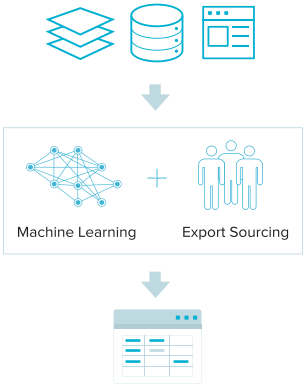
Tamr Technical Whitepaper

1. Executive Summary

Tamr was founded to tackle large-scale data management challenges in organizations where extreme data volume and variety require an approach different from legacy technologies. Whereas most traditional solutions focus on top-down, rules-based methods for managing data, Tamr focuses on a bottom-up, machine learning-based approach to unifying disparate, dirty datasets within an organization.



Tamr focuses on a bottom-up, machine learning-based approach to unifying disparate, dirty datasets.

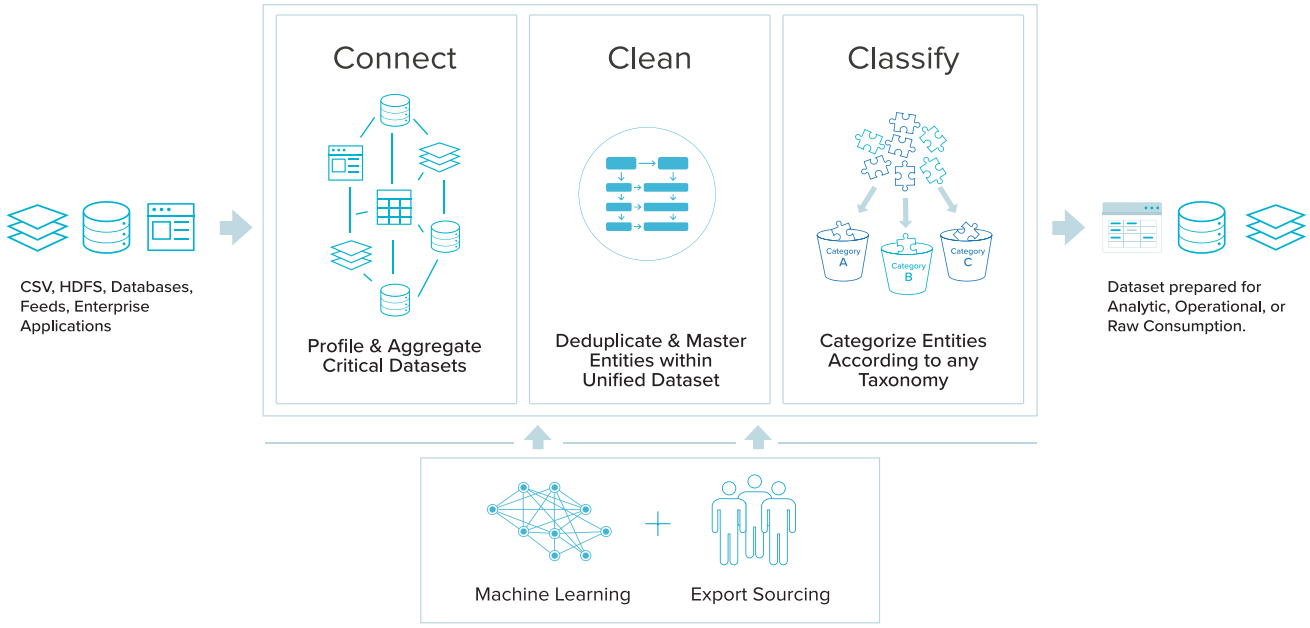
Tamr's enterprise data unification method combines machine learning and human expert guidance to unify data sources across an organization with unmatched speed, scalability, and accuracy. The platform's core capabilities include "connecting" data sources across an organization to align relevant datasets to a unified schema, "cleaning" the unified dataset through entity deduplication and mastering, and "classifying" records within the clean, unified dataset to a client-provided taxonomy for more robust downstream analysis. The resulting dataset can be consumed by multiple endpoints - from analytic tools to data warehouses - and this enables Tamr to be a very complementary data management technology to legacy solutions (such as MDM and ETL) as well as newer technologies (such as Data Catalogs, Self-Service Data Preparation Tools, and Analytic Tools). Ultimately, enterprise data unification is a proven need across a variety of use cases as well as industries and Tamr has been able to unlock significant value for customers - to the tune of hundreds of millions of dollars - through its implementation in these complex environments.

