EBOOK

How to Deliver Better Analytics with Fully-Managed Data Mastering

Tamr

Over the past 10-15 years, organizations have spent countless hours democratizing data and analytics - providing everyone, regardless of their role, a self-service way to easily work with the data, understand the data, and confidently use it to make data-driven decisions.

But as organizations forge down the path of providing analytics to the masses, many have encountered fundamental challenges:

> Data lives in disparate data silos, and integrating it requires a significant number of transformations beyond simple joins, such as data cleaning, standardization, and entity resolution.
>
> Data quality issues are often resolved at the "edge" by analysts for the purpose of a single analysis or dashboard, adding little value to the broader organization.

Traditionally, companies solved these challenges using solutions such as traditional Master Data Management (MDM) and centrally-managed extract, transform, and load (ETL) pipelines. But these solutions are limited by the constraints of the on-premises environments they were initially built to support.

For example, traditional MDM solutions require humans to manually develop and maintain match rules. And they require even more humans to review and approve most changes to master data. This approach works fine if your data is static, but it quickly becomes untenable as the organization adds more tables and attributes that they need to master.

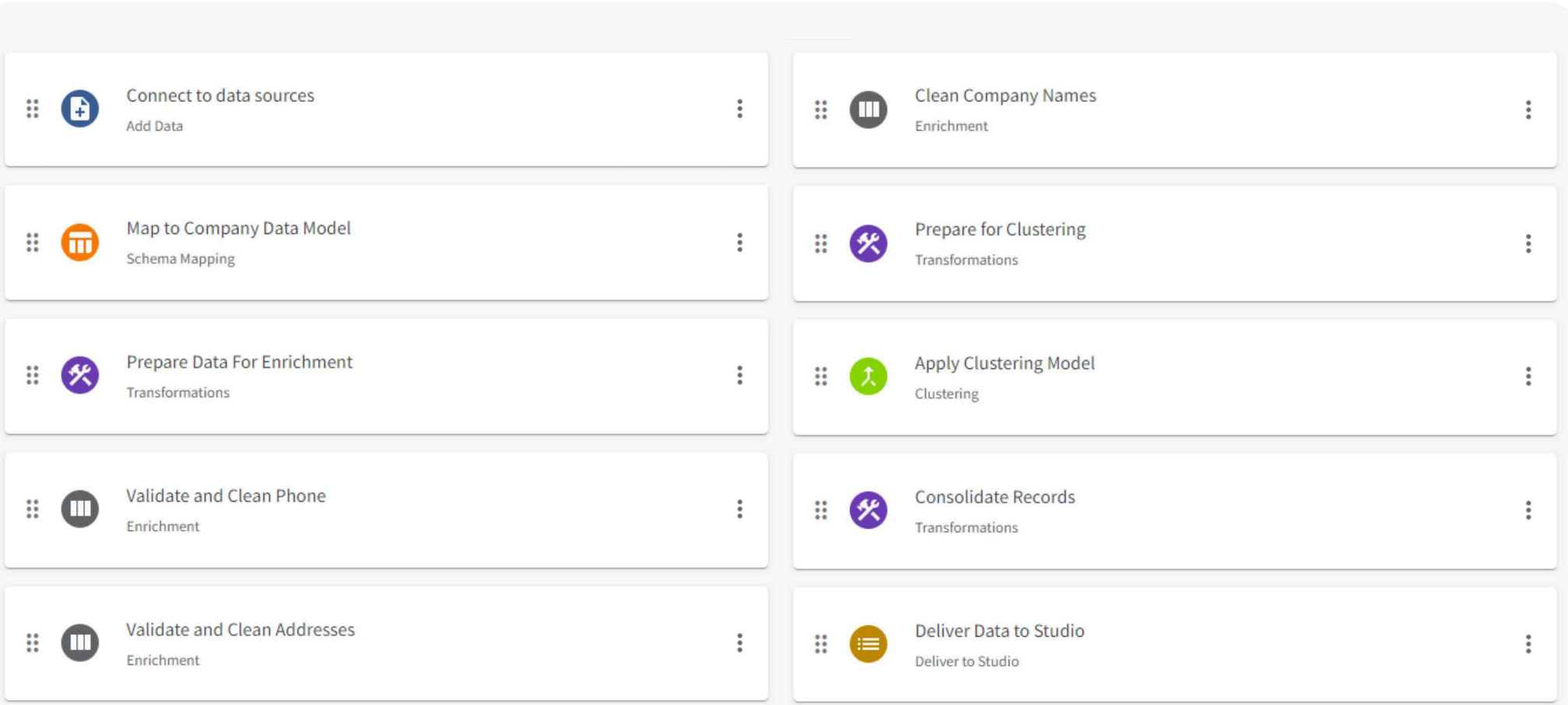
Further, while ETL pipelines are necessary, they are an insufficient solution to the problem. ETL pipelines provide flexibility to integrate with best-of-breed services such as entity resolution libraries and enrichment APIs. But the benefits of that flexibility are short-lived. As the need to support new tables and attributes grows, someone in the organization needs to review every piece of the pipeline - and potentially give it a significant overhaul.

