Tamr

The DataOps Approach to Data Mastering

Taking a Best-of-Breed Approach to Data Management and Analytics

Taking a Best-of-Breed Approach to Data Management and Analytics

The gap between business needs around data quality and availability, and the reality of the state of enterprise data has never been larger. As enterprise data grows exponentially, decades of technologies have failed to address the challenge of large data volume and variety.

Automating data infrastructure and using the principles of DevOps—designed for operations, repeatability, automated testing—is critical to keep up with the dramatic pace of change in enterprise data. DataOps is an agile approach to data management that many data leaders have adopted to accelerate data-driven business outcomes. It addresses both speed and scale, and a key part of DataOps is to take a best-of-breed approach to data solutions. By decoupling key components of data management, such as data mastering and governance, teams are able to tackle key data challenges with tools purpose-built for the task, and stay more agile as the data landscape and analytical projects evolve within the organization.

One key area of focus for DataOps teams is data mastering. Today, many organizations are facing the reality that their significant investments in traditional MDM systems—which served to address the volume of data—have failed to keep pace with the growing number of highly-variable data sources needed to answer critical business questions. The "waterfall" approach to designing rules and iterating based on results haveslowed—and in some cases, failed—data and analytics projects.

Connected Enterprise Data Drives More Value

Tackling data mastering as an iterative process enables organizations to accelerate how quickly they can connect and master new data sources (from CRMs to homegrown databases and third-party data sources). Connected data—with clear relationships established between datasets—such as customers and transactions to products and interactions, drives exponential value to an organization. Connected data is the foundation to meaningful analytics and driving real business outcomes. Tamr's data mastering solutions enables data teams to accelerate their ability to connect datasets and answer critical business questions.

In this document, you'll learn:

□ *How Tamr's data mastering solutions power analytic insights*

□ *An overview of Tamr's capabilities*

□ *Tamr's core competencies in a best-of-breed data management ecosystem*

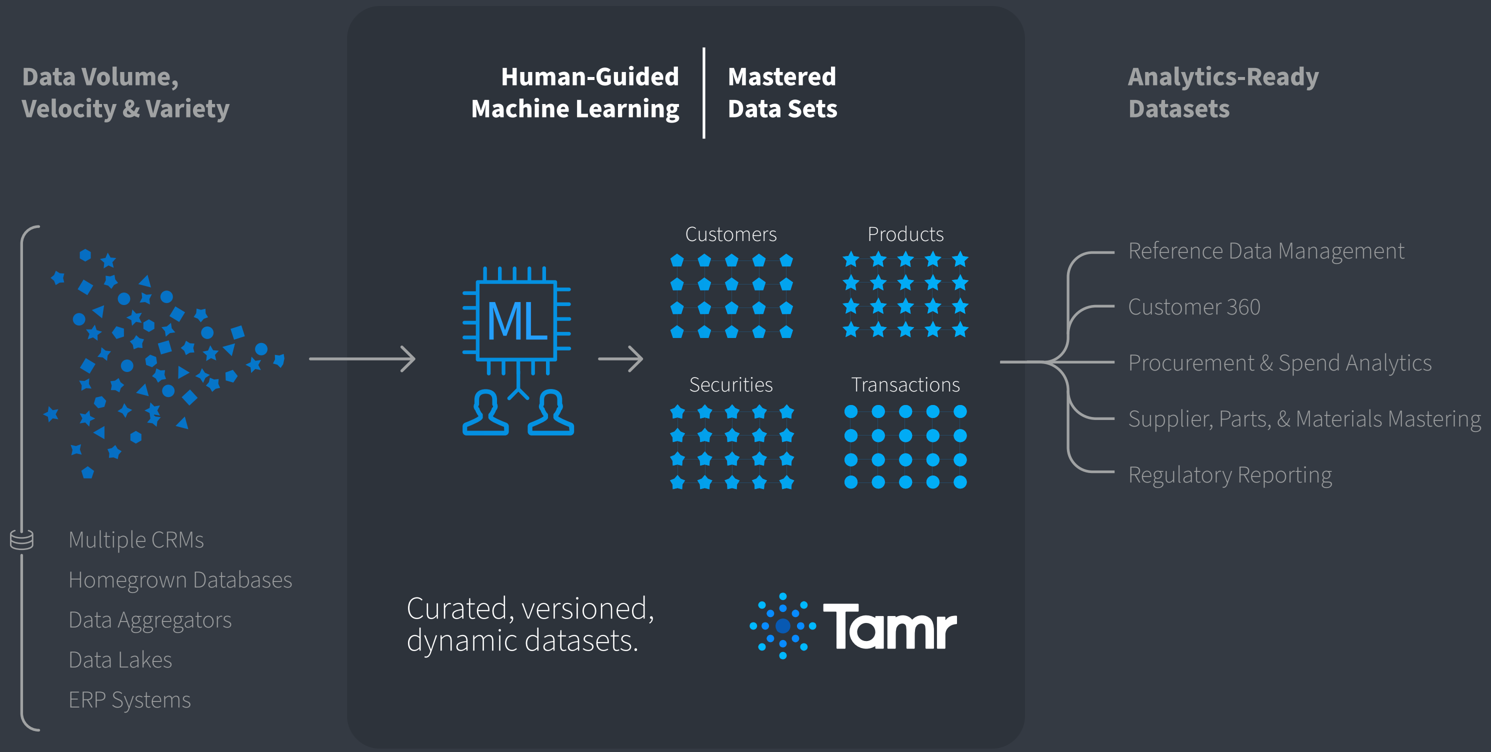
□ *The importance of cloud-native, open architectures for scaling*