2. Tamr Product Overview

a. Company Background

Founded in 2013, Tamr was launched by start-up collaborators and data management veterans Andy Palmer and Mike Stonebraker. The two had previously co-founded Vertica Systems (a high performance database management company that sold to HP for $350M) and worked together on several other related companies. Their shared experiences forged a common belief that the core ideas behind the last 20+ years of data management thinking were failing to meet the needs of today's enterprises. With the amount and variety of data available to enterprises exploding, traditional methods for organizing it for analytics could no longer keep up. Therefore, in 2012 the team began research at MIT's Computer Science & AI Lab on a bottom-up solution for managing the radical data volume, velocity and, especially, variety in the modern enterprise. The resulting 2013 paper, "Data Curation at Scale: The Data Tamer System," described a breakthrough approach for combining machine learning and human expert guidance to unify data across thousands of sources. The paper became the guiding vision for Tamr's product and was the influence as to how Tamr acquired its name.

Key Tamr Capabilities



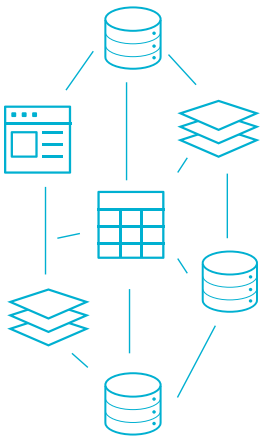
b. Core Capabilities

Tamr is an enterprise data unification platform whose patented software system combines machine learning with human expertise to automate the unification of data silos dispersed across large companies -- delivering previously impossible analytic breakthroughs. It's the only system capable of unifying data at scale and across domains quickly, accurately, and cost-effectively. In order to quickly and accurately prepare data at enterprise scale for downstream analysis, Tamr has architected three core capabilities that utilize its patented human-guided machine learning-based approach.

i. Connect

The "connect" phase starts with a definition of project goals and identification of the entities (e.g. person, place or thing) the user wants a unified view of for the purpose of downstream analysis. Within the connect phase, Tamr aligns all relevant source dataset attributes to a unified schema that is most effective and relevant for project goals. Human-guided machine learning is employed to union these datasets and offers a significant improvement in speed and scale as compared to traditional methods that rely on writing script. Tamr performs this in the following way:

- Tamr ingests data from source systems (e.g. databases, HDFS, CSVs and flat files) via APIs, JDBC connections, or a set of connectors. The Tamr platform requires data to be relatively structured / tagged prior to ingestion and can accept data in data formats such as JSON, Multivalues, Tabular, and XML
- Datasets from these sources are then profiled so users can identify the logical entity types contained within each - such as customers, supplier, etc. - and can assess the quality of their data to ensure its suitable for analysis
- Target schemas are then identified within the Tamr system to ensure the optimal attributes are represented in the final unified dataset. Users can either select attributes from data sources to build their unified schema



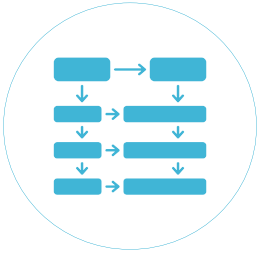
Within the "connect" phase, Tamr aligns all relevant source dataset attributes to a unified schema that is most effective and relevant for project goals.

or load in their own schema with samples of values for each attribute

- At this stage, Tamr will employ human-guided machine learning to align source dataset attributes to the unified schema
    - + Tamr will first profile values within the unified schema as a baseline for comparison. This includes metadata such as field names, descriptions / annotations, data types, and validations
    - + Tamr then profiles the same information within source attributes and uses unsupervised machine learning to identify potential matches between source and target schema attributes based on an initial model Tamr has generated
    - + In order to validate the accuracy of Tamr's matches and allow for the system to learn - turning the corner into supervised machine learning - Tamr will produce "high impact questions" - which are simple yes/no questions about matches between sample attribute pairs from the target schema and source datasets that are highly representative of other potential attribute pairs. For example, if Tamr has low confidence regarding whether or not source attribute "mailing address" is synonymous with unified attribute "street address", it will ask a user within the client organization for their feedback. This not only drives accuracy into the process to ensure trusted results but also enables Tamr's algorithms to learn from the insight so a higher percentage of the next, similar batch of attributes are matched automatically and without the need for human intervention
    - + The curator operating Tamr will identify subject matter experts within the organization to either validate or invalidate Tamr's matches via direct login to the system or out-of-band mechanisms such as email. There is no set limit to the amount of experts permitted to use the system
    - + Expert feedback is then incorporated immediately or goes through a workflow to determine the appropriate action. For example, when assessing whether or not two attributes are matches, the user may want to incorporate feedback from multiple experts. In this environment, they may want to stipulate that the system recognize, for example, the most common yes / no input among the group of experts assigned to that potential attribute pair
    - + Having incorporated expert feedback, Tamr then refines its model for more automated use in the future
- The unified dataset of a logical entity, incorporating the target schema, is now materialized and able to be exported or used in downstream Tamr capabilities

