Podium adds intelligent data identification and cataloging to Data Marketplace

MATT ASLETT
18 OCTOBER 2017

The company has updated its combination of data lake management and self-service data preparation technology to version 3.2, with new intelligent data identification and data cataloging capabilities.

451 Research®

Having previously positioned its combination of data lake management and self-service data preparation as a Data Marketplace, Podium Data has updated its product to version 3.2 with the addition of new intelligent data identification capabilities, expanded data cataloging and support for multiple deployment options (on-premises, cloud and hybrid).

THE 451 TAKE

451 Research believes that the need for data to be filtered, processed, treated and managed to make it suitable for multiple analytics use cases is critical to delivering value from the data lake. Data governance and self-service data preparation are key elements of functional data lakes (or 'data marketplaces' in Podium's vocabulary), with machine-learning-driven insights and recommendations an increasingly important aspect of accelerating the generation of value from enterprise data. Expanding its footprint to databases, file systems, mainframes and other systems (both on-premises and in the cloud) will also cement the strategic nature of Podium Data's adoption to date.

CONTEXT

The last time we checked in with Podium Data, the company had recently begun positioning its combination of self-service data preparation and data lake management as a Data Marketplace, looking to focus attention on the business benefits of an agile and automated approach to data management that provides a focal point for data producers and data consumers within an organization to privately share and consume data resources.

The company has recently released version 3.2 of the Podium Data Marketplace, including new intelligent data identification functionality, as well as an expansion of its data cataloging footprint beyond Hadoop to relational databases and operational systems, whether they are on-premises or in the cloud.

Podium Data was founded in 2014 and notes that it has had customers in production for almost three years. While the company doesn't disclose the number of paying customers, it does have an impressive roster of reference customers, including Astellas Pharma, TD Bank, Sun Life Financial and Cigna, and notes that it has seen all its existing customers renew with a 50% increase in average annual revenue indicative of its growing footprint in those accounts.

The latest additions to Podium Marketplace 3.2 should help extend that footprint further. The software was initially focused on enabling enterprises to load data into a Hadoop-based data lake from multiple sources, while creating a metadata-based inventory of all data to enable self-service data discovery and preparation.

We previously noted that the company's customers were focused less on the underlying technology – the Hadoop-based data lake – but were increasingly more interested in the business agility advantages of the resulting managed self-service environment, or data marketplace. As such, it makes sense that Podium Data would extend its purview beyond Hadoop, and it has done just that with 3.2, specifically via its Data Conductor, which enables the cataloging of data in Hadoop, as well as databases, file systems, mainframes and other systems. The addition of Data Conductor means that the data cataloged by Podium Data Marketplace does not all have to reside in a centralized data lake, although it offers the functionality to move the data as required.

Additionally, Podium has introduced the concept of managed, registered and addressed data. Managed data is data ingested into the Podium Data Marketplace on the lake. Registered data is data that resides in an external data source, but for which Podium has captured the technical metadata, profiled and validated it, and is available for users with the right permissions to load into Podium as required. Addressed data also resides in an external data source, but for which Podium understands the metadata structures. With 3.2, Podium has also extended its footprint to the cloud and can support any combination of on-premises and cloud data processing and storage, enabling support for elasticity and the management of data in cloud storage as part of a broader tiered storage strategy.

Last, but by no means least, Podium Data Marketplace 3.2 also adds a new capability called Intelligent Data ID, which takes advantage of a customizable rules engine to automatically tag, flag and respond to data as it is ingested and indexed by the data catalog based on pattern recognition and matching.

Intelligent Data ID was originally developed in conjunction with a Podium Data customer (an unnamed North American bank) to automatically detect personally identifiable information (PII), but Podium notes that it can also be used to identify other sensitive data, such as payment card and HIPAA data, as well as personal data covered by the GDPR. Intelligent Data ID can also be used to detect duplicate data and potential data corruption problems. Other relevant use cases include fraud detection and improved data governance.

Podium Data currently has about 40 employees, up from 35 in April and previously told us it was heading for 45-50 by year-end. The company raised a $9.5m series A funding round, announced at the end of September 2016, led by Malibu Ventures. It previously raised a $2m seed funding round in June 2015 (led by CommonAngels Ventures, now Converge Venture Partners).

COMPETITION

There are a variety of vendors competing with Podium Data's combination of self-service data preparation and data management/governance, as well as vendors with niche offerings in each category. Several of them are using different terms to describe what they do: while Podium offers the Data Marketplace, Unifi talks about data as a service, while Zaloni talks about operationalizing the data lake.

All three offer a combination of self-service data integration and data management/governance in a single offering (Zaloni used to offer two separate components but has now combined them in its Data Lake Management Platform). Cask is also a player in this space with its Cask Data Application Platform, as is Cambridge Semantics with its Anzo Smart Data Platform. Meanwhile, Immuta is focused on data management, specifically for data science.

Data integration heavyweight Informatica also offers a combination of self-service data integration and data management/governance via Informatica Big Data Management. It includes Intelligent Data Lake for self-service data preparation and Enterprise Information Catalog, the latter of which was recently infused with artificial intelligence functionality via Informatica's CLAIRE AI engine.

IBM is also a key vendor in this space with its combination of IBM Bluemix Data Connect and InfoSphere Information Governance Catalog, while the data cataloging opportunity is also a target for the likes of Alation, Collibra, and Waterline Data as well as Amazon Web Services and Microsoft via Azure Data Catalog.

There are still further competitors more heavily focused on self-service data preparation, including Datawatch, Oracle, Paxata, Pentaho SAP, SAS Institute, Talend and Trifacta, while there are also a variety of self-service analytics products with data preparation capabilities, including the likes of Alteryx, Birst, ClearStory Data, Datameer, Looker, TIBCO Spotfire and Yellowfin.

SWOT ANALYSIS

STRENGTHS	WEAKNESSES
Podium Data's management team has strong experience in enterprise data management, and the company has landed several large-scale deployments of its product.	The term 'data marketplace' has been used elsewhere in the IT industry – notably to describe public commercial services through which companies can buy and sell data sets, so there is some potential for confusion.

OPPORTUNITIES	THREATS
Expanding its focus beyond Hadoop to the wider data landscape was a natural next step for Podium Data and will cement the strategic nature of its adoption to date.	The incumbent data management and integration vendors are looking to expand their reach into data lake management, and have the advantage of established supplier relationships with enterprise decision-makers.