□ *How Tamr complements or can replace MDM solutions*

Driving Business Outcomes Faster with Machine-Driven Data Mastering

Tamr masters data at enterprise-scale so that data is ready and curated for analytics programs and digital initiatives (such as AI/ML programs or shifts to the cloud). Tamr's cloud-native data mastering technology combines machine learning-based models, human feedback from data experts, and rules to curate and accurately publish data from large, diverse data sets, enabling effective data consumption in analytics and business processes.

Tackling Enterprise Data Debt



Data Volume, Velocity & Variety

Multiple CRMs
Homegrown Databases
Data Aggregators
Data Lakes
ERP Systems

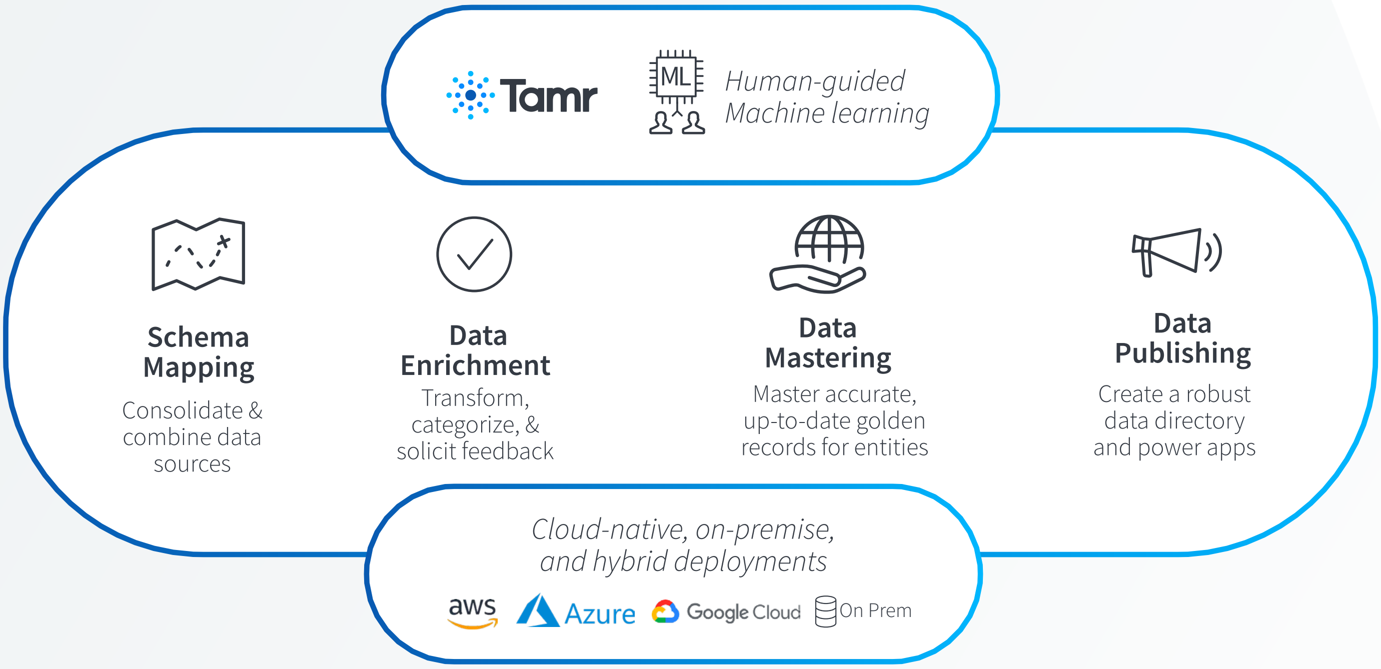
Human-Guided Machine Learning | **Mastered Data Sets**

