5-MINUTE GUIDE TO

ENTITY RESOLUTION

Tamr

Introduction

The goal of data analytics and business intelligence tools is to arm people with reliable data to make better business decisions. However, this can only happen if the data feeding these systems is clean, accurate, curated, and holistic.

Business decisions usually center on certain logical entities, such as customers, suppliers, products, etc. For example, when you are researching companies for potential investments, a holistic view of the target company with a consistent set of identifiers (from data sources such as S&P Capital IQ, or Pitchbook) and attributes (LinkedIn employee count growth) can provide context to the business data that matters most to your investment criteria.

More broadly, as data leaders are designing their pipelines and workflows, they should define the entities so that everyone is “speaking the same language” and agrees upon what the data (and metadata) is and is not. These logical entities are the key to leveraging internal structured data to unlock the value of external data and eventually unstructured data.

In this 5-minute guide we’ll dig into this topic a little further and discuss:

- What entity resolution really is
- Challenges to effective entity resolution
- Key principles to instituting entity resolution into your data strategy
- Real-life examples of entity resolution in practice

As data leaders are designing their pipelines and workflows, they should define the entities so that everyone is “speaking the same language” and agrees upon what the data (and metadata) is and is not.

What is Entity Resolution?

Entity resolution (ER), also known as entity linkage or record matching, is a technique used to associate multiple disparate datasets into a logical entity or, in simpler terms, one real-world thing like a person, organization, address, bank account, device, etc. Entity resolution addresses the challenge of reconciling records across (and within) datasets so that the same records are detected, matched, and assigned a unique ID to ensure they are treated as one unique entity going forward.

For example, if a company sells the same product on Amazon and Walmart, the product may have different names, descriptions, and unique IDs on their respective listing pages. This makes critical analysis difficult to undertake, as products that are actually the same appear to be different. If the company were to scrape the product pages of these websites to track sales volume and price, they would quickly find the task nearly impossible to do manually at scale.

Why is this Challenging

Entity resolution can be hard to implement for a multitude of reasons:

1) **Heterogeneity of data quality.** Data from different sources are likely to have different formats such as different ways to represent dates, different ways to write an address, or different abbreviations. On top of that, misspellings, missing information, and even intentional manipulation are common for large quantities of data. All of these can contribute to duplicate data or data points that are difficult to connect.

2) **The problem of scale.** Theoretically, if there are n records in total, then the number of pairs you need to compare to reconcile the data is n-squared. And the heterogeneity of data makes it almost impossible to write rules to match them. A human can likely look at two records and determine if they refer to the same thing or not, but this does not scale across all records.

3) **Changing business context.** Each use case requires specific data sources and has different requirements around the precision and recall trade-offs. For example, a compliance analyst investigating financial crime would want more fuzzy matches whereas a patient matching use case would try to avoid linking two different patients as one. Therefore, you need to have multiple views of entities for different requirements.

Because of these challenges, we believe there are certain key principles that you should consider when choosing entity resolution tools and building your data tech stack.

On top of that, misspellings, missing information, and even intentional manipulation are common for large quantities of data. All of these can contribute to duplicate data or data points that are hard to be connected.

8 Key principles of Entity Resolution

Humans can perform entity resolution well, just slowly. Algorithms can do things fast and at scale, but are too rigid. A good entity resolution tool should be able to provide the best of both worlds:

1 **Time-to-value** - Start the project with value in mind and select solutions that onboard new data sources quickly and easily so that you can get to value quicker. Additional points if there are built-in integrations to external data sources to save more time.

2 **Scalability** - As your organization grows, so does your data volume. It's vital that the technology you choose utilizes cloud infrastructure and is optimized for scalability to the largest volumes of data possible.

3 **Machine Learning** - Look for solutions that use a machine learning-first approach to entity resolution. Machine learning improves with more data. Rules do not. Machine learning increases automation and frees up technical resources by up to 90%. Rules-only approaches have the opposite effect.

4 **Accuracy** - A person can usually match two records and easily determine they are the same entity, even when dealing with poor quality data and incomplete information. Technology solutions should consistently be at least as accurate as a person doing the work manually, if not better.

5 **Flexibility** - One size does not fit all. Depending on the use case, you might want to split the entities or roll them up to the same entity. Your approach should be flexible in dealing with different use cases and be easily adjustable.

6 **Persistent ID** - Data and attributes tend to change over time. As you reconcile different records into the same entity, it's important to maintain a unique, persistent ID in order to have a longitudinal view of the entity.

7 **Data lineage** - As data records move in and out of a cluster because of data changes, it is important to keep track of the record-level data provenance. Additionally, for compliance and audit purposes, it's important to keep track of any changes a person makes to a record.

8 **Enrichment** - **Data enrichment** integrates your internal data assets with external data to increase the value of these assets. It adds additional relevant or missing information so that the data is more complete and usable

Putting Entity Resolution into Practice at Blackstone

Leading global investment firm **Blackstone** implemented a data mastering strategy with Tamr to create golden records of client, property, and other key-asset data across its $8800+ billion portfolios. With over four decades worth of data assets that span systems, platforms, data vendors, and third parties, Blackstone found that not only did the pace of data accumulation accelerate, but the data itself became messier.

Blackstone embarked on a project to create “golden records,” or fully-mastered data, for its key entities, but faced numerous challenges including:

- Duplicate portfolio company data mixed with client data, resulting from the rapid expansion of the portfolio-company universe
- A highly-manual data mastering process that made it difficult to manage the quality and consistency of data
- Difficulty in scaling, specifically related to the onboarding of additional, third-party reference datasets

With Tamr, Blackstone was able to effectively curate and enrich its customer data at scale, allowing them to create the golden records they desired. They were also able to extend these capabilities to their real estate properties portfolio as well, in a unified, coherent, and scalable way. Blackstone now has access to accurate data that is readily available for business decision-making. The firm is also able to:

- Speed up time to value and reduce human effort from data integration to drive higher accuracy
- Develop a highly-efficient workflow that enriches the company universe with PREQIN, Capital IQ, Pitchbook, Bloomberg, and other external data
- Use a single solution to master multiple entity types, including deals, funds, investment vehicles, and properties

Blackstone

With Tamr, Blackstone was able to effectively curate and enrich its customer data at scale, allowing them to create the golden records they desired.

In summary,

the goal of data analytics and business intelligence tools is to arm people with reliable data to make better business decisions. Entity resolution is the way to make sure data feeding these systems is clean, accurate, curated, and holistic. Entity resolution ensures that the best available data is used throughout the company to manage operations and make critical business decisions.

Reference:
Entity Resolution for Big Data: A Summary of the KDD 2013 Tutorial Taught by Dr. Lise Getoor and Dr. Ashwin Machanavajjhala

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**