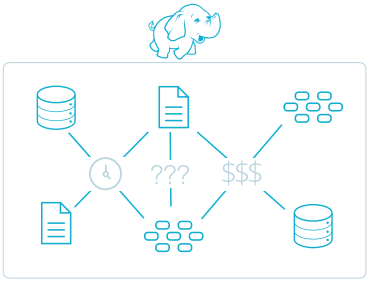
Tamr Unifies Datasets In Hadoop To Unlock Hidden Insights

Companies Struggle With Integrating Data In Hadoop

Hadoop has helped organizations significantly reduce the cost of data processing by spreading work over clusters built on commodity hardware as well as giving companies the ability to host massive amounts of heterogeneous and diverse data sets. With the growing popularity of Hadoop, a significant amount of organizations have been creating Data Lakes, where they store data derived from structured and unstructured data sources in its raw format. However, these companies struggle with connecting and transforming the data into a unified dataset for business analysis without significant investment in time and money. This is largely because schema proflieration is rampant and very rarely are any two datasets structured exactly alike.

Tamr's Matching Engine Unifies Data Within Hadoop

Tamr solves the biggest challenge in unifying datasets in Hadoop, namely connecting and cleaning the data so that it's ready for analytics. Tamr is a data unification platform that leverages machine learning and customer expertise to create integrated, clean datasets with unrivaled speed and scalability. In particular, Tamr focuses on profiling datasets, creating ideal target schemas, and deduplicating records in order to prepare datasets for analysis.



Companies struggle with connecting and transforming the data into a unified dataset for business analysis without significant investment in time and money.

Tamr's core offering for Hadoop consists of two components:

+ A module for training, administration, and expert sourcing that runs on top of a relational database on an edge node of the customer's Hadoop cluster.
+ A matching engine that runs distributed on the Hadoop cluster where pertinent data is stored

Because of the scale of the data, it is very expensive to move Hadoop-based data outside of the Data Lake. Therefore, Tamr avoids this by doing all data-scale processing within the Data Lake, thus eliminating the need to replicate the entire data set.

Tamr's Data Lake capabilities can be best explained by example. Let's take a (simplified) example where customer data is the focus of what's stored in a company's Data Lake and the com is trying to generate a 360 view of their customers. The particular data sets stored are:

+ CRM data with information about each customer the organization does business with, often with several duplicates
+ Clickstream data from the company's website
+ Transactional data such as prior purchases by each customer

Let's assume the following data structures:

CRM Data

Customer ID	Name	Email	Address	Phone Number	Total Spend	Subscription Preferences	...
345	John Doe	john@gmail.com	...	...	...	...	...
80920	Johnny Doe	john@gmail.com	...	...	...	...	...

Clickstream Data

ID	Email	Phone	Page Visited	Time Visited	Time On Page	...	...
Sam Roberts	...	...	...	...	...	...	...
Garry Johnson	...	...	...	...	...	...	...

Transactional Data

Name	Email	Product ID	Purchase Data	Amount Spent	...	...
John Doe	john@gmail.com	...	...	...	...	...
Jonny Doe	John@gmail.com	...	...	...	...	...

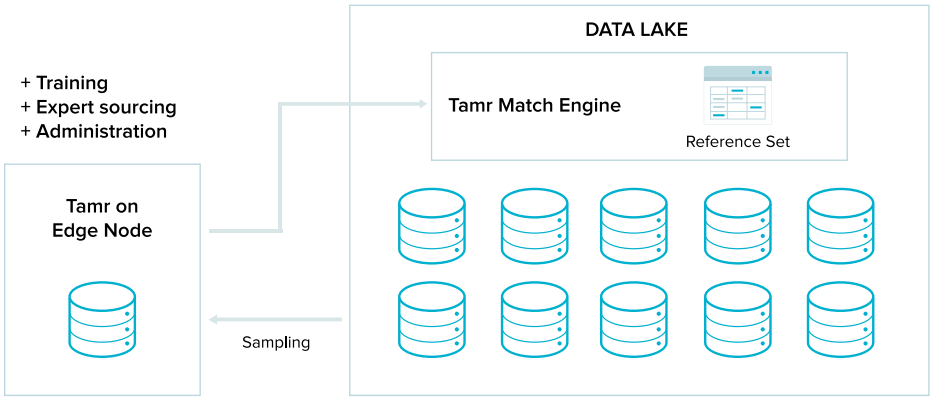
The 'Clickstream Data' and 'Transactional Data' are transactional in nature and, hence, will likely grow to be very large in size. As an example, if a company has 1 million customers, the clickstream data might contain billions of rows. Moving it out of the Data Lake is not an option due to cost and technical challenges.

