SAS throws its hat into the self-service data preparation ring with Data Loader for Hadoop

Analyst: Krishna Roy

28 Jan, 2015

SAS Institute is gearing up to make a self-service data preparation play with its new Data Loader for Hadoop offering. Designed for profiling, cleansing, transforming and preparing data to load it into the open source data processing framework for analysis, Data Loader for Hadoop is a lynchpin in SAS's data management strategy for 2015. This strategy centers on three key themes: 'big data' management and governance involving Hadoop, the streamlining of access to information, and the use of its federation and integration offerings to enable the right data to be available, at the right time.

The 451 Take

SAS is best known and most successful as an analytics company. However, it continues to have ambitions as a data management vendor – albeit one that is focused on analytically driven management uses cases. Data Loader for Hadoop exemplifies this focus. Moreover, it should make other products for Hadoop within the company's arsenal more attractive, given that data preparation is 70-80% of the work involved in any analytic project, and Hadoop-based analysis is no different. But while Data Loader for Hadoop is a good start, we think it could benefit from other capabilities that are increasingly becoming table stakes in self-service data preparation, including machine learning for recommendations and collaboration. These functions – and other features – are on its roadmap, but they need to be delivered relatively swiftly for SAS to keep pace with developments in this vibrant and increasingly crowded sector.

Context

SAS will formally unleash its new offering for preparing data for analysis in Hadoop in the first quarter of 2015. Already purchased by a couple of customers, according to management, Data Loader for Hadoop is SAS's self-service data preparation play for providing user-friendly Hadoop access and management capabilities. The offering is a core product in the company's strategy to better support big-data management in 2015. It also speaks to another key theme SAS has set out for the year: making it easier and faster to on-board information and make it ready for consumption.

Like many offerings of this ilk – SAS's product is designed for business analysts without coding skills. That said, it is also being pitched at a more technical user base consisting of SAS coders, ETL developers, and Hadoop programmers who have scripting skills and want to write MapReduce operations, HiveQL and SAS code.

Central to the offering's self-service capabilities for business analysts is an HTML 5 wizard-driven interface that aims to guide nontechnical end users through data management processes they might want to perform. These processes – known as directives – include transforming data from a Hadoop table, copying data to and from a database into Hadoop, and sorting and de-duping a Hadoop table. Directives are also available for profiling and cleansing data, and generating a report from profiled data in the open source data processing framework.

The interface also houses directives for in-database and in-Hadoop data management processing. These capabilities are part of SAS's data management strategy to reduce data movement, and draw on Hadoop's horsepower to cope with processing demands from ever-growing data volumes. Prior to Data Loader for Hadoop, an SAS user would have had to write deliberate code or process flows to push down data management processing to Hadoop.

SAS is also bringing into play several other capabilities to accelerate data management processing and performance within Data Loader for Hadoop.They leverage the SAS Embedded Process – essentially a special SAS kernel that runs inside of Hadoop and processes SAS code in parallel across all of the nodes of a Hadoop cluster – and SAS Data Quality Accelerator and Code Accelerator.

Users can also load specific columns into the company's LASR Analytic Server for in-memory processing. SAS LASR Analytic Server is an in-memory architecture designed specifically with high performance in mind. It is the technical foundation for many of the company's analytics products, including newer offerings for the open source data processing framework such as SAS Visual

Statistics and SAS In-Memory Statistics for Hadoop, which debuted in 2014.

The company's new Data Loader user interface is a VMware vApp that is designed to be loaded onto a user's desktop. The initial release is for single users, but there are plans for a multi-user version. The first cut is also designed for Cloudera and Hortonworks only. However, SAS plans to support multiple Hadoop distributions over the long term, as part of an overall strategy to be agnostic and support as many distros as possible. Pricing is based on the size of the customer's Hadoop cluster.

SAS also has a number of other enhancements lined up for the offering, which is priced according to the size of the customer's Hadoop cluster. They include a recommendation engine, in order to make it more intuitive, and the introduction of deployment models other than an on-premises rollout for additional flexibility.

The company is, for example, mulling a cloud delivery option as part of its cloud deployment strategy for data management. This strategy is currently focused on making its data management offerings available as vApps – as Data Loader for Hadoop exemplifies. Over the long term, SAS intends to serve up on-demand data quality capabilities under a SaaS model, which will involve licensing for multi-tenant use in the cloud.

SAS also plans to draw on other pieces of the Hadoop framework and apply them to the company's data-loading offering for the open source data processing framework, where it makes sense. The initial release is designed to determine the best execution process, which could be a Hive query or faster SAS code by running it within a Hadoop cluster. The ability to leverage Hadoop's Spark in-memory engine for fast processing is one enhancement in the cards.

Competition

SAS is the latest big gun to join the self-service data preparation fray. Informatica began targeting this sector in September 2014 with the announcement of its Project Springbok cloud service. IBM moved into the space in December 2014 with its DataWorks data refinery cloud service. DataWorks is embedded in offerings such as Watson Analytics, and is now positioned as a dedicated offering for end-user data preparation needs. TIBCO Software is also looking to get a piece of the DIY data-wrangling action with its Clarity data quality cloud service.

We also expect SAS to encounter an emerging group of startups that are similarly focused on providing access and data management capabilities for the open source data processing framework.

One such player is Datameer. The startup is the purveyor of a Hadoop-based analysis platform that has burgeoned into data management with the introduction of data profiling capabilities of late. Datameer has also introduced so-called 'smart execution' to automatically select the optimum compute framework for maximum performance, which is another feature it shares with SAS Data Loader for Hadoop.

Paxata peddles a Hadoop-based data preparation offering with broadly similar capabilities to SAS's new offering for the open source data processing framework, although Paxata is delivered as a cloud service, not as an on-premises offering, which is the current deployment for SAS's offering.

SAS may also square up to Tamr and Trifacta, which are other young guns with eyes on self-service data preparation opportunities for Hadoop. We also wonder whether Data Loader for Hadoop will bump heads with Waterline Data Sciences, which is another startup focused on data management for the open source data processing framework. Waterline is all about preventing users from having to dive for data in Hadoop by providing discovery and profiling capabilities, which are also features of SAS's new offering.

Oracle, SAP, Global IDs, Talend, Actian, Magnitude Software and Syncsort are other data management vendors that are potentially competitive, owing to their ability to support Hadoop data access and management. They may also make a broader dedicated self-service data preparation play in 2015, when we also expect the entrance of fresh startups. SourceThought is one startup that will provide competition to SAS, given that it is revving an automated data shaping offering designed natively for Hadoop.

SWOT Analysis

Strengths

SAS has a broad range of data management offerings to underpin its analytical software. Data Loader for Hadoop showcases the company's solid engineering talent and reputation for building high-quality software.

Weaknesses

SAS's Data Loader for Hadoop is a work in progress. It currently lacks many of the features common to self-service data preparation offerings, including recommendations, cloud deployment and collaboration with coworkers who need to be involved in data management projects.

Opportunities

Existing customers struggling with managing a Hadoop environment are the low-hanging fruit. SAS's new offering should also increase the attraction of its existing analytics portfolio for Hadoop, since data preparation is a vital time-consuming step.

Threats

Informatica, IBM and TIBCO have already planted stakes in the DIY data management turf. There are an increasing number of dedicated startups focused on similar use cases that will be snapping at SAS's heels, making for a stiff competitive environment, particularly outside of the company's installed base.