Global IDs gets big into 'big data' management

Analyst: Krishna Roy

29 May, 2013

Global IDs has so far largely focused on automating a range of tasks such as scanning, integrating, profiling, cleansing, mastering and monitoring structured data environments in order to provide large enterprises with a data management stack for the discovery and governance of data assets. However, the recently unveiled version 8 release of the company's Enterprise Information Management (EIM) offering sees Global IDs applying its structured data management capabilities to 'big data' environments, which are semi-structured or unstructured in nature. EIM v8 is also designed to scale up using public and private clouds. In addition, the company is looking to bump up visual analysis and discovery of enterprise data management environments in subsequent v8 cuts by drawing on Hadoop's Mahout data-mining and machine-learning library as well as the Titan open source graph database, and plans to support streaming big data environments.

The 451 Take

Global IDs is a frontrunner in big data management with its latest release, which has wide-ranging support for structured, semi-structured and unstructured information assets. We also welcome the visual analysis and predictive analytic capabilities it is offering since big data has created an increasingly complex data landscape that is often hard to visualize, analyze, understand and therefore manage. However, despite the progress made, there is still work to be done. The extraction of rules for data quality and other data management purposes is still often manual and not automated. Mapping capabilities are also a work in progress – Global IDs can currently map about 80% of a company's data assets. Nonetheless, EIM v8 and subsequent upgrades will likely provide Global IDs with a leg up against its main rivals, IBM and Informatica, as well as give it an arsenal with which to compete with the

increasingly large group of players also pursuing big data management opportunities (see below).

Context

Global IDs is all about offering large enterprises insight into their data environments using an approach that relies on the reverse engineering of data, semantic profiling, and a number of other agents contained in its data management platform, which includes some 160 applications. The company offers its EIM platform via a direct enterprise-oriented sales cycle, which typically takes six to nine months and involves seven-figure deals. Management has recently been broadening its sales avenues by signing up very large system integrators. Global IDs has about 25 customers, and is looking to bolster its direct sales operation with the result that headcount is likely to be roughly 100 by year-end. It currently has about 65 employees.

The company is now looking to address the need to manage the plethora of big data environments increasingly in use. Having introduced the ability to move data in and out of the Hadoop open source processing framework as an early big data player, Global IDs is now looking to better support Hadoop as an ETL environment. The latest version of EIM sports the ability to scan data in Hadoop's HDFS file system. It also draws on the parallelism of Hadoop's MapReduce programming model and utilizes HDFS's distributed storage. Furthermore, EIM v8 will tap into several other components within the open source data-processing framework in subsequent updates (see below). The latest release is also designed to support database scanning of other big data environments such as SAP's HANA in-memory analysis engine, the open source Apache Cassandra NoSQL database, and the Blueprints graph database.

In addition, EIM v8 provides scanning capabilities for a number of other semi-structured and unstructured data environments, including PDF documents, XML data, files, Microsoft SharePoint and social media, essentially scanning every data environment in order to create a comprehensive repository of an organization's data assets on which to perform detailed semantic analysis. The latest version also introduces an enterprise search feature to enable users to search for any data asset, as well as a business ontology to create maps to show the distribution of master data. Like its predecessors, EIM v8 enables the creation of data-quality controls for validation and cleansing purposes. However, it also allows them to be displayed as data-quality metrics in a dashboard for better data stewardship. The latest release can also run on a VMware virtualized environment and OpenStack's cloud and provides MDM capabilities for big data, as we discussed in an earlier report.

Data discovery and understanding a company's data environment and how a customer ID relates to a product ID, for example, is a longtime core capability. Global IDs is continuing to expand the number of different types of IDs it can support, and is also bolstering its data discovery and analysis capabilities. EIM v8 provides integration with the R open source data-mining language as an initial foray into embedded analytics, which it will build on subsequently via support for Mahout. The endgame for the company's embedded analytics strategy is to provide an environment for learning, predicting, linking and analyzing a variety of structured, semi-structured and unstructured data environments to support the data mining of master data and predictive modeling.

For example, linking master data relationships via the use of the Titan open source graph database for a graph view of relationships between customer and product IDs is coming in an upcoming EIM v8 release. Global IDs also has other Hadoop-related enhancements in the works for later releases. It plans to support Hadoop's Hive data warehouse and Sqoop tool, which is used for the bulk transfer of data between Hadoop and structured repositories such as relational databases. The company also wants to embrace data movement in graph databases as a target environment, as well as HDFS. Support for streaming data sources such as complex-event-processing engines as part of a real-time big data play is also in the cards.

Competition

Global IDs' focus on serving a broad variety of data management needs for large enterprises means that it tends to bump heads with IBM and Informatica, which target similar opportunities. When compared with these rivals, Global IDs is a lesser-known, smaller vendor – albeit one that seems poised for further growth.

Both IBM and Informatica have made big data plays for data management, which include embracing Hadoop. Big Blue illuminated its strategy for big data integration and governance in early 2013, drawing on a portfolio of assets such as its InfoSphere Information Server for integration and cleansing as well as its Data Explorer for big data content navigation. InfoSphere BigInsights and InfoSphere Streams also play a key role in addressing Hadoop-based analysis and the streaming of events for real-time customer analysis. Meanwhile, notable offerings inside Informatica's big data portfolio include HParser, the company's data-transformation environment optimized for Hadoop, and PowerCenter Big Data Edition for semi-structured, unstructured and structured data-integration needs.

Other data management players looking to meet big data needs include Syncsort, Talend, Pentaho, Cirro and Actian's Pervasive Software data-integration business. Syncsort is pursuing a similar

market to Global IDs in the enterprise sector and provides competition when it comes to Hadoop-related ETL and sort scenarios. Talend is targeting data profiling, data quality and integration for Hadoop in its Platform for Big Data, which is also designed to support HBase, HDFS, Hive, Cassandra, Couchbase, CouchDB and MongoDB databases. Pentaho provides its Instaview module for discovering, visualizing, exploring and analyzing big data environments such as Hadoop, Cassandra and MongoDB databases in tandem with structured data. Cirro offers an enterprise data hub for the integration and analysis of relational and non-relational database environments, and is focused on enabling access to data in Hadoop and other environments and repositories via tools that users already have on their desktops, such as Excel and Tableau. Actian's Pervasive provides big data predictive analytics for Hadoop, as well as data integration and profiling for the open source data-processing framework. Javlin's CloverETL tool and SnapLogic also support Hadoop integration.

SWOT Analysis

Strengths

Global IDs' agent-based approach to a range of data management tasks, which can be linked together via a workflow, makes for a broad-based yet integrated approach that covers a wide variety of data sources, both emerging and traditional.

Weaknesses

Some of the company's big data capabilities are a work in progress, and yet to be delivered. It doesn't really provide an offering for organizations other than large enterprises, which precludes deals with SMBs.

Opportunities

EIM v8 should find favor with companies that have Hadoop and other big data environments in production and therefore require management capabilities. VMware and OpenStack support increased scalability and therefore build on EIM's big data appeal.

Threats

IBM and Informatica have a higher market profile and global reach, and are already serving some of the big data scenarios that Global IDs is now targeting. The big data management space is also being pursued by a variety of other players, including Syncsort, Talend, Pentaho, Cirro and Actian's Pervasive, which all provide competition on one or more fronts.