ii. Clean

Tamr's "clean" phase is designed to deduplicate and master the entities within the unified dataset efficiently and accurately through the use of human-guided machine learning. The issue of dirty, duplicative data across enterprise data systems is extremely common and a one that is very difficult to solve using conventional data management techniques. The principle function of this phase of Tamr is record linkage –allowing users to identify duplicates and / or groups within the core, unified dataset and master the records contained within it – resulting in accurate, complete analysis downstream. Within the cleaning phase of the platform:



Tamr's "clean" phase is designed to deduplicate and master the entities within the unified dataset efficiently and accurately through the use of human-guided machine learning.

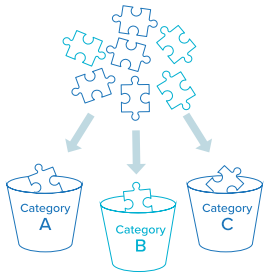
- Tamr will start with a unified dataset that is yet to be mastered and likely contains significant duplicative records. If training input regarding identification of duplicative records is available, Tamr will incorporate that into the model
- In a process similar to "connect", Tamr will apply human-guided machine learning to the unified dataset in order to cluster or group records that likely relate to the same entity
    - + Unless training data is provided upfront, Tamr will employ unsupervised machine learning to detect record similarity by analyzing all attributes / attribute values for a pair of records
    - + Tamr will then generate suggestions as to which records may be duplicative based on its modeling efforts and generate simple high impact questions regarding record pairs that are representative of other potential record pairs in the dataset. For example, if Tamr has low confidence that the person referenced in record "J. Smith" is the same person referenced in record "John S" then it will ask an expert in the organization who deals with clients to provide feedback as to whether these are distinct or matching records. Like the "connect" phase, this not only drives accuracy into the process to ensure trusted results but also enables Tamr's algorithms to learn from the insight so a higher percentage of the next, similar batch of attributes are matched automatically and without the need for human intervention

+ The curator operating Tamr will identify subject matter experts within the organization to either validate or invalidate Tamr's matches and then expert feedback is incorporated either immediately or goes through the previously defined workflow to determine the appropriate action
+ Having incorporated expert feedback, Tamr then refines its model for more automated use in the future

- For some downstream use cases, grouping clusters of records is sufficient for consumption. In other use cases, organizations may prefer the groups of records be mastered and merged into a single record. Tamr's merge logic is robust and includes options for selecting values for attributes within clustered records that include:
  + Most common value for an attribute
  + Value selection from a known 'trusted source' dataset (i.e. a dataset the user knows to be most trustworthy)
  + Steward nominee - where experts can select the appropriate value for a certain attribute
- The cleaned, unified dataset of a logical entity is now materialized and able to be exported or used in downstream Tamr capabilities

iii. Classify

Once a clean, unified dataset of a particular entity has been produced by Tamr, the user has the option of "classifying" the records to a company-specific or commonly used taxonomy for more in-depth analytic capabilities downstream. This is particularly true within use cases such as supply chain or procurement analytics - where taxonomies help organize entities into logical groupings for business and analytic purposes. Tamr's classify phase operates in the same manner as the connecting and cleaning phases do, leveraging Tamr's unique blend of human-guided machine learning to rapidly and accurately categorize records to the deepest levels of a provided taxonomy. Within the classification phase:

