Tamr lands $16m series A round for data preparation service with mapping, matching slant

Analyst: Krishna Roy

27 Jun, 2014

Tamr (formerly Data-Tamer) has emerged out of stealth mode with series A funding and an offering in the marketplace. The startup – the brainchild of database industry pioneers and entrepreneurs Mike Stonebraker and Andy Palmer – is the latest to enter the data preparation sector with a machine learning-based approach to the issue of ensuring that data, in a variety of formats, is ready for analysis. More specifically, Tamr is looking to take the heavy lifting out of the attribute mapping and recording matching process for semi-structured and structured data, using an enterprise-oriented multi-tenant cloud service designed for data scientists.

The 451 Take

As we have said before in our Total Data Integration report, semi-structured information needs to be united with structured data in order to provide insight, which wasn't possible or present when only focusing on the integration, mapping and matching of data of the structured variety. With its sterling management pedigree, cloud service designed to serve a genuine market need, and financing to further product development, Tamr is off to a strong start. However, it is entering an emerging field courted by other new breeds of data management players, which will not offer quite the same thing and are similarly focused on serving analysts' data preparation requirements at scale and on diverse data types.

Copyright 2014 - The 451 Group

Context

Tamr has launched into the data management sector, having generated a great deal of buzz over the past year, in part due to the industry credentials of its founders. The startup took the wraps off the first release of its multi-tenant mapping and matching service in May, when it also announced series A funding. The culmination of four years' development work – two years within Tamr and two inside the Massachusetts Institute of Technology – the startup's offering was conceived by its co-founders Stonebraker and Palmer.

For those outside the IT industry or new to it, Stonebraker is a database research pioneer, one of the original developers of Ingres and PostgreSQL databases, and a professor at MIT. He has also founded a number of other database companies, including Vertica Systems, which was acquired by HP in 2011, and VoltDB, which we examined more closely here. Fellow serial entrepreneur Andy Palmer also co-founded Vertica and VoltDB, and has a history of being a founding investor, advisor or board member of a number of startups. He also founded Koa Labs, a shared workspace for early-stage startups in Harvard Square, Cambridge, in 2012.

Management says that the idea behind Tamr – which is headquartered very close to Harvard Square and has also just opened a sales, business development and engineering operation in San Francisco – dates back to Vertica days. At Vertica, Palmer et al. observed an increasingly common problem. Data scientists and analysts had forgotten attributes and/or wanted to add new data sources to a warehouse, after already implementing an ETL process, which management noted required them to go back to square one. Tamr was conceived to address this issue by providing a bottom-up probabilistic approach to integrating data sources, rather than the top-down deterministic modus operandi that is common in ETL.

Funding

The startup, which was founded in January 2013, has roughly 25 employees and is funded to the tune of $16.5m. It scored $16m series A funding in May. The latest financing came from Google Ventures and New Enterprise Associates (NEA) and is being used for engineering and product development.

Strategy/customers

Tamr is focused on serving up hard core plumbing to data scientists, analysts and other technical folk, who need to meld together data in a variety of formats and want an environment that essentially presents them with integrated data and metadata, which is already related.

It is currently using a direct sales model, although the long-term game plan is to focus on indirect sales opportunities with OEM customers. For now, the startup has a small business team peddling the Tamr mapping and matching cloud service, which is billed as a data connection and enrichment offering, and designed to be complementary to master data management (MDM), and an enabling technology behind analytics.

Tamr is using an annual subscription-based pricing model. The aim behind the pricing model is to get enterprises started for a relatively small investment – in the hundreds of thousands of dollars per annum – with a view to expanding the deployment over time. The endgame is to enable every data scientist to use it to get started on some type of data curation process, while simultaneously promoting and fostering collaboration between these individuals and those with knowledge of the particular data sets in question.

This startup's offering is available on-premises and in the cloud on Amazon EC2 and Google's GCE cloud platform. Management notes that although Tamr was designed to be cloud native, many current customers deploy it on-premises.

The company chose to get customers on board before it had a formal coming-out party for the startup. As a result, it claims to have 5-10 paying customers using the offering. Although many use it as a multi-tenant cloud service, management notes that it does have customers that have installed it on-premises. Furthermore, it will provide customers with a single-tenant environment, if requested.

A key early use case, according to management, is using Tamr to get a unified view of a customer. Thomson Reuters is one of its first marquee accounts. The media and information conglomerate is using Tamr to address a record matching problem to connect previously 'siloed' data as part of an internal data integration process, which requires a high level of automation and precise match rates. Thomson Reuters is also using Tamr for attribute matching and record de-duplication.

Product/technology

So what's actually under the hood? Machine learning algorithms are a central lynchpin, and fundamental to Tamr's ability to ease and automate attributing matching and record matching. However, Tamr doesn't rely exclusively on machine learning. Instead, it also requires human intervention to provide the domain-specific expertise required when mapping and matching at scale, for precision purposes.