To address these challenges, Tamr developed **Tamr Cloud**, a machine learning-driven data mastering platform that enables data architects, engineers, and analysts to efficiently curate and enrich disparate data sources for accurate, complete analytics. It's built on top of Tamr's patented technology that is in production at nearly 100 organizations, including the US Air Force, Blackstone, GSK, Toyota, and Staples.

Unlike traditional MDM and ETL solutions, Tamr built Tamr Cloud to handle a diverse and changing set of attributes through an approach that is scalable and minimizes overhead.

Tamr Cloud uses a modular architecture where the data transformation processes are abstracted into small reusable blocks that are based on attributes, which simplifies the process and makes it customizable.

And because it uses a machine learning-based approach, Tamr Cloud easily supports the addition of new attributes without having a human review the whole process and rewrite the entire data pipeline.



Example modules ("steps") in Tamr Cloud

Each of these modular steps is like a "Lego piece" that can be swapped in and out of the pipeline depending on the dataset and the specific use case, allowing for further simplification and customization of the process.

Creating an Entity Layer with Tamr Cloud

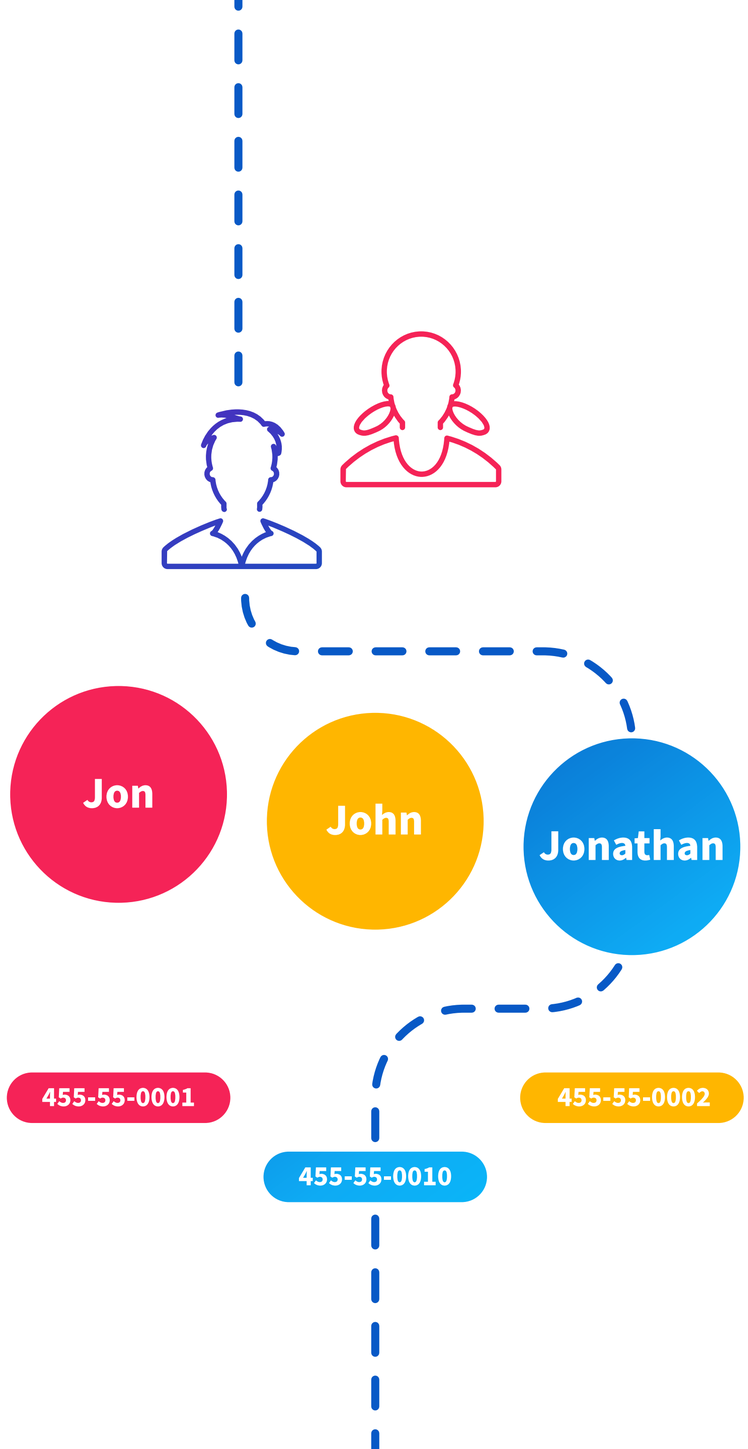
At Tamr, we believe that the only way to keep the data that users consume accurate is to manage it first at the entity level and then at the attribute level. Why? Because that is how we track and manage data in the real world.

