MIT CSAIL ALLIANCES

Tamr

Member Success Case Study

Tamr

intelligence, and machine learning help businesses stay competitive. The more data you have available from different sources, though, the more you have to think about organizing it and analyzing it so that users can have a unified view. How do you "tame" unruly, large-scale data so that it is manageable for analytics?

One data integration tool to come out of the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Data Tamer, has been so successful in solving this problem that 2014 Turing Award-winner and CSAIL Professor Michael Stonebraker built a company around it with co-founder Andy Palmer.

About the Company

Founded in 2013, Tamr, Inc. addresses the challenge of unifying and mastering large volumes of highly variable data within an organization. Prof. Stonebraker observed that traditional solutions focused on top-down, rules-based methods for managing data. Moreover, these solutions work well at small scale, but fail miserably in the "big data" space. Specifically, rule systems do not scale, because humans cannot "grok" more than a few hundred rules. Big data problems cannot usually be solved by small rule sets, and a different approach is needed. Stonebraker, in collaboration with researchers at the Qatar Computing Research Institute (QCRI), began to investigate applying machine learning (ML) to unifying large amounts of data.

This research project at MIT CSAIL resulted in a 2013 paper, "Data Curation at Scale: The Data Tamer System." Data Tamer is an end-to-end data curation system that combines machine learning to scale and human expert guidance to solve unification problems in big data.

Based on successful prototype applications of Data Tamer at multiple companies, Stonebraker recruited co-founder Andy Palmer to form the commercial entity Tamr to extend, robustize, and commercialize the Data Tamer system. Tamr is now seven years old with headquarters in Harvard Square and is going strong, even with the headwinds resulting from the presence of COVID-19.

The Challenge

"Data integration is one of the big challenges in this big data space," said Stonebraker. He breaks down big data into the "three V's": Volume, Velocity, and Variety. According to Stonebraker, too much volume or velocity can be solved by "big money," whereas variety is "just plain hard."

He explained: "If you've got too much variety, you've got a horrible data integration problem, and that's what's killing a lot of enterprises. They organize their businesses into independent business units (IBUs) so they can achieve business agility. Otherwise, every decision has to go to the top of the management chain, and agility goes out the window. IBUs tend to set up their own IT systems, and as a result create 'data silos'. After the fact, there is tremendous business value in integrating the silos." *(continued)*

The Challenge (continued)
For example, he continued, "One IBU might sell refrigerators in the USA; another sells air conditione in Asia. Each of them would like to know the customers of the other so that cross selling is possible However, that requires independently constructed customer databases to be integrated. These customer databases invariably use different names, different terms, different units and store different data; hence data integration is a very hard problem."

The Solution
Tamr unifies data sources across an organization with great speed, scalability, and accuracy. The platform is capable of "connecting" data sources to first align relevant datasets — "cleaning" these datasets by getting rid of duplicates, mastering entities within the unified datasets, and "classifying" records within the datasets according to any taxonomy. Users across the organization can then use the resulting datasets to fit their needs, whether for analytic, operational, or raw consumption.

For example, suppose that a Fortune 500 diversified manufacturing company has 75 IBUs, each with their own procurement system for buying parts. Ideally, a company would have only one procurement system, but M&A activities and IBU agility might leave this company with many more. The CFO of the company realizes that he can save $1 million a year with the company's paperclip supplier by first finding out what terms and conditions the 75 procurement officers managed to negotiate, and then having each demand "most favored nation status." Over all suppliers, the savings add up. However, obtaining these savings requires the company to integrate 75 independently written supplier databases. Tamr helps to integrate all this data, so the company can enjoy significant savings.

The Engagement
In addition to the Data Tamer curation system, Prof. Stonebraker has developed many database prototypes during his years at MIT CSAIL, including the Aurora/Borealis stream processing engine, the C-Store column-oriented DBMS, the H-Store transaction processing engine, and the SciDB array DBMS. "I'm a big fan of doing exploratory research at CSAIL," he said. "Some of it works out, some of it doesn't. And the stuff that has commercial value — throwing it over the wall to a startup to get commercialized is fabulous."

After frequently being asked how to start a company, the Tamr co-founders worked with CSAIL to teach an edX course on the topic. "It's a collection of steps to go through, in order to get a commercial company off the ground — including how to negotiate with venture capitalists," they said. "It's sort of a recipe for getting a company off the ground." These steps are practical and come from experience (Stonebraker has started nine companies and Palmer has also started several companies), and it shows the stages from building a prototype to getting a product backed by VCs.

Tamr's academic, research-oriented perspective is a clear advantage in the business of big data. After growing from research at CSAIL, the startup continues to apply tools and develop methods to "tame" data so that we can actually interpret and use big data practically as the volume, velocity, and variety of data continue to expand.