Tamr, in essence, generates questions for the data expert, aggregates responses, and then feedsthem back into the system. There's also a data inventory, where the user can explore all of the connected data sources and pick out an expert to review the data in the directory of those available. The directory is also designed to provide the user with the data expert's area/areas of expertise and levels of expertise, which it tracks automatically over time.

That said, Tamr is also designed to learn and therefore the algorithms get better over time, and start predicting when a new data source is added, in a bid to cut down on the manual review process, and improve the confidence level of the mapping. Tamr essentially uses unsupervised and active learning techniques in order to generate a high classification model, without requiring a technician to oversee the training process.

The Tamr cloud service is also underpinned by a homegrown triplestore. A triplestore is a database, which is purpose-built for the storage and retrieval of triples; i.e., data entities composed of a subject, a predicate and an object, such as 'Jill is 28.' The aim behind the data store is to enable it to accommodate a wide variety of independently constructed data, as well as schema-less data, dirty or missing data, and information that is poorly formatted.

Tamr also implements an automated feature extraction system in a bid to enhance both the precision and recall of its machine learning algorithms. The cloud service essentially builds a two-stage classifier using a single set of training data in order to generate candidates for connections and achieve the desired precision and recall. It also uses cluster analysis to resolve many types of conflict and uncertainty. The final set of integrated metadata and data, which has been produced by the mapping and matching service, is consumed via a RESTful API, which is designed to be hooked into an existing data management infrastructure in the enterprise.

Last, it is worth noting that the Tamr service isn't for unstructured data. It relies on partners to tag the data, so that it essentially becomes semi-structured and can be handled by its cloud service. Current data sources supported include Oracle and SQL Server, Hadoop's File System (HDFS), MongoDB, CSV files, XML files and IBM's Cloudant database as a service. The game plan is to continue to build out data source connectivity in a bid to be as open and connected as possible.

Competition

Tamr is likely to elicit comparisons with Trifacta and Paxata. Why? All three startups share a similar focus on the data preparation requirement for traditional structured data, as well as emerging sources of information in Hadoop and other 'big-data' environments, and use machine learning to boot.

Management notes that Tamr and Trifacta are no strangers to each other because the latter's cofounder and CEO, Joe Hellerstein, was one of Mike Stonebraker's PhD students. Mike and Joe also discussed founding a company together before Hellerstein went off to establish Trifacta. While we certainly think Trifacta is more focused on the front-end user interface challenges of transforming semi-structured and structured information, while Tamr is concentrating on back-end data quality-oriented plumbing, the market might not see the distinctions so clearly. First, both players frame themselves as big-data startups. Second, they also concentrate on significantly improving the productivity of data scientists and analysts.

Moreover, it is also worth pointing out that Paxata is also positioning around similar messaging, and furthermore also has a multi-tenant cloud service for the process of joining different data types and cleansing them, as its primary play. Paxata also has a strong collaborative angle, and while technically different to Tamr, could be used as an alternative in certain situations.

When it comes to the need to integrate semi-structured information with structured data, we also see competition coming from Cirro, which is another startup whose wares could be used to gain a single customer view, for example. However, technically Cirro is using data federation and an optimized query engine, which is a different technical base from Tamr's, and complementary in certain ways.

We also wonder whether the startup will bump heads with Global IDs, which is focused on a number of issues associated with the management of semi-structured and structured data, including data mapping.

We also see both a complementary and competitive role for Tamr against ETL offerings, which are traditionally used to map and match data because profiling and data cleansing have become table stakes in most offerings of this ilk, including those from IBM, Informatica, SAP and Talend. To add another wrinkle, ETL tools are increasingly embracing semi-structured data sources, starting in the main with Hadoop, as exemplified by Pentaho Data Integration (PDI), which further blurs the lines with Tamr.

Copyright 2014 - The 451 Group

SWOT Analysis

Strengths

Tamr has an elite management pedigree and track record. The startup also has customers in production and using its wares prior to launch, which is refreshing and unusual in these highly marketing-tastic days.

Weaknesses

We're not convinced all data scientists will embrace a multi-tenant cloud service. Data source connectivity currently covers only the basics.

Opportunities

Hadoop distributors seem like a good first start for OEM accounts. In the short term, the startup is wise to build up a bank of direct customers to validate its technical smarts and differentiation, which may take some evangelizing.

Threats

Tamr is not alone is looking to make data scientists more productive by lessening the data preparation burden. It is also likely to be joined by other fellow young guns – and more established players including Informatica with Project Springbok – making the competitive environment increasingly tough.

Reproduced by permission of The 451 Group; © 2014. This report was originally published within 451 Research's Market Insight Service. For additional information on 451 Research or to apply for trial access, go to: www.451research.com