- **At the entity level**, different attributes will identify different entities. For example, you may identify a person by name, gender, or social security number. And you may identify an organization by attributes such as name, addresses, tax ID, or website.

- **At the attribute level**, each attribute is also treated differently. For example, the people monitoring phone number attributes (national numbering service) are very different from the people who care about address attributes.

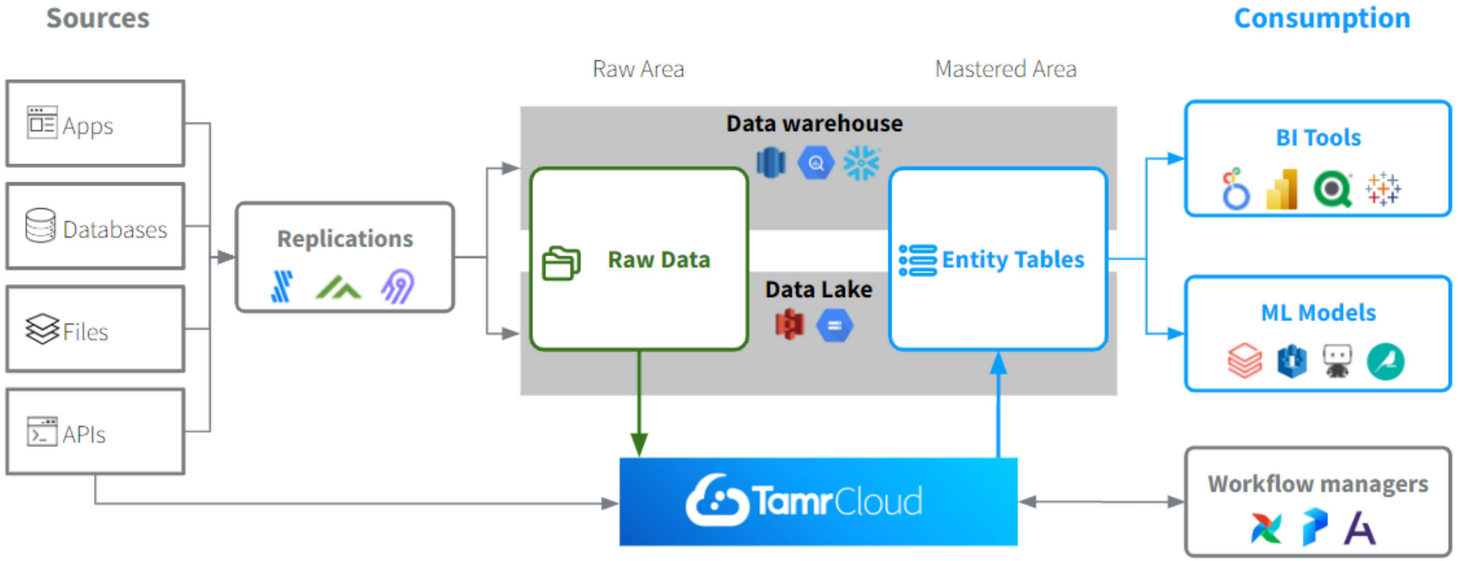
We treat the data this way because there is no 'one-size-fits-all' solution for every entity and every attribute. Currently, Tamr Cloud natively supports organizations and people entities as well as their respective main attributes. We're planning to release many more entities in the future by taking advantage of the modular architecture that enables us to continuously release new entities and attributes.