Curated, versioned, dynamic datasets.

Analytics-Ready Datasets

Reference Data Management
Customer 360
Procurement & Spend Analytics
Supplier, Parts, & Materials Mastering
Regulatory Reporting

Tamr makes it easy for organizations to connect internal and external data sources, cleanse and consolidate them, and create curated datasets that power analytic outcomes. Tamr takes a machine learning-first approach to data mastering, with intuitive workflows for data experts and business users to train the ML models.



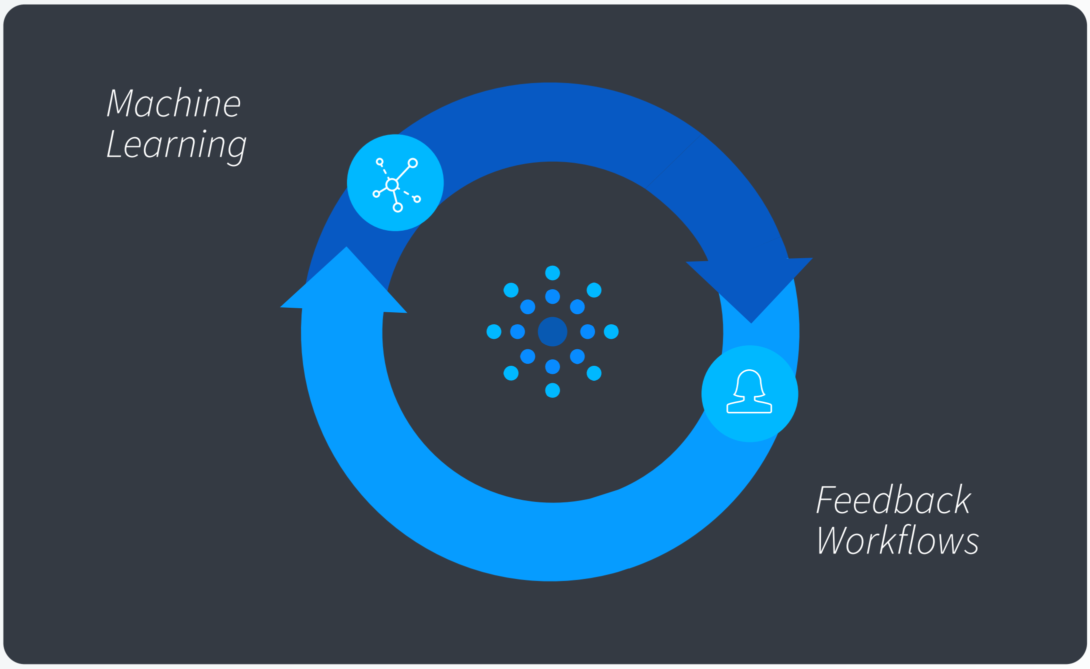
Tamr
ML
Human-guided Machine learning
Schema Mapping
Consolidate & combine data sources
Data Enrichment
Transform, categorize, & solicit feedback
Data Mastering
Master accurate, up-to-date golden records for entities
Data Publishing
Create a robust data directory and power apps
Cloud-native, on-premise, and hybrid deployments
aws
Azure
Google Cloud
On Prem

The technology reduces manual workflows needed to consolidate, categorize, and create golden records by up to 90%. And with workflows to engage key stakeholders early and often, organizations can stay more agile and accommodate emergent data requirements. The result? Lower cost of ownership for data mastering projects, and faster delivery of cleansed, up-to-date enterprise data.

