oracle Database and MySQL. There are also new connectors and adapters for cloud services that include Amazon Redshift, Snowflake and Azure.

Cask is reticent to discuss customer numbers, but told us in February that it's in the double digits, and that the company is expecting to triple the total number this year. Cask did discuss the details of one well-known media-publishing organization (that we are not in a position to name) that has adopted CDAP as part of a major strategic shift to self-service. In a test environment, Cask was able to prove that it could get up and running with a CDAP-managed environment run by two people with no prior Hadoop experience, saving approximately 80% of both time and code compared to a Hadoop environment without CDAP.

Cask has raised more than \$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 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.

## COMPETITION

Cask's combination of application development, deployment and data integration and management functionality is a differentiator, but that doesn't mean the company is without competitors. There are a variety of products and services that compete with CDAP in terms of specific functionality areas. The company notes that its greatest competition comes from large enterprises trying to build CDAP-like platforms themselves from open source components.

In terms of commercial vendors, data management and data integration incumbents such as Informatica and Talend are likely to be among the regular competitors due to their existing customer relationships, along with IBM, Oracle, SAS Institute and Hitachi's Pentaho.

We would also expect CDAP to be compared to a combination of products 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.

The fact that CDAP 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 – is a further differentiator, along with the fact that it offers data preparation/integration and data management functionality as part of a larger strategic development and deployment platform.

## SWOT ANALYSIS

### STRENGTHS

Cask 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, and with CDAP 4.2, Cask has delivered options to ease adoption.

### THREATS

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