To enable flexibility in the Tamr Cloud, we utilize a customizable schema that enables custom attributes. We've designed Tamr Cloud's underlying architecture so that it can support any entity imaginable as well as an ever-changing number of attributes.



Tamr Cloud and the Modern Data Infrastructure

Tamr architected Tamr Cloud to integrate seamlessly with today's cloud warehouse and lakehouse architectures, coupling with products such as Fivetran, Snowflake, BigQuery, and Looker. Let's take a look at how Tamr Cloud fits into the ecosystem.



Tamr Cloud takes unmastered data from the raw zone of the data warehouse or lakehouse and delivers the entity layer for downstream consumption through the continuous cleaning and curation of data.

For users who consume data curated by Tamr Cloud, there are two main benefits:

1. **Clean persistent identifiers:** Tamr Cloud provides each matched cluster with a Tamr ID, a persistent identifier that is mapped to original source primary keys, so that users can easily join the output entity table to other source attributes once it's mastered.

2. **Full access to the attribute catalog of an entity:** Through Tamr Cloud, data consumers can have a holistic view of an entity across data sources and understand what is known - or not known - about this entity (such as a customer) and the variations between sources to identify business risks and opportunities.

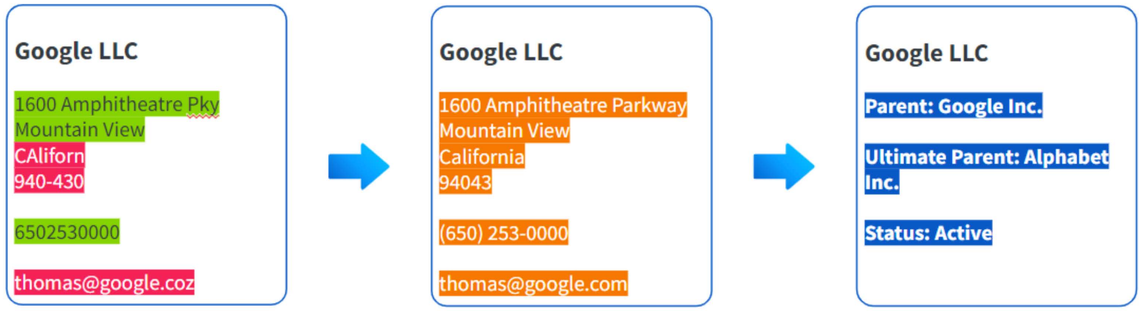
Further, Tamr Cloud is a tool that helps you combine disparate data sources without having to write a significant number of transformations. Tamr Cloud's function is to take data from raw sources, clean it, standardize it, and create an entity layer out of the raw sources for users to consume.

Clean, Standardize, and Consolidate your Data

Data comes in many different schemas and formats. But in order to use the data in machine learning models and business analytics, organizations need to clean it and standardize it. Tamr Cloud has built-in validation and standardization modules for company names, addresses, phone numbers, and more.

Once standardized and validated, Tamr Cloud uses machine learning to cluster records from different sources based on similarities in input attributes. Unlike traditional data integration approaches, Tamr's patented technology does not require rules. Instead, it uses random forest models developed through learnings from human feedback.