Outcomes: A US Financial Institution estimates ~$20M in annual savings from deploying Tamr for one data mastering project, due in large part to hours saved on manual data preparation and lower compute costs.

Engaging Data Experts Effectively

At the core of Tamr's technologies is the ability to engage data experts and data stewards through simple yes or no questions to train the machine learning models. Tamr's ML algorithms have been honed over seven years to master data on customers, products, suppliers, and more. The machine performs most of the heavy lifting to consolidate disparate data sources, categorize them (e.g., classifying spend), and transforming data (e.g., dollars to euros). When Tamr's models do not meet a configured probability score (e.g., the model places a schema mapping match at 75% probability), a workflow begins to engage data experts in data remediation decisions. As data experts answer questions over time and train the ML models, probability matching increases.



This approach drives higher data matching accuracy than traditional rules-based models; our studies have shown 90%+ accuracy for with Tamr's technology as compared to 50-80% accuracy with rules-based models. This accuracy accelerates time-to-insights for critical business decisions, saving data scientists a significant amount of time on data preparation and manual consolidation workflows. And data teams can stay more tightly-aligned with business teams to drive analytic outcomes faster.

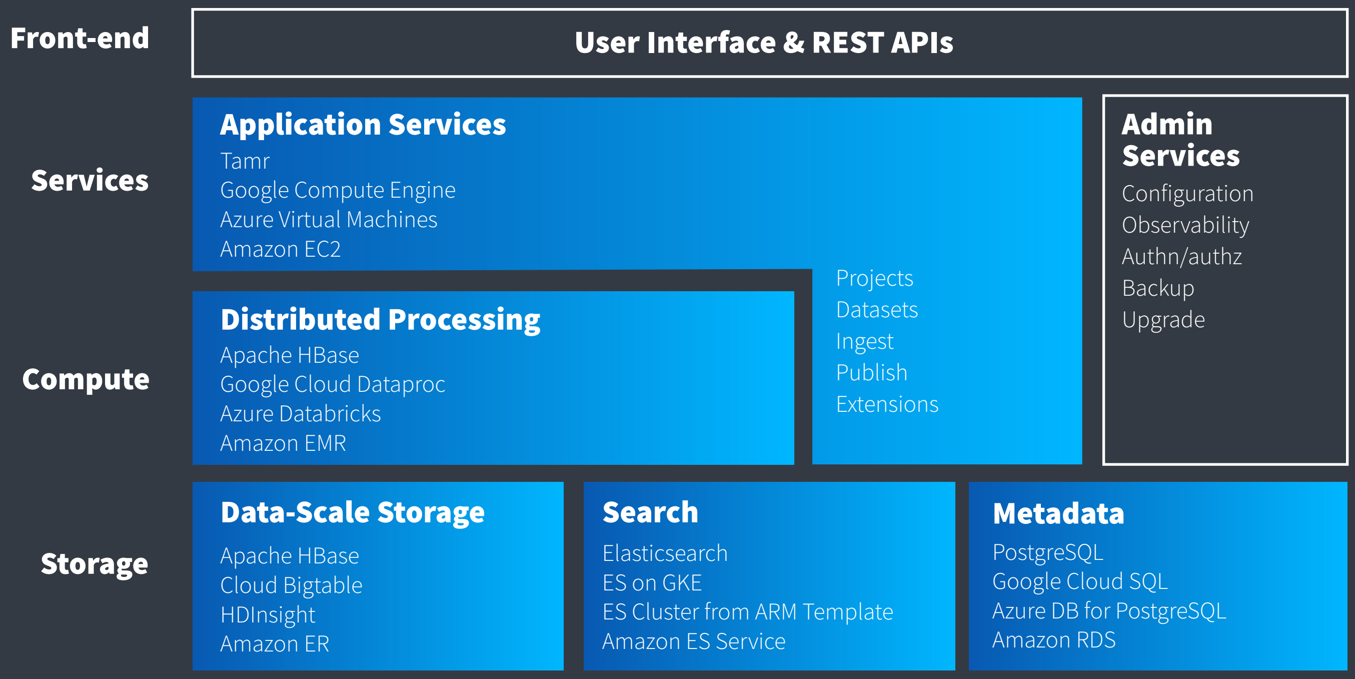
Outcomes: In 30 hours of work with a global financial institution, Tamr accurately classified 75% of $12 billion Euros in spend, representing 6 million records. The engagement reduced manual workflows by 90% for ERP consolidation

.