Tamr Deployment

Below is the process for how Tamr will be deployed in the environment described:

Registration

To begin, a user would deploy a Tamr instance on the edge node of the company's Data Lake (Hadoop). The user then registers data sources within Hadoop that are relevant to the particular analysis being conducted. For example, in this case the three nodes with the data sources mentioned above would be registered.



Schema Mapping

Once registration of sources is complete, Tamr will read all relevant data sources in Hadoop and pull in samples of the data in order to conduct schema mapping. This data set will only have a small sample of rows to help the user conduct a schema map. Through a unique combination of machine learning and expert sourcing, Tamr will work with the end user and business experts to create a schema for the unified dataset, coined the 'Unified Schema'.

Let's assume that the schema mapping is identified as follows. Note that most attributes from transactional data sources are not mapped, as they do not help to identify a customer. The table generated through this schema could later be joined with a transactional database for analytics.

Entity Matching

Unified Attribute	CRM Data	Clickstream Data	Transaction Data
Name	Name	Name	Name
Email Address	Email Address	Email	Email
Phone Number	Phone Number		Phone Number
Address	Address		Address

Once critical attributes are identified, Tamr will bring in records from each dataset for entity matching. Because only the unified attributes are relevant, Tamr can bring in records that only differ in those aspects through its intelligent sampling capabilities. As an example, the billions of rows in the clickstream data will be reduced to millions of unique records.

Tamr will then use machine learning to automate most of the entity matching, while ensuring the highest levels of accuracy through Tamr's native capabilities around expert sourcing. Tamr's machine learning-based approach has the added benefit of learning as experts feed insight into the product (i.e. verification of matches) such that future matching processes require continuously less human intervention.

Matching Engine Deployment On Hadoop To Create Clean, Integrated Datasets

Tamr produces 'reference maps' of keys as a result of this sampling exercise, which contain IDs related to all unique entities Tamr has identified as well as IDs related to the unified attributes identified. These reference maps are then pushed to Tamr's matching engine, which is natively deployed on Hadoop.

Tamr's matching engine will leverage the reference maps to:

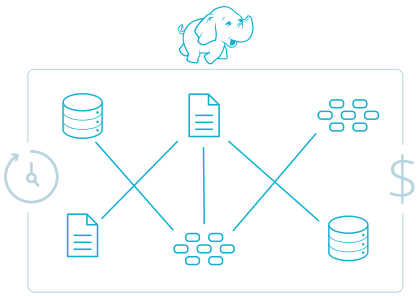
+ Align all relevant source attributes from across datasets registered in Hadoop to unified schema attributes, regardless of the multitude of naming conventions associated with the source attributes. In the Customer 360 example, Tamr could identify all source attributes that relate to a unified attribute, for example 'Customer Name', even if the source attributes are called 'Cust_Name', 'Full Name', or 'Consumer_Name'.

+ Identify and cluster records referencing the same entity across multiple datasets in Hadoop. In the Customer 360 example, Tamr would identify a cluster of names that are all related to 'Robert Smith' even if the source records listed the customer as 'R. Smith', 'Bob Smith', or ' Rob S'.

The application of Tamr's matching engine on the Hadoop nodes will enable a customer to produce data tables of interest. In the Customer 360 example, Tamr would help produce a customer table which can be used in the construction of a data mart or in downstream analytics tools. This table represents the clean, integrated dataset missing from many Hadoop implementations today.

Iterative, Ongoing Analytics

As additional data gets added to registered data sets, Tamr continues to match it to the reference data set initially created. In the Customer 360 example, as new customers get added, there might not be a match to the existing reference data set. In this case, Tamr will automatically add the record to the reference data set or send it to experts for review. This is enabled through Tamr's ability to create master lists of entities (in this case, customers) being modeled and match new data sets to it at scale. The capability allows the organization's downstream analytics to remain updated as they continue to use Hadoop as a source for critical enterprise data.



Tamr integrates related datasets across Hadoop nodes very efficiently and enables customers to finally realize a return on their Big Data investments.

Unleash The Power Of Your Hadoop Implementation

Investment in Hadoop is expected to continue for some time as companies look to capture and analyze data that was not previously accessible to them. One of the major factors that turn Data Lakes into data swamps is the failure to integrate related datasets across Hadoop nodes. Moreover, there is a need to conduct this integration without having to spend the time and money extracting the data from Hadoop. Tamr solves both of these problems very efficiently and enables customers to finally realize a return on their Big Data investments.

About Tamr

Tamr, Inc., provides a data unification platform that dramatically reduces the time and effort of connecting and enriching multiple data sources to achieve a unified view of siloed enterprise data. Using Tamr, organizations are able to complete data unification projects in days or weeks versus months or quarters. For your own personalized Tamr demo, visit www.tamr.com.