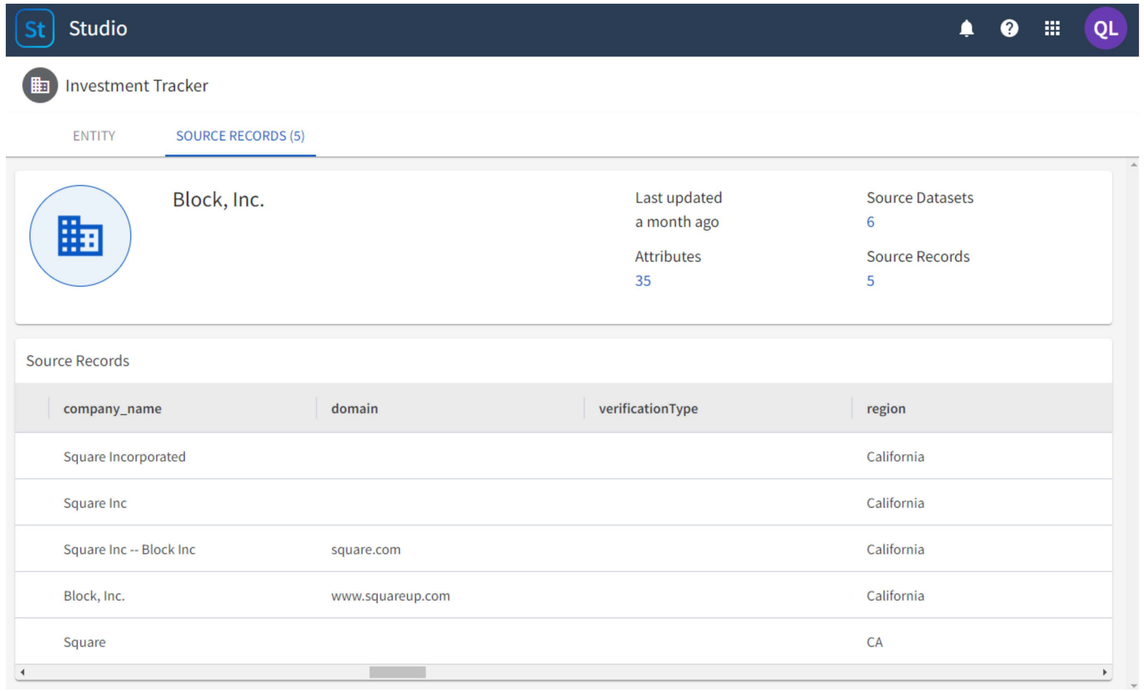
In this example, Tamr Cloud clustered the organization entity, Block, Inc., from five different sources and showed that they are, indeed, the same entity and could be grouped together to provide a holistic view. Tamr Cloud then assigns each entity and cluster a Tamr ID that is persistent and trackable, even as records are merged to or split from clusters.



Validate
Verify syntax and legitamacy

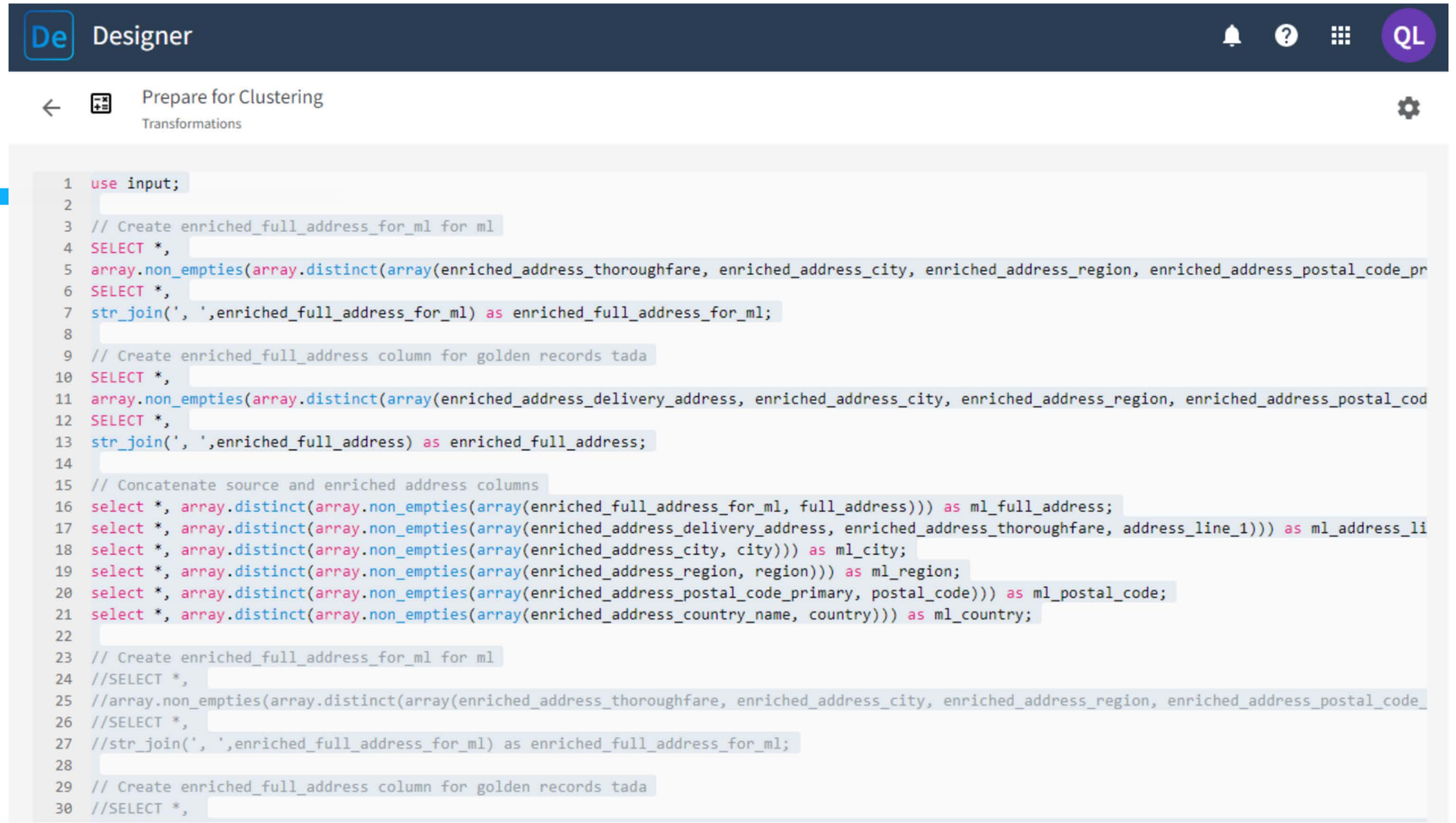
Standardize
Convert values to common, structured formats