The Importance of Cloud-Native, Open Architecture for Data Mastering

The DataOps ecosystem should resemble DevOps ecosystems; modular, interoperable components that can scale over time. This approach offers more flexibility as teams modify and grow data pipelines and introduce new technologies.

Tamr's architecture is built on the same principles: interoperable, best-of-breed technologies comprised of RESTful APIs that sit on top of proven big data components like Hadoop, Spark, Elastic, and Postgres. In addition, Tamr partners with leading cloud providers (Google GCP, Amazon Web Services, Microsoft Azure) and leverages cloud-native capabilities to improve scalability and lower compute and storage costs.



Front-end

User Interface & REST APIs

Services

Application Services

Tamr
Google Compute Engine
Azure Virtual Machines
Amazon EC2

Projects
Datasets
Ingest
Publish
Extensions

Admin Services

Configuration
Observability
Authn/authz
Backup
Upgrade

Compute

Distributed Processing

Apache HBase
Google Cloud Dataproc
Azure Databricks
Amazon EMR

Storage

Data-Scale Storage

Apache HBase
Cloud Bigtable
HDInsight
Amazon ER

Search

Elasticsearch
ES on GKE
ES Cluster from ARM Template
Amazon ES Service

Metadata

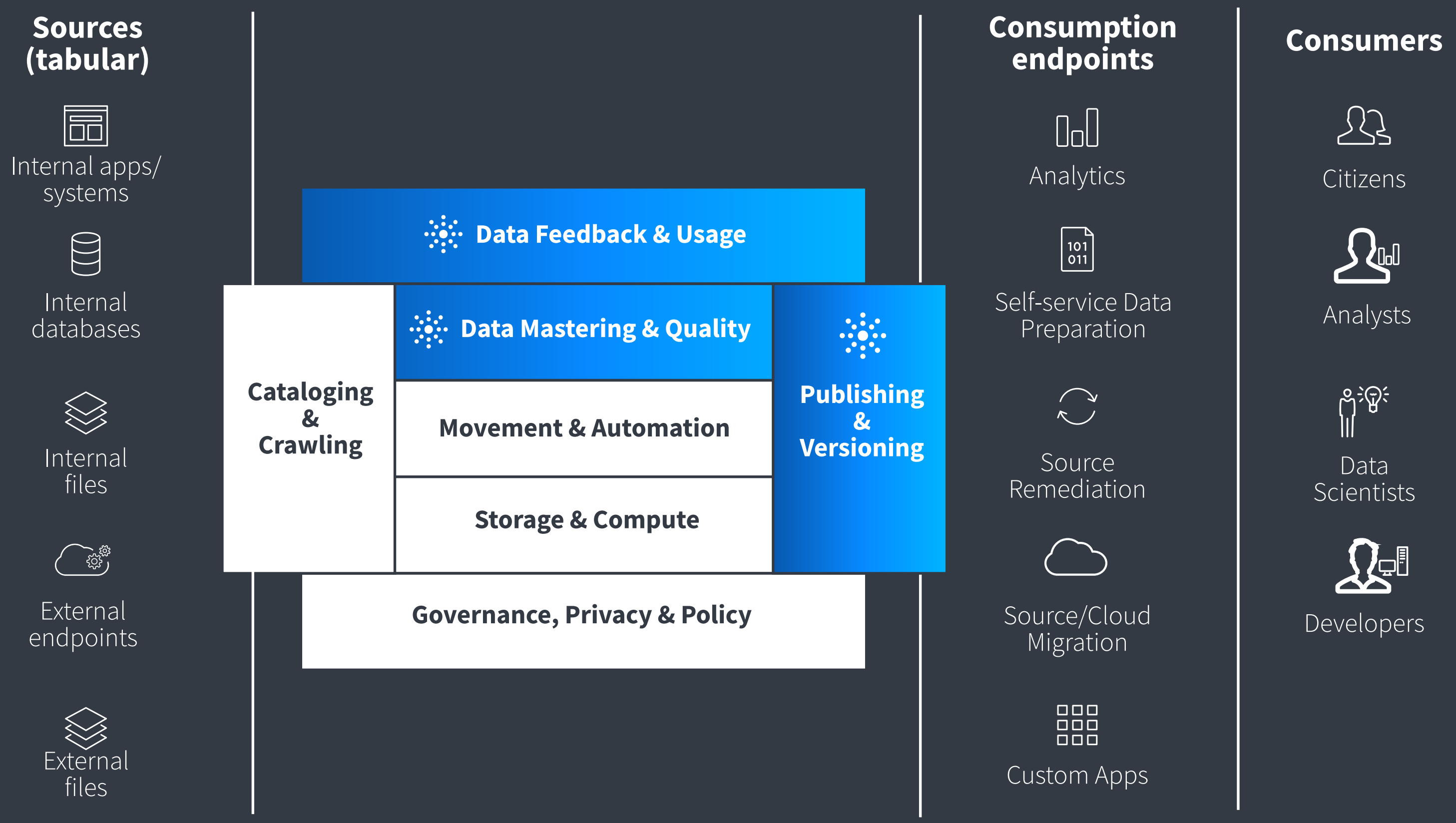
PostgreSQL
Google Cloud SQL
Azure DB for PostgreSQL
Amazon RDS