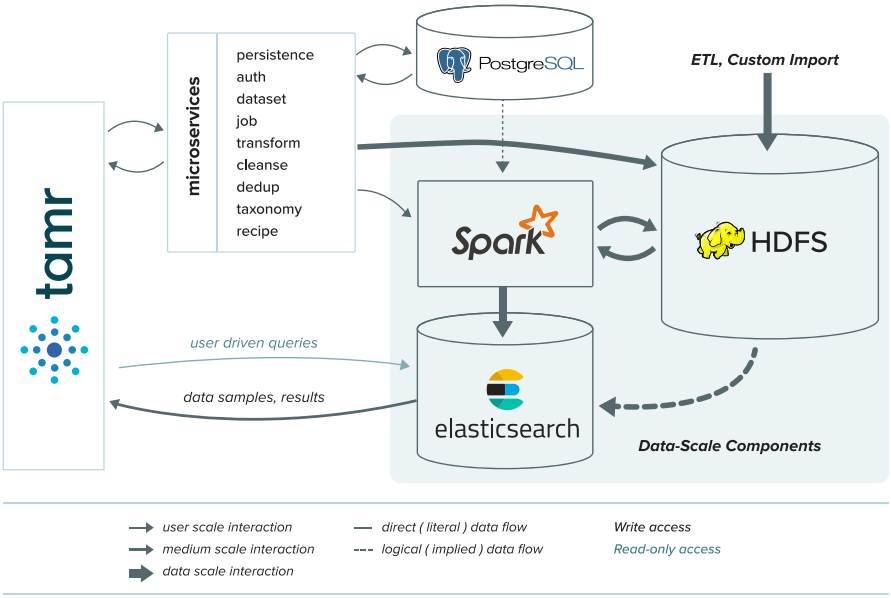


Once a clean, unified dataset of a particular entity has been produced by Tamr, the user has the option of "classifying" the records to a company-specific or commonly used taxonomy for more in-depth analytic capabilities downstream.

- Tamr will start with a clean, unified dataset focused on logical entities (such as parts) that have yet to be categorized to an organizational taxonomy. The first step in the classification process is to load the target taxonomy into Tamr with sample records of entities related to each branch included - where Tamr can then identify the words / tokens related to each branch of the taxonomy
- Tamr will then apply human-guided machine learning to the unified dataset in order to appropriately categorize each record in the unified dataset to a specific taxonomy
  + Tamr will profile the values within each record of the unified dataset and use it's machine learning algorithms to identify matches between words contained within values of each dataset record and words associated with each category of the taxonomy. This will enable Tamr to accurately suggest a classification for each record based on the initial model it generates
  + Tamr will then produce simple high impact questions regarding whether or not certain records, that are representative of a large portion of the unified dataset records, are categorized appropriately. For example, if Tamr has low confidence regarding whether or not a record pertaining to "1 inch turbine bolts" is in fact part of the "Bolt" category within the organization's taxonomy, it will ask an expert within the client organization for their feedback - driving accuracy and enhancing future automation
  + Expert feedback via direct login or out-of-band mechanisms such as email is then incorporated into the dataset and Tamr's models in accordance with the client's expert feedback workflow
- Once records have been classified via Tamr's workflow, Tamr will add new fields to each unified dataset record indicating how that record is categorized. For example, if Tamr is classifying a record to the 4th level of an organization's taxonomy, it will add 4 fields to each record indicating how it is categorized

After classification, the connected, cleaned, and classified dataset is ready for consumption. Tamr is flexible in its consumption options via its APIs - whether it be via an analytic tool, operational database, or simple Excel / CSV file. This is in large part due to the flexibility Tamr has in supporting multiple operational and analytic projects within the enterprise - as the need for centrally curated, trusted datasets of key organizational entities is virtually limitless.

Tamr Product Architecture



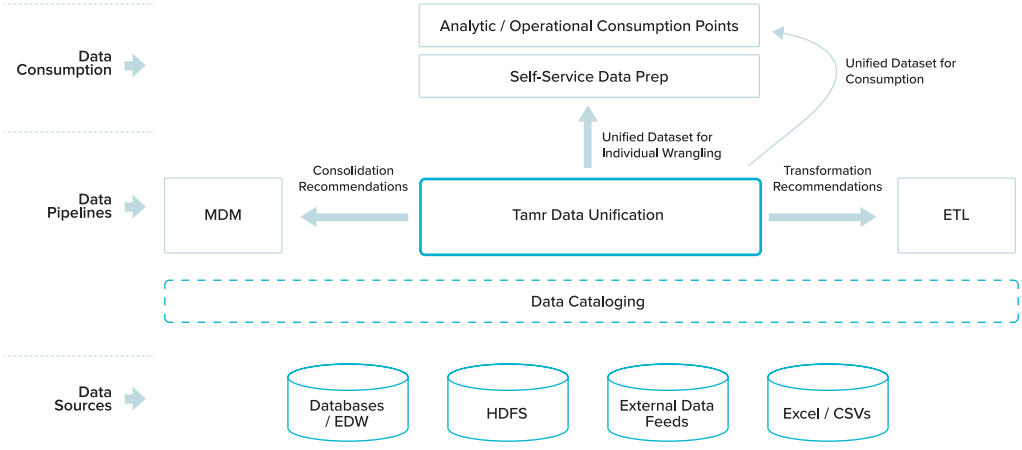
Tamr's Product Architecture Is Enterprise-Grade, Flexible, And Incorporates Innovative Technologies

3. Product Architecture

The Tamr platform is designed to take advantage of recent advances in flexibility, scalability, and ease of administration. The application layer is composed of an array of loosely-coupled microservices providing a broad array of capabilities, while simultaneously allowing flexibility in how the application is deployed and scalability of individual application components. The data processing layer assembles highly scalable components to provide both high-volume data processing and low-latency search and filtering. The overall system scales down to a single, modest server for trials and up to multi-node data lake infrastructure to tackle large, production challenges.

