REPORT REPRINT

Cask Data boosts self-service data integration capabilities, expands portfolio

MATT ASLETT

05 DEC 2017

Cask Data recently updated its data integration platform for big data and launched two new complementary, but separately licensed, products: a distributed rules engine and a microservices framework.

Cask Data, a specialist in data integration for big data, recently updated its Cask Data Application Platform with the addition of new self-service capabilities and data science improvements. Cask (as the company is more commonly referred to) also launched two new complementary, but separately licensed, products: a distributed rules engine and a microservices framework.

THE 451 TAKE

It is a little difficult to gauge Cask's commercial progress given that it hasn't divulged many details about customer numbers or deployments; however, the company has established an impressive – and differentiated – combination of application development and deployment, and data integration and management functionality, with a growing emphasis on the latter. Cask also continues to expand its portfolio with the addition of a distributed rules engine and microservices framework capabilities, illustrating its potential value as a provider of a big-data platform that abstracts the integration and management of data from the underlying data-processing engines.

CONTEXT

Cask Data is now firmly established as a provider of data integration and management software for big data, having expanded its purview from an early focus on application development and deployment. The company launched the latest update to its software in August. Version 4.3 includes enhancements to the platform's self-service capabilities, as well as a new complementary distributed rules engine and a microservices framework offering.

The company's Cask Data Application Platform (CDAP) already provided a unified data and application management platform for managing Hadoop-based data lake environments and the applications that run on them. 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, and also the original application development and deployment environment. CDAP features Cask Market, which is essentially an app store for big-data applications and components, including data transformation and processing pipelines, plug-ins, sample datasets and drivers.

With version 4.3, Cask introduced a new data-preparation framework, including new user-defined directives to enable users to develop, deploy and use custom data-processing directives. The latest version also has improvements to the product's pipeline studio, making it easier for users to create large numbers of data integration pipelines, as well as support for conditions in pipelines, and automatically triggered pipelines.

Also new in version 4.3 is support for the Apache Ranger data security projects, along with improved support for Apache Spark, including support for Spark Dataframes and PySpark – the Spark Python API – enabling PySpark users to apply their Spark transformation logic into a CDAP data pipeline, run the code and get results without leaving the Spark user interface.

In addition to making enhancements to CDAP, Cask introduced two new complementary products that are available licensed separately from the data integration pipeline itself. The Cask Distributed Rules Engine for CDAP is, as the name suggests, a distributed rules engine that is built on top of CDAP. It acts as an 'if-then-else' statement interpreter to enable business users to specify and manage data transformations and policy enforcements and is designed to run natively on Apache Spark, Apache Hadoop, Amazon EMR, Azure HDInsight and Google Compute Engine. The Cask Microservices Framework for CDAP is similarly self-explanatory and is designed to enable developers to build and manage loosely coupled connected services based on Java APIs and the underlying parallelism and scalability of CDAP to enable real-time applications, including those related to the Internet of Things.

Cask is still reticent to disclose details about the number of paying customers it has, but the company did recently name Thomson Reuters as a reference customer. While we were not in a position to name the company, we previously noted that Thomson Reuters had adopted CDAP as part of a major strategic shift to self-service, and got up and running with a CDAP-managed environment run by two people with no prior Hadoop experience. Adopting CDAP saved Thomson Reuters roughly 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 that involved 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 continues to see its greatest competition come from large enterprises trying to build CDAP-like platforms themselves from open source components. However, the company also faces competition from a variety of vendors targeting the data lake management and self-service data-preparation functionality.

Data management and data integration incumbents such as Informatica (Informatica Big Data Management) and Talend are likely competitors due to their existing customer relationships, along with IBM (Data Connect and InfoSphere Information Governance Catalog), Oracle (Big Data Preparation), SAS Institute and Hitachi Vantara's Pentaho portfolio.

There are a variety of emerging specialist vendors, including Trifacta, Paxata and Datawatch, for self-service data preparation; Alation, Waterline Data and Tamr for data management/governance; as well as Unifi Software, Podium Data, Zaloni and Cambridge Semantics, which offer both. Immuta is a recent market entrant focused on data management, specifically for data science, while Infoworks is focused on automating the data-processing pipeline.

Cask's combination of application development, deployment and data integration and management functionality is a differentiator, as is the fact that CDAP is primarily adopted by developers and engineers, rather than data management professionals. Cask notes that CDAP can be integrated with stand-alone tools – for self-service data preparation, for example – if a customer already has an investment.

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 it will need to raise its profile given the breadth of competition in this space.

OPPORTUNITIES

The Cask Distributed Rules Engine and Cask Microservices Framework for CDAP extend the company's potential value for customers to support a holistic approach to data integration and management-driven application development and deployment.

THREATS

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