In addition to loosely coupled technologies, and avoiding one-size-fits-all platforms for all data management needs, the shift to the cloud is a primary focus for organizations looking to scale operations and lower cost of ownership. The core compute services available from the large cloud providers are powerful and easy to scale up/ out quickly as required with little to no capital investment. Tamr offers the only data mastering solutions on the market today that support cloud-native, on-premise and hybrid deployments, supporting organizations at all phases of their digital transformations.

Outcomes: Organizations save nearly $400,000 yearly by running Tamr in the cloud and leveraging its cloud-native capabilities to scale workload based on need. Running traditional data mastering solutions on-premise cost approximately $460,000 every year, compared to $64,000 a year by deploying Tamr in the cloud.

Tamr and the DataOps Ecosystem

Tamr can operate in a variety of capacities within an enterprise's data environment, including both as a system of record and a system of reference. The platform is designed to operate in a complementary nature to big data investments, ensuring that data across the stack is complete, up-to-date, and cleansed.



In the sample reference architecture above, Tamr's core competencies are highlighted in blue. In addition, below are examples of how Tamr complements or is differentiated from common data technologies:

- **MDM:** Tamr combines machine learning with data expert input to accelerate data mastering at scale and lower the total cost of ownership of data projects. Most MDM solutions on the market consolidate and master data by relying on rules, which are costly to build and maintain, as well as often failing to address the variety of data in large datasets. Tamr can integrate with and augment existing MDM solutions, among other data sources, yielding improved mastered data, while leveraging the traditional MDM platforms for other capabilities such as data governance.

- **Data Catalogs:** Tamr includes capabilities that profile datasets, including our ability to power analytic strategies by categorizing data with proprietary or third-party taxonomies. As data catalogs can crawl multiple sources and discover data and schemas, they can serve as an input for Tamr to map schemas and create cleansed, de-duplicated golden records for downstream systems. Tamr's mastered datasets can also be integrated in the catalog, improving visibility and access to these curated sources along with other enterprise data.

- **ETL:** One area of overlap with ETL solutions is that Tamr includes a full-featured data transformation capability in service of unifying and mastering highly variable datasets. Some customers leverage Tamr as a system of reference for executing transformations, given the scalability and performance of the Tamr functionality, with other customers relying upon their incumbent ETL tooling to move data through the Tamr system and to orchestrate automated data pipelines.

- **Data Quality:** Tamr's data mastering solutions enable organizations to create complete, de-duplicated records that enhance data quality overall. To improve data quality further and promote accessibility, Tamr offers several data feedback capabilities for data experts to provide data remediation input and feedback at the point of engagement (e.g., from within Tableau or Salesforce). Organizations typically leverage Tamr as a data source for further data quality initiatives, such as data validation and data repair.