Enrich
Correct values and inject new attributes



Finally, you can write custom SQL between each of the steps for bespoke transformations.

Tamr is the pioneer in machine learning-driven data mastering and the only solution that uses machine learning to solve data mastering problems. As of May 2022, Tamr has 14 Patents Issued / Allowed in 10 Patent Families, all in the area of machine learning-based data curation.



Power Your Analytics with Continuously Mastered Entities

Tamr Cloud delivers mastered records for each entity using the best of all available data. And, it continuously updates the pipeline. So as the clusters and input sources change, so will the mastered records.

On top of the mastered records from internal sources, Tamr Cloud also has built-in enrichment from reference data sources such as BvD, S&P, Dun & Bradstreet, and others. Tamr selects and curates these datasets to ensure not only the depth and breadth, but also the timeliness of data to achieve the highest data quality. Tamr Cloud will automatically match your internal data to the same entity in the enrichment sources.

Tamr Cloud supports data orchestration at two levels to better fit into your data ecosystem. At the module level, you can swap each of the “Lego pieces” in and out of the flow to better fit the use case and the underlying entity. At the pipeline level, you can schedule Tamr Cloud’s flow and easily insert it into an existing data pipeline. Alternatively, you can use Tamr Cloud to access data orchestration tools such as Prefect or Airflow for workflow management.

