tamr

Data curation at scale

Information Services
Case Study

Information Services Leader Uses Tamr to Automate Data Curation, Slashes Manual Reviews by 90% and Cuts Process Time by Months

A multinational media and information company faced challenges maintaining critical, accurate data. It had outgrown its manual curation processes and looked to Tamr to provide a better solution. Using Tamr, one project, estimated to take six months, was completed in only two weeks, requiring just forty man hours of manual review time—a 12x improvement over the manual process. The number of records requiring manual review shrunk from 30% to 5%, and the number of identified matches across data sources increased by 80%—all while clearing the company's 95% precision benchmark. The disambiguation rate—the rate of resolving conflicts—rose from 70% to 95%. Furthermore, the knowledge Tamr gleaned from its machine learning activities means that future data integration will take even less time per source.

"Tamr proved that fast and accurate data integration results in tremendous benefits. By combining the system's machine learning with the knowledge of our data experts, we can dramatically improve the quality of our services."

— VP of Data Services

	Existing process (estimates)	Tamr Results (actual)	Benefits of using Tamr	
Project Time	6 Months	2 Weeks	92%	Reduction in time required for the project
Automated Matches	210,000	375,000	79%	Increase in matches
Disambiguation Rate	70%	95%	35%	Increase in rate of matches or uniqueness
Records Requiring Review	30%	5%	96%	Reduction in time required for the project
Manual Review Time	>50 Days	5 Days	>90%	>Reduction in time to review

Note: Based upon 5.4 million records analyzed

A Need to Replace Manual Curation

An information services company's reputation, brand and market position depend on the breadth and quality of its data. With a huge customer base—media organizations around the world, 80% of Fortune 500 companies and over 400,000 end users—the company's data challenges limited the efficiency of its business operations and restricted its ability to capture additional market share.

As the number of available data sources grew from tens to hundreds to thousands, several issues became increasingly clear:

+ **Manual process limited scalability.**
  Integrating just a few data sources required extensive manual efforts and often relied on people conducting manual curation—essentially relying on spreadsheets to try and solve the problem. These slow, manual processes also left many crucial relationships between datum unmapped.
+ **Curation did not include all data experts and owners.**
  The people most qualified to understand and make authoritative decisions about data were spread across the company, and the curation processes did not have a way to consistently involve them and leverage their knowledge of the data.
+ **Low-quality and the slow pace of curation impacted customers.**
  Using its existing curation approaches, the amount and the speed of incoming data made it hard for the company to meet the contractually-obligated service level agreements of its customers. Manual curation had become a governor on growth.

Massive Improvements with Tamr

Company executives agreed that their business goals were not achievable without a new approach to integrating and curating data sources. Understanding that the ability to quickly integrate new sources at incremental or sub-linear cost could create tremendous and sustainable competitive advantage, they looked to Tamr to help them with two short-term goals: improve the quality of their data curation and achieve radically better scalability.

The company wanted to focus on integrating three of its core data sources—factual data on millions of organizations with more than 5.4 million records. Previous in-house curation efforts, relying on a handful of data analysts, found that 30%-60% of entities required manual review. The company estimated that if the project was completed in the existing, manual manner, it would require two months of man-hours to fully ameliorate the sources. Additionally, it was thought that the old process would identify 95% of duplicate matches (precision) and 95% of suggested matches that were, in fact, different (recall). Overall, the best guess for completing the activity with the manual process was six months.

Tamr kicked off the project by converting the company's XML files to CSVs. Next, Tamr ingested the three sources to de-duplicate the records and find suggested matches, with a goal of achieving high accuracy rates while reducing the number of records requiring review. In order to scale the effort and improve accuracy, Tamr applied machine learning algorithms to a small training set of data.

Completed in two weeks, the project resulted in better matching results and dramatically less human intervention. The company, impressed with the significant increase in results and substantial decrease in required man-hours, is expanding its use of Tamr, integrating even more data sources. As the new, Tamr-driven curation processes expands in the organization, the benefits increase as the system continually learns and improves.

Tamr gets Smarter with Each Engagement

Tamr has developed an intuitive user interface that allows data analysts to easily send curation tasks to users who know and people who produce the data—and to build that step into a company's workflow. As machine learning algorithms understand more about users and data environments, organizations spend less time preparing data—even as they scale up the number of sources to hundreds or thousands—and more time competing on analytics.

Tame Your Curation Challenge Today

To learn how Tamr can dramatically lower the cost, improve the quality and boost the speed of your data integration activities, call **617-413-6551.**