- **Data Governance:** Tamr's data mastering solutions contribute to data governance programs in several ways: First, Tamr supports role-based access controls. Second, by ensuring that datasets are cleansed and curated with surrounding metadata on data sources for a mastered record, organizations are positioned to report effectively on data lineage. Similar to data quality programs, many customers leverage output from Tamr to focus their data governance efforts (e.g., policies) downstream.

- **Self-Service Data Prep:** Tamr complements downstream data prep by curating the mastered data being consumed by analysts and data scientists to improve productivity and analytic veracity. Tamr is typically a source for downstream data prep activities, where analysts leverage the curated, mastered datasets from Tamr in self-service data prep tools for data discovery and profiling.

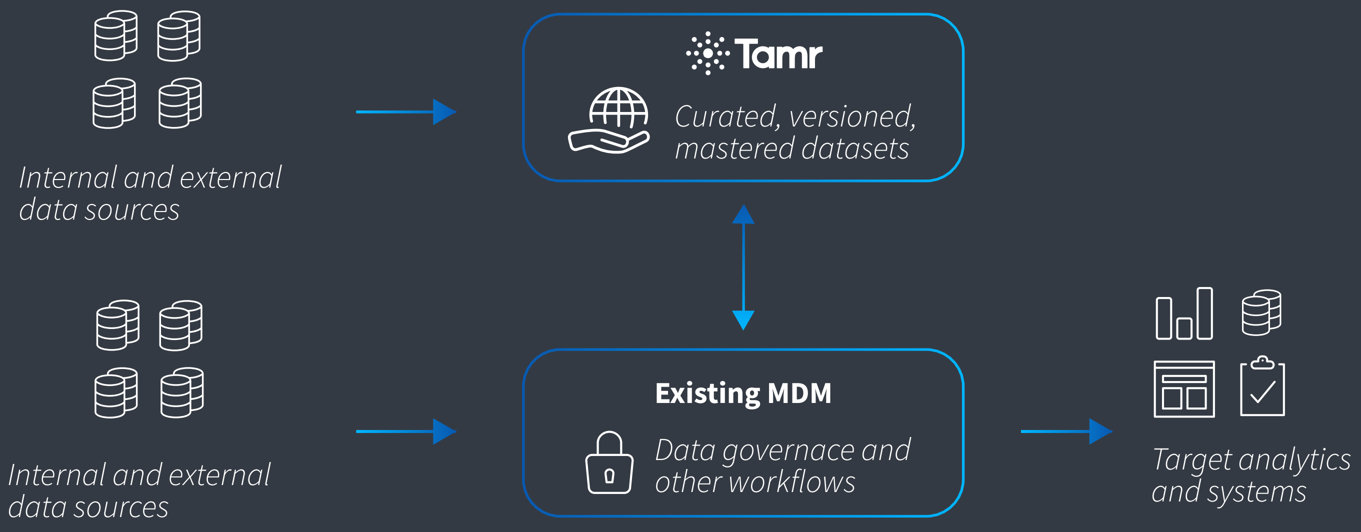
Tamr's open architecture, APIs, and cloud-native capabilities provide flexible integration with legacy and new data pipelines.

Tamr Data Mastering and Traditional MDM Systems

Tamr is interoperable with existing MDM solutions, or can be deployed as a MDM solutions for organizations focused on data mastering.