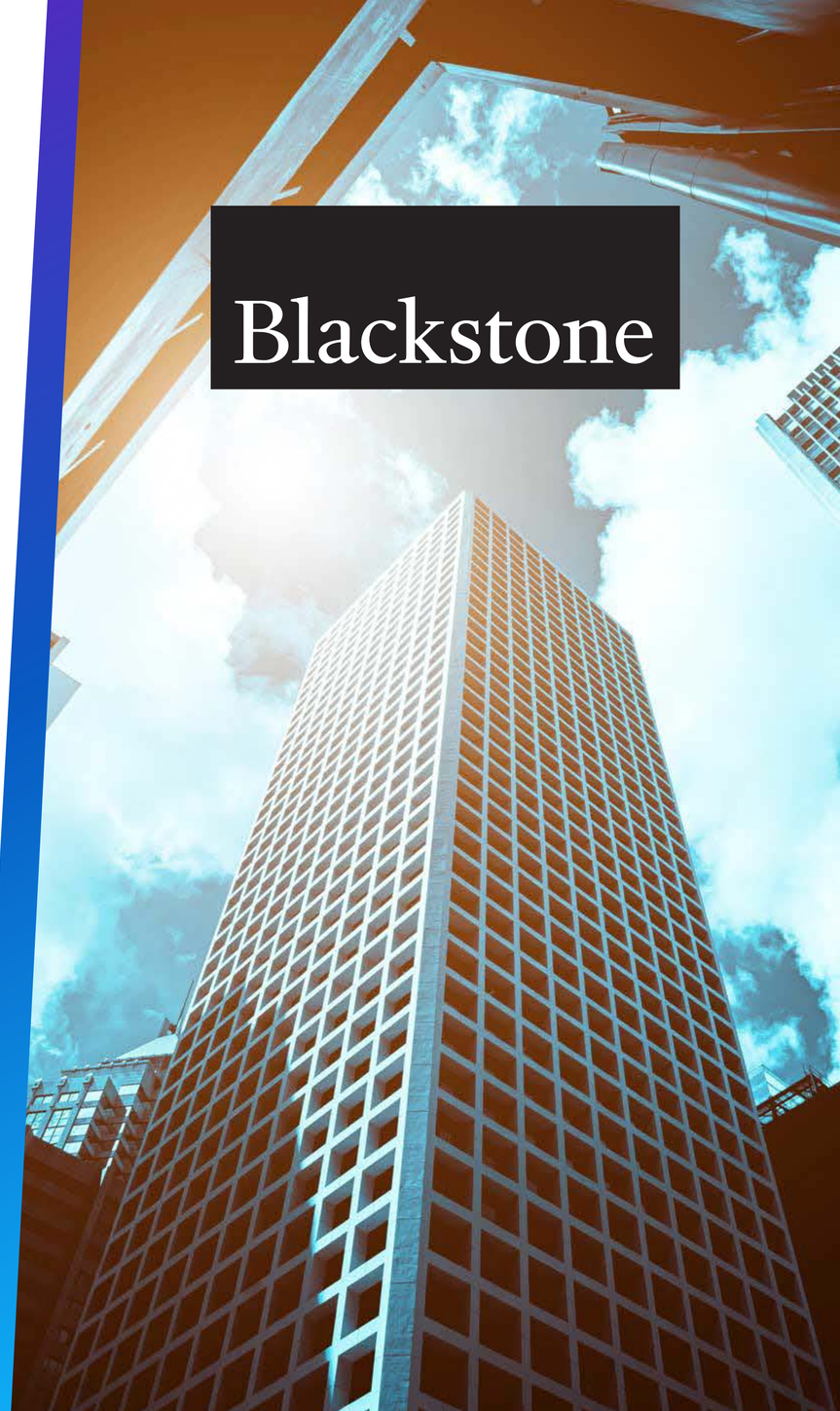
Blackstone: Mastering a Growing Data Ecosystem

Blackstone, one of the world's leading investment firms with over $900B in assets under management, faced a challenge: as the company's portfolio universe grew, so did its data. Duplicate data was persuasive across their data ecosystem, and their existing data mastering processes were manual and unsustainable. It became increasingly difficult to scale their operations to manage the quality and consistency of data, especially as they onboarded additional, third party datasets.

To solve this challenge, Blackstone partnered with Tamr to implement cloud-native, machine learning-driven data mastering on AWS, enabling downstream consumption in Snowflake. The solution, implemented with just two people in 50 days, provides a highly-scalable, cloud-native technology solution and approach that the firm could reuse across use cases. They implemented mastering models around a portfolio with a golden record, enriched by 3rd party sources. And, they took advantage of easy UI workflows and Tamr's patented technology to direct expert time to high-impact edge cases.

As a result, Blackstone realized a quick time to value by mastering multiple types of entities including deals, funds, investment vehicles, properties, and more. Data integration allowed them to reduce human effort while achieving much higher data accuracy. And highly-efficient workflows enriched the company universe with PREQIN, Capital IQ, Pitchbook, Bloomberg, and others.

To see how Tamr Cloud can work for your data, please **request a demo** today.



Tamr

Tamr is the world leader data mastering. We accelerate business outcomes for leading organizations by powering analytic insights, boosting operational efficiency, and enhancing data operations. Tamr's cloud-native solutions offer an effective alternative to traditional Master Data Management (MDM) tools, using machine learning to do the heavy lifting to consolidate, cleanse, and categorize data. Tamr is the foundation for modern DataOps at large organizations including Industry leaders like Toyota, Santander, and GSK. Backed by investors including NEA and Google Ventures, Tamr is transforming how companies get value from their data.

Learn more at **tamr.com**