The individual microservices that comprise the application run behind a single facade that makes them look and feel like a single application to the end user. This assembly also interacts with organizational backing services, such as LDAP for user authentication, and a relational database for storing user preferences and the like. The data itself stays within multi-node, scale-out infrastructure taking advantage of technologies such as HDFS for distributed, reliable storage, and Spark for distributed, in-memory computation. Finally, an Elasticsearch cluster powers a richly interactive front end, while keeping query latency and page load time short so that users can focus on gaining insights from their data.

Tamr Market Positioning



Tamr complements existing Data Management & Analytics Investments

4. Market Positioning

Tamr can operate in a variety of capacities within an enterprise's data environment - as both a system of record and a system of reference. The platform is designed to operate in a complementary nature to most big data investments and solve the large "garbage in, garbage out" issues. Tamr is most often compared to and can complement the following technologies:

- **Data Catalogs** - Tamr has some capabilities around profiling of datasets; however, dedicated data catalogs that can discover datasets of interest related to a particular entity and foster secure collaboration among users can enhance Tamr's value proposition downstream. The data sources discovered and analyzed within a data catalog can serve as an input to Tamr

- **Master Data Management** - While Tamr could be deployed as a de facto master data management solution, it most often complements legacy master data management solutions. In particular, Tamr can help MDM systems by acting as a system of reference for record consolidation - making the operation faster and much more scalable

- **ETL** - Much like master data management, Tamr acts as a system of reference for ETL solutions. In particular, Tamr can suggest transformations for ETL solutions regarding which certain records are in fact referencing the same entity - helping with the speed and scale of executing transformations

- **Self-Service Data Preparation** - Tamr is complementary to self-service data preparation tools in the market. Often times, these tools are targeted at data sets that are fairly connected and cleaned already and no robust machine learning is needed. Self-service data preparation tools do, however, allow for individual user data curation (for example, eliminating unwanted records) which is a valuable downstream function of the Tamr platform

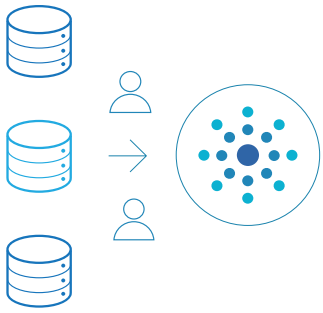
- **Analytic Tools** - Tamr is complementary to analytic and visualization tools - which can be used as a Tamr consumption method. Unified, clean datasets are critical to analytic and visualization platforms - as it solves the "garbage in, garbage out" dilemma that plagues most large organizations undertaking analytic initiatives

5. Use Case Examples

Tamr's domain-agnostic approach to enterprise data unification makes it a great fit for companies across all verticals and applicable to a wide variety of use cases. Tamr has enabled enterprises to centrally curate data from suppliers and parts to products and customers. Below are a couple of examples:

Multinational Industrial Company (Suppliers, Parts, & Services): Tamr was working with a multinational industrial organization that wanted to gain enhanced visibility into their supply base - in particular what parts they were purchasing across the entire enterprise and from whom. This was extremely difficult to do using traditional approaches to integrate and clean datasets given the size and complexity of their data environment. The company approached Tamr to help with this data unification problem and in doing so, Tamr connected and cleaned their procurement data (representing $60 billion in spend) across 8 business units to fuel analytic outcomes. The results included first-time visibility into spend (suppliers, parts, services) that enabled the company to unlock $380+ million in cost-savings opportunities (projected $500+ million), including a 10x ROI in Year 1

Large Media Company (Company Entities): Tamr had engaged with a large media company that was undertaking an initiative to create an enterprise-wide data model after years of growth both organically and inorganically. The company sells organizational data and had an existing data integration process internally, but due to its heavily manual nature, it lacked the speed and scale needed to keep up with the changing environment. The client needed help with record deduplication and Tamr was able to deliver highly impactful results. This included expediting data integration efforts by several months while reducing the manual effort needed to integrate datasets by over 40%. Finally, given Tamr's human-in-the-loop workflow, the client accomplished this while achieving remarkably high accuracy (precision and recall rates of over 95%)



Tamr's domain-agnostic approach to enterprise data unification makes it a great fit for companies across all verticals and applicable to a wide variety of use cases. Tamr has enabled enterprises to centrally curate data from suppliers and parts to products and customers.