Integration with Existing MDM Solutions

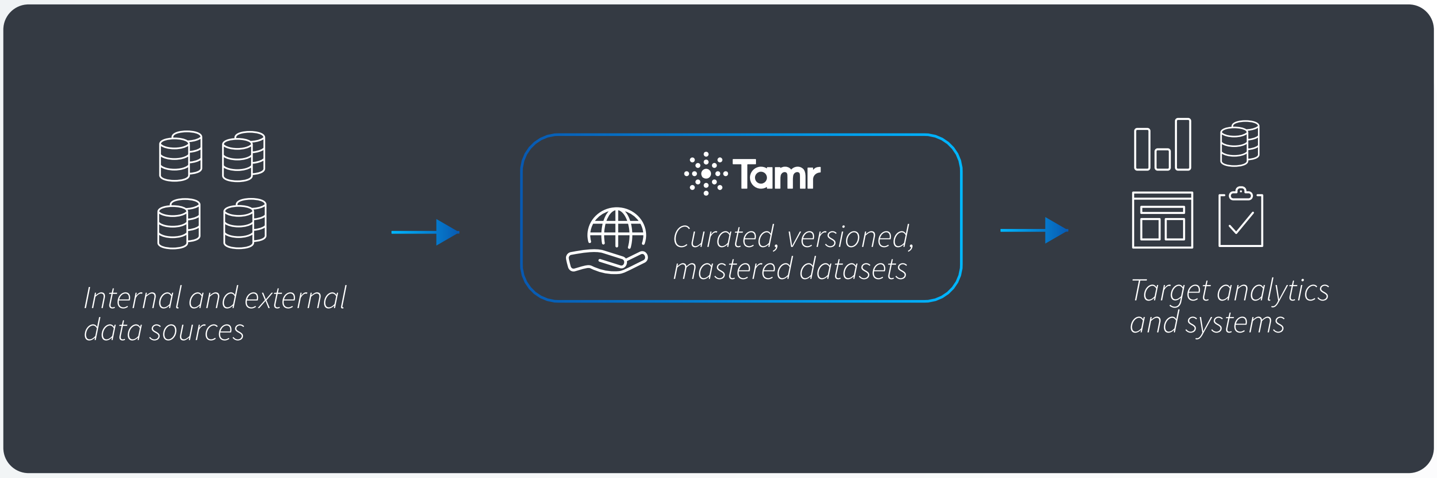
Some organizations may have a traditional MDM solution deployed along with business processes tightly built around it. In the example below, Tamr masters key data entities, ingesting data from internal sources (including the MDM system) and external, third-party sources:


Internal and external data sources
Tamr
Curated, versioned, mastered datasets
Internal and external data sources
Existing MDM
Data governance and other workflows
Target analytics and systems

The output from Tamr is curated, versioned datasets that are integrated back into the MDM solution for data governance capabilities, or to support other business workflows that are aligned with the MDM solution.

Tamr Deployed as MDM

Tamr can be deployed in place of an MDM solution to ingest data from disparate data sources, consolidate the data, and output curated, mastered data sets to power business intelligence platforms or other downstream systems.


Internal and external data sources
Tamr
Curated, versioned, mastered datasets
Target analytics and systems

Tamr generates golden records and clusterIDs, which group together matching source records, and serves as a system of record for downstream systems. Tamr's mastered datasets (including golden records and clusterIDs) reference the original data source so that data teams can track data lineage in Tamr and downstream systems.

Connecting Data to Drive Better Business Outcomes Faster

With Tamr, data scientists and analysts spend less time on manual data processing and preparation, enabling them to connect enterprise data far more efficiently than ever before. Through reduced manual workflows, data teams are empowered to drive business outcomes. From Customer 360 to supply chain optimization, Tamr helps leading organizations across the globe solve several business challenges that all tie to the need for timely, connected, accurate enterprise data to power analytic outcomes. Visit www.tamr.com to learn more.

U.S. DEPARTMENT OF HOMELAND SECURITY

TOYOTA

gsk

Scotiabank

Johnson & Johnson

NOVARTIS

SOCIETE GENERALE

CAA CreativeArtistsAgency

HESS

Next Steps

Are you ready to tackle data-driven business challenges? Connect with Tamr to bridge the gap between data and analytics outcomes.

To learn more about Tamr, please visit www.tamr.com or contact us to schedule a meeting and a demo.

SCHEDULE DEMO

Tamr

Tamr

About Us

Tamr is the leading data mastering company to accelerate data-driven usiness outcomes. Industry leaders like: Toyota, Societe Generale, GE, and Thomson Reuters trust Tamr to manage their enterprise data as an asset. Tamr's unique approach of using human-guided machine learning algorithms to accelerate data mastering projects lets the world's largest organizations enhance their data operations, rapidly activate latent data, and increase the velocity of business outcomes through data-driven insights. With a cofounding team led by Andy Palmer (founding CEO of Vertica) and Mike Stonebraker (Turing Award winner) and backed by investors including NEA and Google Ventures, Tamr is transforming how companies get value from their data.

To find out more, visit **tamr.com**